water

MDPI

Article

Developing a Data Quality Evaluation Framework for Sewer Inspection Data

Hossein Khaleghian * and Yongwei Shan *

School of Civil and Environmental Engineering, Oklahoma State University, Stillwater, OK 74078, USA;
hossein.khaleghian@okstate.edu (H.K.); yongwei.shan@okstate.edu (Y.S.); Tel.: +1-(405)-762-1709 (H.K.);
+1-(405)-744-7073 (Y.S.)

Abstract: The increasing amount of data and the growing use of them in the information era have raised questions about the quality of data and its impact on the decision-making process. Currently, the importance of high-quality data is widely recognized by researchers and decision-makers. Sewer inspection data have been collected for over three decades, but the reliability of the data was questionable. It was estimated that between 25% and 50% of sewer inspection data is not usable due to data quality problems. In order to address reliability problems, a data quality evaluation framework is developed. Data quality evaluation is a multi-dimensional concept that includes both subjective perceptions and objective measurements. Five data quality metrics were defined to assess different quality dimensions of the sewer inspection data, including Accuracy, Consistency, Completeness, Uniqueness, and Validity. These data quality metrics were calculated for the collected sewer inspection data, and it was found that consistency and uniqueness are the major problems based on the current practices with sewer pipeline inspection. This paper contributes to the overall body of knowledge by providing a robust data quality evaluation framework for sewer system data for the first time, which will result in quality data for sewer asset management.

Keywords: data quality; sewer infrastructure; pipeline assessment certification program; sewer asset management

check for updates

Citation: Khaleghian, H.; Shan, Y. Developing a Data Quality Evaluation Framework for Sewer Inspection Data. *Water* **2023**, *15*, 2043. https://doi.org/10.3390/w15112043

Academic Editors: Keshab Sharma and Oriol Gutierrez

Received: 19 April 2023
Revised: 16 May 2023
Accepted: 22 May 2023
Published: 27 May 2023

1. Introduction

The general condition of America's infrastructure is alarmingly poor. According to the American Society of Civil Engineers (ASCE) 2021 Infrastructure Report Card, the average grade point for the overall infrastructure is C-. It is estimated that $2.6 trillion is needed for the next 10 years to restore the nation's infrastructure systems to good condition. Among these systems, wastewater received a grade of D+ and needs more than $270 billion in improvements over the next 10 years. Sewer pipelines are the primary component of wastewater systems, and they consume approximately 80% of the capital investment for wastewater. There are nearly 800,000 miles of public sewer pipelines, and many of them are at the end of their service life [1]. As a result, understanding the current condition of the sewer system is a critical step for infrastructure asset management strategies and improving national wastewater systems [2].

Quality data that document the current condition of sewer pipelines is fundamental for the development of sewer asset management tools and strategies [3]. Significant efforts have been made to evaluate the condition of sewer systems and determine the factors affecting them. Several different deterioration models have been developed to assess pipe conditions. These models can be divided into two groups: (1) function-based models, and (2) data-based models [4,5]. Function-based models use statistical methods such as regressions and Markov chains, while data-based models use artificial intelligence (AI) and machine learning techniques such as artificial neural networks (ANN) or random forests [6–11]. These models determined the main factors that have significant effects on pipe condition, such as age, depth, length, soil type, location, size, and material; however, a common concern raised by those studies was data availability and data quality [7,9,12,13].

Artificial intelligence (AI) and machine learning have recently gained popularity as methods for data analysis; yet, in most cases, the necessary data infrastructure is not present to use such tools. Before adopting AI and machine learning algorithms, a solid foundation for data is necessary [14]. Data science requirements are shown in Figure 1. Data collection is at the bottom of the pyramid. Then, reliable data flow and structured data storage are needed to make it accessible. Data quality management, an underrated side of data science, and data preparation is the next step in making it reliable for optimization and analytics. Although these two procedures are essential to data science, they are frequently neglected. While the amount of collected data is increasing rapidly, evaluating data quality is becoming a big issue [15]. Any data analytics tools, charts, and algorithms will be worthless if they have been developed based on low-quality data.

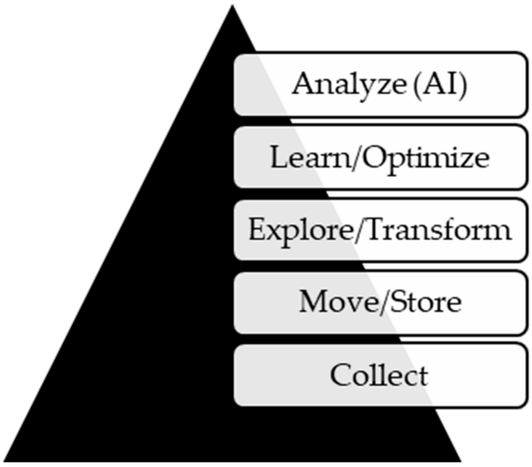


Figure 1. Hierarchy of needs in data science [14].

The accessibility and reliability of the current sewer inspection data are questionable due to several different factors, such as operator experience, input errors, data schema problems, data standard versions, and data collection software incompatibilities. Due to the quality issues in the current data collection practices, it is estimated that between 25% and 50% of data is eliminated to make the data ready for analysis [16,17]. This approach may result in underestimation of the severity of the current condition of the system and false outputs. To address this problem and to resolve reliability issues in the collected databases, a comprehensive data quality evaluation framework must be developed.

The evaluation of data quality is typically taken into account in response to the issues that arose throughout the decision-making process. This reactive strategy may address the problems with the present database, but it will not go to the root of the issue and prevent further quality problems. Since data defects can happen at any time and have an impact on the quality of the data, data quality evaluation is a continual endeavor [18].

Data quality should be considered through the intended use of the data and will be defined based on the relevance to the context of the data to be used [19]. Data quality is a long-lasting issue in the field of civil infrastructure condition assessment [20].

