Pavement Condition Data Integrity Quality Assurance Using an Information Management Approach

Ghim Ping ONG
Lecturer
Department of Civil and Environmental Engineering
National University of Singapore
1 Engineering Drive 2
Singapore 117576, SINGAPORE
Fax: +65-6779-1635
E-mail: ceeongr@nus.edu.sg

Kumares C. SINHA
Olson Distinguished Professor
School of Civil Engineering
Purdue University
550 Stadium Mall Drive
West Lafayette, IN 47907, USA
Fax: +1-765-496-7996
E-mail: sinha@ecn.purdue.edu

Abstract: Highway agencies are collecting an enormous amount of pavement condition data and have to face constantly the laborious task of ensuring its completeness and accuracy. Traditionally, most data-quality studies are focused on ensuring the accuracy and precision of delivered pavement condition data and little attention has been paid to ensure that the data is of high integrity. This paper proposes the use of information management concepts to investigate the integrity of a delivered pavement condition database. A set of performance metrics is developed to assess data integrity from both macroscopic and microscopic perspectives. Using these metrics, a quality assurance framework is proposed to test for data integrity before accepting the database for use in the pavement management system. Through the use of a case study, the applicability of the developed quality assurance procedure is demonstrated.

Key Words: Pavement Condition Data, Data Quality, Data Integrity, Quality Assurance

1. INTRODUCTION

Many state highway agencies are collecting and managing an enormous amount of pavement condition data that is collected via automated techniques. Despite the laborious nature of ensuring that the collected data is complete and accurate, agencies recognize the consequences of poor pavement data quality (incomplete or having poor accuracy). Based on a recent survey on automated pavement condition data collection practices among state highway agencies, McGhee (2004) found that most agencies did not possess adequate experience on automated condition data quality control and assurance. This is despite the fact that these agencies indicated the importance of data quality during data collection and processing. In another recent study, Jordan et al. (2009) summarized current quality management practices and found that some agencies have started to develop methods and practices to ensure the appropriateness of the collected data for use in their respective pavement management system. However, existing practices are far from universal and there is a lack of a formalized data collection quality management plan within the agencies.

The lack of data quality control can be addressed from two perspectives: data accuracy/precision and data integrity (TRB, 2009). Data accuracy and precision refer to effective representation of the collected data of the actual pavement condition in the field. Data integrity refers to the collected database being complete and free-of-error when fed into
the pavement management system. A high-integrity database ensures that there is no missing data that could potentially jeopardize the optimality of the decisions made at the project and network levels. Numerous research studies were conducted to evaluate pavement condition data quality in terms of accuracy and precision, and quality assurance plans were developed to ensure that high quality data is collected and processed (Stoeffels et al., 2001; Morian et al., 2002; Ong et al., 2009). However, limited attention has been given to data integrity (Larson et al., 2000; Wolters et al., 2006; Shekharan et al. 2007). Wolters et al. (2006) proposed the use of preliminary checks [which consists of division check (a division operation with its denominator having a value of zero) and field type check (entity with an incorrect value type – integer, floating number, or character)] to test for the completeness of the database. Similarly, Shekharan et al. (2007) proposed a data completeness check within the “independent verification and validation” module for the Virginia Department of Transportation’s quality management system.

Recognizing the need to have a systematic approach for data integrity quality assurance (especially before importing collected data into the pavement management system), this paper proposes a method to assess the integrity of the delivered pavement condition data. A set of performance metrics is developed and a quality assurance framework proposed to test for the integrity of the pavement condition database. A case study is presented to illustrate the applicability of the developed framework.

2. ROLE OF INFORMATION MANAGEMENT IN DATA INTEGRITY QUALITY ASSURANCE

A typical state agency data collection process in a given data collection season yields millions of data entities, as summarized in Figure 1 (Ong et al., 2009). A typical database consists of many sub-databases containing information about the location of a highway segment, its roadway inventory, pavement condition, and others. In terms of pavement condition data, there could be different data collection sources. A straightforward differentiation would be data collected by contract and data collected in-house. From the figure, it can be observed that each data table contains:

- A set of primary keys (rows) which are unique identifiers that link each table to the other tables in a relational database. Typically, this identifier refers to a specific segment.
- A matrix containing the foreign keys which describe the characteristics of a primary key (row). In pavement applications, the characteristics could include location references, roadway classification, inventory information, pavement condition indicators, pavement structure data and others.
- Headers describing each column (or indicator describing a given characteristic).
- Data entities which provide a value for each cell (i.e. a value corresponding to an indicator for a given characteristic and roadway segment).