Currently, closed-circuit television inspection (CCTV) is the major source of information (more than 60%) for defining maintenance and rehabilitation projects for sewer systems [21]. As a result, the quality of the obtained CCTV data plays a crucial part in the correctness of the final conclusions. It has been found that the likelihood to overestimate a pipe in bad condition is 20%, and the probability to underestimate a pipe in good condition is 15%. It has been noticed that to generate the data for this evaluation, only 65% of total inspections have been analyzed, and the rest of the data has been neglected due to inconsistency, incompleteness, and lack of reference keys [17]. This data elimination practice would result in underestimating the severity of the system by neglecting the assets that could have more severe conditions in the system.

It has been acknowledged that each database's data quality needs to be assessed to remedy this issue in sewer inspection data. The objective of data quality evaluation is to ensure that the inspection data are accurate and consistent with other datasets. This process is a significant step in developing sewer systems data inventory by integrating existing datasets [22,23].

The objective of this research is to provide a framework for evaluating the data quality of the collected sewer system databases for the first time. The data quality metrics were developed based on the literature and sewer inspection data requirements. Then, the data were evaluated based on the defined metrics to determine the quality problems within the database. The results were reported, and the root cause of each quality issue was identified to provide the correction suggestion and implement the resolution.

2. Literature Review

Municipalities have been recording multiple forms of data on sewer pipe conditions, including closed-circuit TV, sonar, laser, and acoustic, as a basis for capital improvement and asset management plans. However, the benefits of data-driven decisions can only be obtained if data quality is guaranteed. In previous studies on the quality issues of sewer inspection data, it was concluded that the quality problems mainly occurred due to the operators' level of experience [24–26]. Fisher explained that the quality of inspection data depends on the skill and motivation of the operator [24]. Comparing sewer pipe inspections by various operators, only 16% of the 307 inspections found similar numbers of defects. The following suggestions have also been made to improve the quality of sewer pipe inspections [26].

1. The inspection coding system should be simplified to avoid misclassification of the defects.
2. The defect image should be evaluated with the defect information to avoid misinterpretation of the defects.
3. The sewer inspectors should be provided with reliable feedback on their inspection evaluations.

Although these suggestions can improve the quality of the inspection records, they do not address the current issues within the sewer inspection databases.

As the usage of data analysis of the collected infrastructure data for asset management decisions is trending up, poor data quality can have a negative impact on the condition of infrastructure due to ineffective decisions and poorly performing decision models [27]. It is challenging to assess data quality if it is not quantitatively defined. Moreover, the data context should be taken into account when enhancing data quality [28]. The data quality evaluation consists of three steps: (a) identify, (b) measure, and c) resolve. Decision making and data quality management are facilitated by this procedure. The procedure for evaluating data quality is shown in Figure 2 [29,30]. This study focused on identifying and measuring data quality problems.

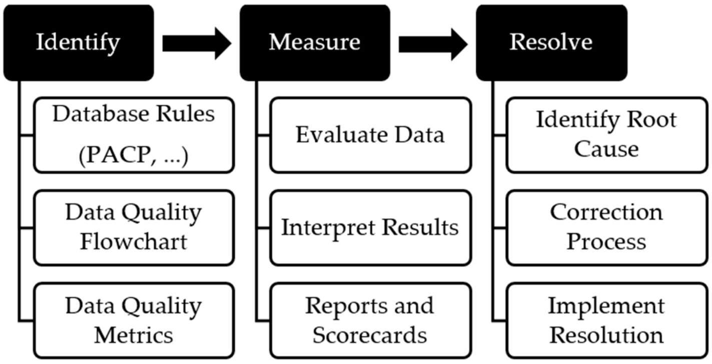


Figure 2. Data quality evaluation process.

The quality evaluation for every database differs. Thus, domain experts should identify the database rule, metrics, and evaluation process. These rules and metrics could be domain-independent or domain-specific based on the database requirements [31]. Then, the quality measures identified and established in the previous step were used to evaluate a database. Finally, the cause of data quality problems can be identified and addressed to avoid future problems as well. The quality assessment process is the main focus of this work.

Data quality evaluation has become the center of attention, specifically in business and healthcare sectors where data analysis is the main decision support tool [32–35]. Developing an evaluation process and defining a set of data quality metrics is a regular practice in many academic and professional fields. Data quality is a rational approach to defining a set of dimensions to measure and improve the quality of data. Defining the data quality dimensions for a database is the first difficult step [36], since dimensions should consider specific applications and uses of the data.

Data quality is a multi-dimensional concept that includes both subjective perceptions and objective measurements. The experience of the individuals involved with the data forms a subjective assessment of the data quality. Objective assessment can be divided into two categories: (1) task-dependent or (2) task-independent. Task-independent metrics are developed without considering database rules or restrictions, while task-dependent metrics include them. Pipino et al. [23] provided three functional forms for objective data quality metrics that consider objective and subjective assessments:

1. Simple Ratio: The ratio of the positive outcome to the total outcome is a simple way to measure different dimensions. It considers that 1 or 100% is the total desired outcome, and the ratio will show positive outcomes.
2. Min or Max Operations: This form is used when the data quality dimension is a combination of several variables. The min or max values will be compared to the preassigned values.
3. Weighted average: The weighted average can be calculated for dimensions with multiple variables. Each variable is weighted according to its importance between 0 and 1 with the sum of 1. This form can provide an appropriate measurement if precisely developed.

In previous studies, several different sets of data quality metrics have been developed for data quality evaluations. Table 1 shows the most common data quality metrics.

Table 1. Data quality metrics examples.

References	Data Quality Metrics
Pipino, Lee [23]	Accessibility, Appropriate Amount of Data, Believability, Completeness, Concise Representation, Consistent Representation, Ease of Manipulation, Free-of-Error, Interpretability, Objectivity, Relevancy, Security, Timeliness, Understandability, Value-Added
Piprani and Ernst [37]	Accuracy, Completeness, Consistency, Precision, Reliability, Temporal Relatability, Timeliness, Uniqueness, Validity
Nousak and Phelps [30]	Validity, Completeness, Consistency, Uniqueness, Timeliness, Accuracy, Precision
Loshin [17]	Uniqueness, Accuracy, Consistency, Completeness, Timeliness, Currency, Conformance

These metrics are used to assess different data quality dimensions. In order to measure these metrics, definitions should be provided, and the measurement techniques should be defined. These techniques can be quantitative or qualitative based on the provided definition [35].

3. Data and Methods

3.1. Data

In 2002, the National Association of Sewer Service Companies (NASSCO) developed the Pipeline Assessment and Certification Program (PACP) to become a standard for the evaluation of sewer pipe closed-circuit television (CCTV) inspections [38]. Prior to the PACP, there was no standardized protocol in the United States for the collection and management of data related to pipelines' internal inspections. The primary purpose of PACP is to ensure that all data describing the conditions within a pipe are collected and coded in a consistent and reliable manner. PACP became the industry standard for sewer condition data, and it was implemented by more than 200 cities and utility districts in the United States and Canada [39].

Four different sewer pipe inspection databases (shown in Table 2 were used to test the data quality evaluation framework to be described in Section 3.2. The data were collected from municipalities in the US. Three of them were collected through PACP-Certified software and are in the PACP 6 standard. However, DB3 was not in the PACP data structure, and the data transformed into the standard format. The other database is for a small municipality that has a total of 775 miles of sanitary sewer lines (ranging from 6″ to 54″), 9100 manholes, and 70 lift stations. Data are not in PACP format and are hardly understandable. However, it provides cross-check references to evaluate data quality metrics. The city provided 2.2 miles of inspections from the downtown area. In Table 2, the number of conditions refers to the total number of codes associated with inspections. In other words, an inspection record typically has multiple condition records to describe the defects that appear in one inspection.

Table 2. Sewer pipe databases.

Database	State	No. of Inspections	No. of Conditions	PACP Version
DB1	Texas	72	724	N/A
DB2	Indiana	5232	84,785	PACP 6
DB3	California	6418	85,255	PACP 6
DB4	Pennsylvania	1169	14,341	PACP 6

3.2. Data Quality Metrics

The development of well-defined data quality metrics is essential for making significant data-driven decisions. These metrics could evaluate the data based on the context and schema and provide subjective and objective assessment [40]. To improve the data quality of sewer inspection data, this paper follows a 5-pillar data quality management technique defined by Lebied [41]:

1. The people: The quality of data relies on the individuals who implement it
2. Data profiling: Reviewing data and comparing data to metadata
3. Defining data quality: Developing data quality rules and metrics based on the context and use of data (business rules).
4. Data reporting: Identifying data errors and reporting for the resolution process
5. Data repair: Addressing data error in the most efficient way

In this study, more than 100,000 inspection records were evaluated in order to build an efficient data quality management strategy for sewer system data, and more than 50 industry experts offered their opinions on the efficacy of the gathered sewer data (Pillar 1). Based on industry needs and regulations, primary data quality issues were identified (Pillar 2). Then, based on earlier discoveries (Pillar 3), data quality metrics were developed and reported (Pillar 4). Finally, a workable solution for several data quality challenges has been offered and put into practice (Pillar 5).

As mentioned above, a set of data quality metrics is required to assess the data quality of the collected sewer inspection data based on the rules and database requirements. The following rules were considered during the development process of data quality metrics [28]:

metrics should (1) be insensitive to changes in the number of records in the database, (2) accurately reflect the degree of the data quality requirements, (3) be independent of each other, (4) be limited to a reasonable number, and (5) address database rules.

Defining a proper set of data quality metrics simplifies the measurement of the quality of the data and provides a quantitative structure for data quality evaluation. Data quality rules are integrated into quality metrics and provide a tool for data quality management [17]. Based on the five rules described above, the authors developed a proper set of data quality metrics, shown in Table 3, based on relevancy to the data collected on sewer systems. However, it is important to mention that these metrics can be calculated based on data availability.

Table 3. Data quality metrics.

Name	Description
Accuracy	Data element values are properly assigned
Consistency	Data element is free from variation and contradiction based on the condition of another data element
Completeness	Data element is required based on the condition of another data element and database rules (required and optional data)
Uniqueness	Data element is unique
Validity	Data element passes all requirements for acceptability (PACP Rules) *

Note: * Pipeline Assessment Certification Program.

The description of each metric is as follows:

Accuracy (ACC) indicates whether the data have significant errors. It can be measured by source documentation or a comparison of the attributes in a database. It can also be checked logically to see if it falls within accepted bounds and makes sense. The metric is defined as the number of errors divided by the total number of attributes subtracted from 1:

$$Accuracy = 1 - \frac{(total\ number\ of\ errors)}{(total\ number\ of\ attributes)} \tag{1}$$

Consistency (CNS) indicates whether the data are presented in the same format. The metric is defined as the number of violations divided by the total number of consistency checks subtracted from 1:

$$Consistency = 1 - \frac{(total\ number\ of\ violations)}{(total\ number\ of\ consistency\ checks)} \tag{2}$$

Completeness (COM) indicates whether there are any missing values in the database. Completeness is defined based on the database rules, as not all the attributes are required for database fields. The metric is calculated by the ratio of the incomplete units to the total number of units and subtracting from 1:

$$Completeness = 1 - \frac{(incomplete\ units)}{(total\ number\ of\ units)} \tag{3}$$

Uniqueness (UNI) indicates whether a data record is represented uniquely in the database and no entity exists more than once. In other words, uniqueness captures redundancy in the database. It is important to identify duplicates and either merge them or delete the duplicates. Redundancy measures the occurrence of data, and uniqueness is calculated by subtracting redundancy from 1:

$$Uniquness = 1 - \frac{(Number\ of\ occurrance)}{(total\ number\ of\ entities)} \tag{4}$$

Validity (VAL) indicates whether the database complies with standards. Pipeline Assessment Certification Program (PACP) is a widely accepted standard for collecting

CCTV inspections in the US. Thus, this metric is only applicable to databases that have been collected in the PACP standard. To measure this metric, all PACP requirements should be evaluated to calculate the validity of the database. The metric is calculated as follows:

$$Validity = 1 - \frac{(Number\ of\ invalid\ attributes)}{(total\ number\ of\ attributes)} \tag{5}$$

To calculate the invalid attributes in the collected databases, a referential database was created that included all PACP (V6) requirements in 5 tables (Figure 3). The 'pacp_code' table assigns a primary key to each of the codes that can be referenced in the 'pacp_rule' table for different versions of PACP. The 'pacp_rule' table records the PACP requirement for every defect, including clock positions, values, and Joint. The 'rule_details' table is used to evaluate the limits for each field in the 'pacp_rule' table. For example, deformation (D) requires a value_percent of less than or equal to 40%. This information is retrieved from 'rule_details'. There are also some requirements in the PACP that some codes can or cannot be used together. The common defect is the collapse (XP) in the pipe, and no code other than MSA (survey abandoned) should be used after that. These data are recorded in the 'cross_check' table. Some codes are related to specific materials, and the 'material' table provides this information.

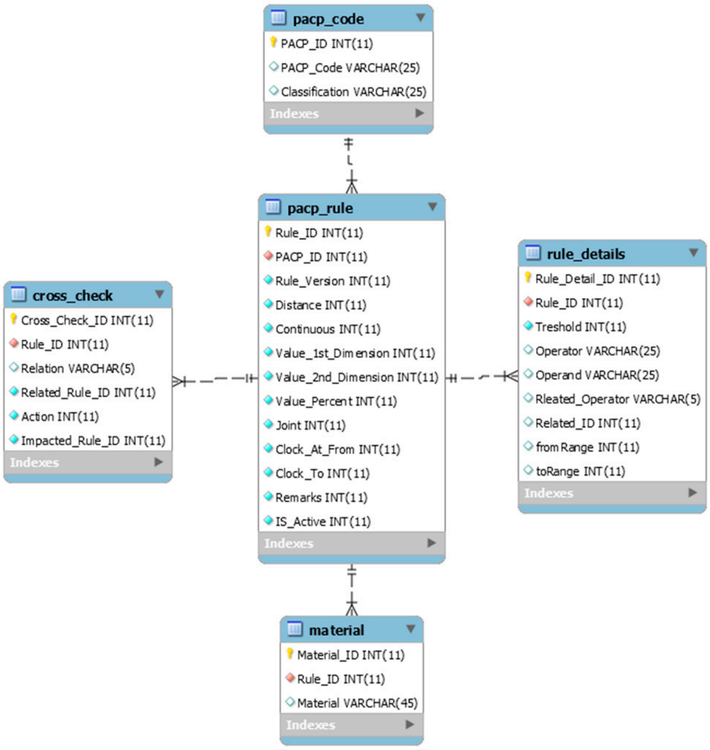


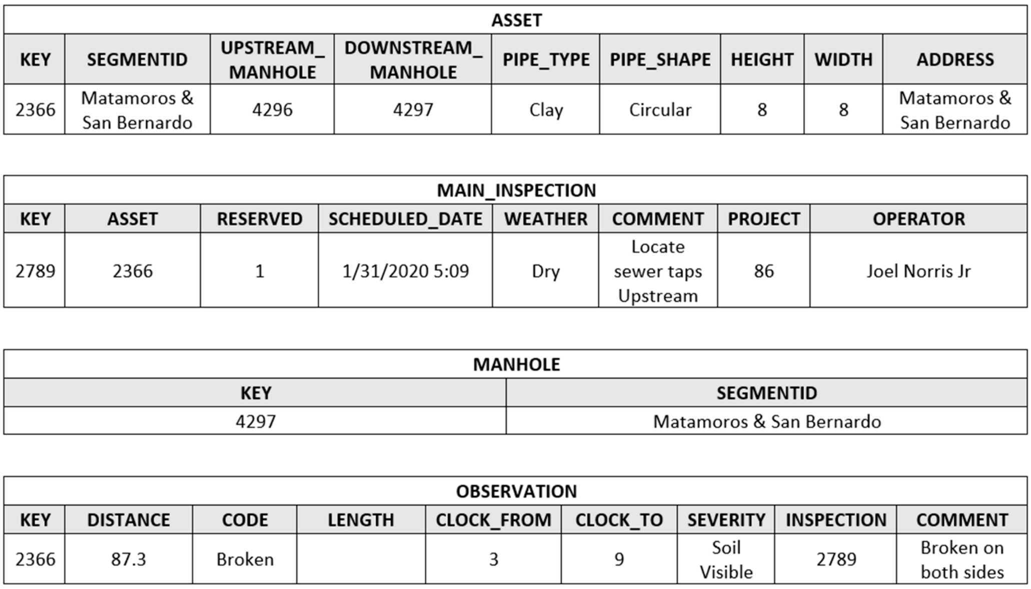
Figure 3. PACP referential database.

In order to evaluate the validity metric, a Python script was developed that checked all PACP requirements in the database. Moreover, a sample database was developed for the validation process of the developed code for all types of material and all the structural defects in the PACP manual. The developed code detected 100% of the invalid attributes in the sample database and can be applied to the collected sewer inspection data.

4. Results

4.1. Accuracy, Consistency, Completeness, and Uniqueness Metrics

In order to evaluate the first four data quality metrics that were developed for sewer system data, DB1 was used. The reason that the first four metrics were only calculated for DB1 is that multiple databases were available, including manhole data, asset data, and CCTV inspections. Cross-referencing is the main criterion for calculating the first four metrics. Since the PACP databases only included the CCTV inspections (Inspection table and Condition table), only validity metrics were calculated. Figure 4 shows the data structure preferred by the city.



ASSET

KEY	SEGMENTID	UPSTREAM_ MANHOLE	DOWNSTREAM_ MANHOLE	PIPE_TYPE	PIPE_SHAPE	HEIGHT	WIDTH	ADDRESS
2366	Matamoros & San Bernardo	4296	4297	Clay	Circular	8	8	Matamoros & San Bernardo

MAIN_INSPECTION

KEY	ASSET	RESERVED	SCHEDULED_DATE	WEATHER	COMMENT	PROJECT	OPERATOR
2789	2366	1	1/31/2020 5:09	Dry	Locate sewer taps Upstream	86	Joel Norris Jr

MANHOLE

KEY	SEGMENTID
4297	Matamoros & San Bernardo

OBSERVATION

KEY	DISTANCE	CODE	LENGTH	CLOCK_FROM	CLOCK_TO	SEVERITY	INSPECTION	COMMENT
2366	87.3	Broken		3	9	Soil Visible	2789	Broken on both sides

Figure 4. DB1 data structure.

The 'ASSET' table includes data related to the sewer pipe constant features. The 'MANHOLE' table provides manhole IDs. 'MAIN_INSPECTION' stores the inspection data each time it is done, and the 'OBSERVATION' table contains the pipe's internal conditions for each inspection. These tables are connected based on the primary key of each table as a point of reference. In order to evaluate the data quality of this database, each table was assessed separately. The percentages show the metric values and do not applicate (N/A) represent the fields that could not be calculated due to lack of references. Figure 5 shows the data quality evaluation process of DB1.

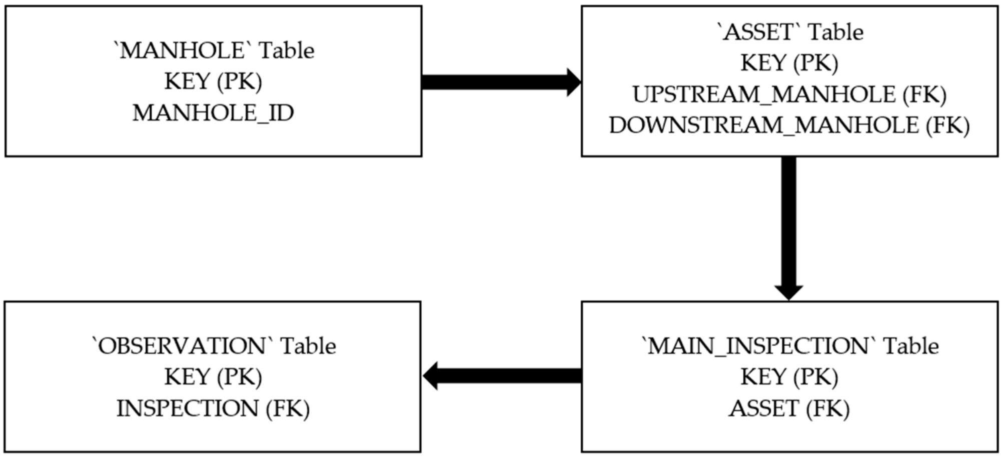


Figure 5. Data quality evaluation of DB1.

4.1.1. 'MANHOLE' Table

This table contains 143 records and includes the manhole IDs that should be unique for each manhole.

Key: A unique number is assigned to each manhole to identify them.

MANHOLE_ID: The ID is not well defined. The ID is supposed to be unique, so it can be used as an identifier for the manholes in the system. However, the ID is developed based on the address, and only the street name or the intersection has been used. It has been noticed that the software required unique values; therefore, different variations of manhole names have been used by the operators, such as “Iturbide M/H”, “Iturbide M/H ,.”, “Iturbide M/H,,”, and “Iturbide M/H...”.

As mentioned above, all the metrics are not applicable to each data set. There is one missing manhole ID, 117 inconsistencies in data format, and only 68 unique values. Table 4 provides the results of the data quality evaluation.

Table 4. Data quality evaluation of the ‘MANHOLE’ table.

Column	Description	COM	CNS	UNI	ACC
KEY	Primary Key	100%	100%	100%	N/A
MANHOLE_ID	It represents the manhole address	99.30%	18.75%	47.55%	N/A

Note: Com = Completeness, CNS = Consistency, Uni = Uniqueness, Acc = Accuracy.

4.1.2. ‘ASSET’ Table

This table has 72 records and provides pipe features, including manhole information, material, shape, dimension, and length. Each field was evaluated separately, and the results of the data quality evaluation are described in Table 5.

Table 5. Data quality evaluation of the ‘ASSET’ table.

Column	Description	COM	CNS	UNI	ACC
KEY	Primary key	100%	100%	100%	N/A
SEGMENTID	Developed based on MAHONHOLE_ID	100%	25%	59.72%	62.5%
WIDTH	The width is not required for circular pipes (all pipes in the database)	100%	100%	0%	91.67%
ADDRESS	Based on SEGMENTID	100%	30.56%	0%	94%
ASSET_LENGTH	The total length of the pipe	80.56%	100%	0%	0%

KEY: It is the primary key of the table and any data error in this field can result in data integrity issues.

SEGMENTID: This ID is generated by the upstream or downstream MANHOLE_ID. However, the data are not consistent with each other and the source field (MANHOLE.MAN HOLE_ID). This field should be redefined to assign a unique ID to each pipe.

UPSTREAM_MANHOLE/DOWNSTREAM_MANHOLE: The current information is continuous numbers assigned to the manholes, and each manhole should have its own identification number to be distinguished from the others. This field cannot be evaluated because the reference to each manhole is not available (GIS data are required).

WIDTH: For circular pipes, the width is considered redundant information and can be eliminated. In addition, some inputs do not match the height of the pipes, which makes the accuracy of the pipe shapes questionable.

ADDRESS: The current information is repeated data from another field (SEGMENTID). Moreover, the data are not consistent with each other and the source data.

ASSET_LENGTH: For all the inspections, MAIN_INSPECTION.SURVEYED_FOOTAGE has been assigned as ASSET_LENGTH. Since these two variables are different from each other, it can be concluded that this field is inaccurate.

4.1.3. ‘MAIN_INSPECTION’ Table

This table contains 72 records, providing general information on the inspection such as operator, weather, date, direction, and comments. It can be considered the header in the PACP database. Table 6 provides an analysis of the data quality of this table.

Table 6. Data quality evaluation of the 'MAIN_INSPECTION' table.

Column	Description	COM	CNS	UNI	ACC
KEY	Primary key	100%	N/A	100%	N/A
ASSET	Foreign key to ASSET table	100%	100%	100%	100%
COMMENT	Direction and location	100%	70%	18%	87%
OPERATOR	The name of the inspector	100%	N/A	N/A	N/A
REASON	The purpose of the inspection	98%	N/A	N/A	N/A
SURVEYED FOOTAGE	The length of the surveyed segment	98%	100%	N/A	100%

KEY: It is the primary key of the table and any data error in this field can result in data integrity issues.

COMMENT: The location and direction are considered redundant information. 100% accurate direction can be extracted from the 'OBSERVATION' table. Other information on the pipe condition and material is useful. This information can be extracted.

4.1.4. 'OBSERVATION' Table

This table contains 724 records, providing condition information, including defect type, distance, clock positions, and severity. The data quality of the observations could not be assessed through the same approach applied before. This table is similar to the 'Condition' table in the PACP database. In order to have a comprehensive data quality evaluation of this table, each defect type should be evaluated separately. For all the codes, no information was provided on the length, width, or percentage. Approximately 77% of the observations are related to construction and miscellaneous features such as manhole location, water level, direction, etc. Each code has been evaluated separately, and the results are presented in Table 7. Accuracy is the only metric that could not be evaluated for the codes based on the current database, as it requires another reference document.

Table 7. Data quality evaluation of 'OBSERVATION' table.

Code	COM	CNS	UNI	ACC
Abandoned Survey	75%	51%	75%	61%
STOP	100%	98%	0%	100%
Water Level	57%	50%	28%	N/A
Water Mark	75%	25%	25%	N/A
Camera Under Water	0%	100%	50%	N/A
General Observation	88%	35%	12%	N/A
Pipe Type	81%	52%	52%	N/A
MANHOLE	91%	78%	13%	91%
Lateral	97%	88%	9%	N/A
Lateral Connection Problem	100%	66%	0%	N/A
Intruding Sewer Tap	100%	75%	0%	N/A
Broken	65%	60%	90%	N/A
Crack	37%	37%	100%	N/A
Joint Offset	46%	46%	100%	N/A
Joint Separated	27%	27%	100%	N/A
Sag	50%	50%	100%	N/A
Root in Lateral/Root in Joint	66%	66%	100%	N/A
Obstacle	100%	100%	100%	N/A

4.2. Validity Metric

In order to calculate the Validity metric, three sewer inspection databases were evaluated. These databases were collected based on the PACP 6 standard through two different

PACP-certified software programs. It was recognized that the PACP data collected with one software is not compatible with another software due to the software's violation of the PACP rules. Table 8 shows the validity metric in each of these databases based on the number of invalid attributes.

Table 8. Validity metric calculation.

	DB2	DB3	DB4
Total Number of Attributes	836,313	857,147	145,435
Total Number of Invalid Attributes	90	55,216	7551
Validity	100%	94%	95%

5. Discussion

Data quality is one of the major challenges in the asset management process since decision-makers rely more on data to implement their objectives. Data error rates exceeding 75% have been observed in the civil engineering industry, and errors of up to 30% are usual [42]. Asset management's primary goal is to provide a proper level of service by effectively managing the infrastructure through repair and replacement. The structural and hydraulic performance of the sewer network serves as the basis for these initiatives, with structural performance serving as the primary budgetary consideration [17,43].

The data quality evaluation framework was developed based on five data quality metrics. These metrics are defined quantitatively to measure the different quality dimensions of sewer inspection data. Each metric was calculated based on data availability and relevancy. Table 9 shows the total quality evaluation of DB1.

Table 9. Total quality evaluation of DB1.

Data Table	COM	CNS	UNI	ACC	MEAN
'MANHOLE' Table	99.65%	59.38%	73.78%	N/A	77.60%
'ASSET' Table	96.11%	71.11%	31.94%	62.04%	65.30%
'MAIN_INSPECTION' Table	99.33%	90.00%	72.67%	95.67%	89.42%
'OBSERVATION' Table	82%	75%	34%	94%	71.25%

Redundancy and inconsistency are major issues in the data. Redundancy is mainly related to the schema and can be addressed through normalization. This metric can help municipalities avoid further mistakes in the decision-making problem. One significant problem regarding redundancy is determining asset location. It has been noticed that poor data management has resulted in mislocating assets. High-quality GIS data can resolve this problem. In DB1, it has been calculated that manholes have only 47.55% uniqueness and pipes have 59.72%.

Inconsistency can cause issues when integrating databases into a common repository. The same data has been stored in different locations and formats. This can also cause problems in query and script development to retrieve data. In the 'ASSET' table, the SEGMENTID was developed based on inconsistent criteria. Some were related to the Downstream Manhole or Upstream Manhole, and the others were a combination of those two or even unrelated. This type of inconsistency can also cause incompatibility issues when different databases are being analyzed together. Developing proper metadata can resolve the inconsistency problems.

Validity evaluates the database to comply with the data rules and standards. While municipalities are using the PACP coding system, their final database is not always PACP compatible. The validity metric can provide the PACP compatibility of the database. This can help municipalities understand the condition of their data and provide them with a solution to access their data across different software platforms. It can also help operators understand the common mistakes at the time of collecting data.

6. Conclusions

The effectiveness of sewer asset management decisions hinges upon the quality of condition assessment data collected on the sewer infrastructure. However, the quality of the data collected by the municipalities varies among municipalities. This research showed that 11% to 35% of sewer inspection records have quality issues that need to be addressed in order to develop optimum asset management decisions. Thus, it is important to develop a data quality evaluation framework to identify and measure the current problems within the collected databases and to provide a feasible resolution to address them and prevent similar problems in the future.

In this paper, a data quality evaluation framework was developed based on five quality metrics to provide a quantitative assessment of current problems. Each metric was calculated based on the data context. It has been noticed that data consistency and uniqueness are the major problems in the collected databases. These two can be addressed by implementing robust database management practices. Database normalization can help reduce data redundancy and improve data integrity. In addition, GIS integration will resolve inconsistencies and improve the accuracy of the data. By addressing these problems, the development of infrastructure asset management plans can be facilitated.

One of the main problems in evaluating data quality is data accessibility. This problem was found in all the collected databases, specifically in DB1, where the city could only extract a few amounts of data. This is one of the major limitations in the evaluation and analysis of sewer system databases. Differences in data management practices among municipalities are another challenge. For instance, while some providers stored their data in a single data repository (DB1 and DB2), others kept their data in separate datasets based on different criteria (DB3 and DB4). This practice resulted in several small databases, which made the proposed data quality evaluation more complicated. Although most of the databases were collected in the PACP standard, some interoperability issues occurred because the data were exported from different software into nonstandard data structures, which proved to be a common problem.

This study contributed to developing a quantitative analysis of the quality problems in sewer inspection data and, for the first time, providing tools for industry stakeholders to address these problems.

Author Contributions: Conceptualization, H.K. and Y.S.; methodology, H.K.; software, H.K.; validation, H.K. and Y.S.; formal analysis, H.K.; investigation, H.K.; resources, Y.S.; data curation, H.K.; writing—original draft preparation, H.K.; writing—review and editing, Y.S.; visualization, H.K.; supervision, Y.S.; project administration, Y.S.; funding acquisition, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation, Grant No. 2141184, and the National Association of Sewer Service Companies, Grant No. AA-5-59015.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the confidentiality agreement signed with the data providers.

Acknowledgments: The authors would like to thank the municipalities that generously provided the data for this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. ASCE. 2021 Report Card for America's Infrastructure. American Society of Civil Engineers. 2021. Available online: www.infrastructurereportcard.org/ (accessed on 10 March 2023).
2. Lewis, P.; Shan, Y.; Khaleghian, H. One Voice for Sewer Condition Assessment and Asset Management. In *Pipelines 2016*; American Society of Civil Engineers: Reston, VA, USA, 2016; pp. 515–526.
3. Khaleghian, H.; Shan, Y.; Lewis, P. Development of a Quality Assurance Process for Sewer Pipeline Assessment and Certification Program (PACP) Inspection Data. In *Pipelines 2017*; American Society of Civil Engineers: Reston, VA, USA, 2017; pp. 360–369.

4. Mohammadi, M.M.; Najafi, M.; Kaushal, V.; Serajiantehrani, R.; Salehabadi, N.; Ashoori, T. Sewer pipes condition prediction models: A state-of-the-art review. *Infrastructures* **2019**, *4*, 64. [CrossRef]
5. Caradot, N.; Sampaio, P.R.; Guilbert, A.S.; Sonnenberg, H.; Parez, V.; Dimova, V. Using deterioration modelling to simulate sewer rehabilitation strategy with low data availability. *Water Sci. Technol.* **2021**, *83*, 631–640. [CrossRef]
6. Abraham, D.; Wirahadikusumah, R. Development of prediction models for sewer deterioration. In Proceedings of the Eighth International Conference on Durability of Building Materials and Components, Vancouver, BC, Canada, 30 May–3 June 1999; Volume 8.
7. Ariaratnam, S.T.; El-Assaly, A.; Yang, Y. Assessment of infrastructure inspection needs using logistic models. *J. Infrastruct. Syst.* **2001**, *7*, 160–165. [CrossRef]
8. Chughtai, F.; Zayed, T. Infrastructure condition prediction models for sustainable sewer pipelines. *J. Perform. Constr. Facil.* **2008**, *22*, 333–341. [CrossRef]
9. Chughtai, F.; Zayed, T. Structural condition models for sewer pipeline. In *Pipelines 2007: Advances and Experiences with Trenchless Pipeline Projects*; American Society of Civil Engineers: Reston, VA, USA, 2007; pp. 1–11.
10. Khan, Z.; Zayed, T.; Moselhi, O. Structural condition assessment of sewer pipelines. *J. Perform. Constr. Facil.* **2010**, *24*, 170–179. [CrossRef]
11. Syachrani, S.; Jeong, H.S.D.; Chung, C.S. Decision tree–based deterioration model for buried wastewater pipelines. *J. Perform. Constr. Facil.* **2013**, *27*, 633–645. [CrossRef]
12. Opila, M.C.; Attoh-Okine, N. Novel approach in pipe condition scoring. *J. Pipeline Syst. Eng. Pract.* **2011**, *2*, 82–90. [CrossRef]
13. Scheidegger, A.; Hug, T.; Rieckermann, J.; Maurer, M. Network condition simulator for benchmarking sewer deterioration models. *Water Res.* **2011**, *45*, 4983–4994. [CrossRef]
14. Rogati, M. *The AI Hierarchy of Needs*; Hacker Noon: Edwards, CO, USA, 2017.
15. Kleindienst, D. The data quality improvement plan: Deciding on choice and sequence of data quality improvements. *Electron. Mark.* **2017**, *27*, 387–398. [CrossRef]
16. Salman, B.; Salem, O. Risk Assessment of Wastewater Collection Lines Using Failure Models and Criticality Ratings. *J. Pipeline Syst. Eng. Pract.* **2012**, *3*, 68–76. [CrossRef]
17. Caradot, N.; Rouault, P.; Clemens, F.; Cherqui, F. Evaluation of uncertainties in sewer condition assessment. *Struct. Infrastruct. Eng.* **2018**, *14*, 264–273. [CrossRef]
18. Loshin, D. *Monitoring Data Quality Performance Using Data Quality Metrics*; Informatica Corporation: Redwood City, CA, USA, 2006.
19. USEPA. *Fact Sheet—Asset Management for Sewer Collection Systems*; Office of Wastewater Management: Washington, DC, USA, 2002.
20. Westin, S.; Sein, M.K. Improving data quality in construction engineering projects: An action design research approach. *J. Manag. Eng.* **2013**, *30*, 05014003. [CrossRef]
21. Van Riel, W.; van Bueren, E.; Langeveld, J.; Herder, P.; Clemens, F. Decision-making for sewer asset management: Theory and practice. *Urban Water J.* **2016**, *13*, 57–68. [CrossRef]
22. Lewis, P.; Khaleghian, H.; Shan, Y. Development of a Sustainable National Sewer Inventory. *Procedia Eng.* **2016**, *145*, 1410–1415. [CrossRef]
23. Pipino, L.L.; Lee, Y.W.; Wang, R.Y. Data quality assessment. *Commun. ACM* **2002**, *45*, 211–218. [CrossRef]
24. Fischer, B.; Hunger, W.; Lehmann, T.M.; Muller, K.; Schafer, T. Objective condition establishment of sewer systems. In Proceedings of the 2nd Conference on Sewer Operation and Maintenance, Vienna, Austria, 26–28 October 2006; Ertl, T., Pressel, A., Kretschmer, F., Haberl, R., Eds.; SIG Eigenverla: Vienna, Austria, 2006.
25. Ertl, T.; Gangl, G.; Bölke, K.P.; Kretschmer, F. Implementing quality management and EN 13508-2 for CCTV sewer inspection in Austria. In Proceedings of the NOVATECH 2007-Sustainable Techniques and Strategies in Urban Water Management 6th International Conference, Lyon, France, 25–28 June 2007.
26. Dirksen, J.; Clemens, F.H.; Korving, H.; Cherqui, F.; Le Gauffre, P.; Ertl, T.; Plihal, H.; Müller, K.; Snaterse, C.T. The consistency of visual sewer inspection data. *Struct. Infrastruct. Eng.* **2013**, *9*, 214–228. [CrossRef]
27. Veregin, H. Data Quality Parameters. In *Geographical Information Systems*; Longley, P.A.G., Michael, F., David, J.M., David, W.R., Eds.; Wiley: New York, NY, USA, 1999; pp. 177–189.
28. Buchheit, R.B. Vacuum: Automated Procedures for Assessing and Cleansing Civil Infrastructure Data. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2002.
29. Wang, R.Y. A product perspective on total data quality management. *Commun. ACM* **1998**, *41*, 58–65. [CrossRef]
30. Nousak, P.; Phelps, R. A Scorecard approach to improving Data Quality. In Proceedings of the SUGI-27: Data Warehousing and Enterprise Solutions, Orlando, FL, USA, 14–17 April 2002.
31. Debattista, J.; Auer, S.; Lange, C. Luzzu—A methodology and framework for linked data quality assessment. *J. Data Inf. Qual.* **2016**, *8*, 1–32. [CrossRef]
32. Ardagna, D.; Cappiello, C.; Samá, W.; Vitali, M. Context-aware data quality assessment for big data. *Future Gener. Comput. Syst.* **2018**, *89*, 548–562. [CrossRef]
33. Pezoulas, V.C.; Kourou, K.D.; Kalatzis, F.; Exarchos, T.P.; Venetsanopoulou, A.; Zampeli, E.; Gandolfo, S.; Skopouli, F.; De Vita, S.; Tzioufas, A.G.; et al. Medical data quality assessment: On the development of an automated framework for medical data curation. *Comput. Biol. Med.* **2019**, *107*, 270–283. [CrossRef]

34. Weiskopf, N.G.; Bakken, S.; Hripcsak, G.; Weng, C. A data quality assessment guideline for electronic health record data reuse. *Egems* **2017**, *5*, 14. [CrossRef] [PubMed]
35. Xia, J. Metrics to measure open geospatial data quality. *Issues Sci. Technol. Librariansh.* **2012**, *68*. Available online: https://journals.library.ualberta.ca/istl/index.php/istl/article/view/1542 (accessed on 21 May 2023). [CrossRef]
36. Batini, C.; Cappiello, C.; Francalanci, C.; Maurino, A. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* **2009**, *41*, 1–52. [CrossRef]
37. Piprani, B.; Ernst, D. A model for data quality assessment. In *On the Move to Meaningful Internet Systems: OTM 2008 Workshops*; Springer: Berlin/Heidelberg, Germany, 2008.
38. National Association of Sewer Service Companies (NASSCO). *Pipeline Assessment Certicication Program (PACP) Reference Manual*; NASSCO: Marriottsville, MD, USA, 2010.
39. Thornhill, R.; Wildbore, P. *Sewer Defect Codes: Origin and Destination*; U-Tech Underground Construction Paper; The Free Library: Philadelphia, PA, USA, 2005.
40. Ehrlinger, L.; Wöß, W. A novel data quality metric for minimality. In Proceedings of the Data Quality and Trust in Big Data: 5th International Workshop, QUAT 2018, Held in Conjunction with WISE 2018, Dubai, United Arab Emirates, 12–15 November 2018; Revised Selected Papers 5. Springer: Berlin/Heidelberg, Germany, 2019.
41. Lebied, M. The Ultimate Guide to Modern Data Quality Management (DQM) for an Effective Data Quality Control Driven by the Right Metrics. 2018. Available online: https://www.datapine.com/blog/data-quality-management-and-metrics (accessed on 21 May 2023).
42. Baskarada, S.; Gao, J.; Lin, S.; Yeoh, G.; Koronios, A. Data Quality Enhancing Software for Asset Management-State of the Art Evaluation. In Proceedings of the 4th WSEAS International Conference on Applied Mathematics & Computer Science (AMCOS 2005), Rio de Janeiro, Brazil, 25 April 2005.
43. Van Riel, W.; Stanić, N.; Langeveld, J.; Clemens, F. Pipe quality information in sewer asset management: Use and uncertainties. In Proceedings of the 9th World Congress on Engineering Asset Managemen, Pretoria, South Africa, 28–31 October 2014.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.