In order to ensure informed decisions are made, quality control and assurance has to be employed on each core elements of the pavement management database (including location, inventory, condition, design, traffic, weather, work history and programmed work sub-databases) shown in Figure 1. It is further recognized that existing pavement infrastructure condition is a key in the decision-making process, ensuring that money is spent on where improvements are needed the most. Yet, little effort is made in practice in ensuring the quality of the delivered data from the highway agency perspectives, hence vindicating the need for a
quality assurance program for pavement condition data collection. Evaluating the quality of the delivered data is undoubtedly a tedious task for any highway agency. To this end, concepts in information and database management can be applied. Information management (IM) is the collection and management of information from one or more sources and the distribution of that information to one or more audiences, who have a stake in or a right to that information (Data Management Association, 2007). Of particular relevance to pavement condition data quality are concepts based in the sub-area of data quality management.

![Figure 1: Overview of data found in a pavement management system (Ong et al., 2009)](image)

Data quality management, from an information management perspective, involves the checking of:

- **Data integrity**: A term used in computer science and telecommunications that can mean
ensuring data is "whole" or complete. Data integrity is normally enforced in a database system by a series of integrity constraints or rules. In terms of data integrity, three types of integrity constraints (known as the Codd integrity constraints) are an inherent part of the relational data model: entity integrity, referential integrity and domain integrity as summarized in Table 1. Based on the definitions given in Table 1, the rating can be evaluated using the following form:

\[
\text{Rating} = 1 - \left( \frac{\text{Number of Undesirable Outcomes}}{\text{Total Outcomes}} \right)
\]

A rating of 1 represents the most desirable condition and 0 is the least desirable score.

- Data cleansing: The act of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Incomplete, incorrect, inaccurate and irrelevant parts of the data is identified and then replaced, modified or deleted (i.e. the act of cleansing).

The data quality management concept is useful in investigating the integrity of the delivered data from vendors or data collector and in quality assurance.

### Table 1: Definitions for Codd Integrity (Codd, 1970)

<table>
<thead>
<tr>
<th>Aspects of Codd Integrity Constraints</th>
<th>Explanation</th>
<th>Mathematical Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity Integrity</td>
<td>Every table must have a primary key and each primary key must be unique and not null.</td>
<td>[ C_1 = 1 - \left( \frac{n_1}{n_2} \right) ] where ( C_1 ) is the entity integrity rating, ( n_1 ) is the number of null or non-unique primary keys, and ( n_2 ) is the total number of rows in the pavement condition database.</td>
</tr>
<tr>
<td>Referential Integrity</td>
<td>Relationship between the primary key (in a referenced table) and the foreign key (in each of the referencing tables) must be maintained.</td>
<td>[ C_2 = 1 - \left( \frac{n_3}{n_4} \right) ] where ( C_2 ) is the referential integrity rating, ( n_3 ) is the number of non-matching values excluding nulls in the dependent table, and ( n_4 ) is the total number of rows in the dependent table.</td>
</tr>
<tr>
<td>Column (Domain) Integrity</td>
<td>Values in columns must be drawn from a set of permissible values.</td>
<td>[ C_3 = 1 - \left( \frac{n_5}{n_2} \right) ] where ( C_3 ) is the column integrity rating, ( n_5 ) is the number of invalid column values excluding nulls in the dependent table and ( n_2 ) is the total number of rows in the table.</td>
</tr>
</tbody>
</table>

### 3. FRAMEWORK FOR PAVEMENT CONDITION DATA INTEGRITY EVALUATION AND QUALITY ASSURANCE

While the concept of data quality management has been applied frequently in organizations and businesses (Wang et al., 1993; Wang and Wang, 1996), its formalized use in pavement management is rare. Past research tends to focus on developing a series of integrity checks to ensure the data is of sufficient quality (Larson et al., 2000; Shekharan et al. 2007), which do not allow agencies to explicitly evaluate data integrity, nor are there any tools that allow agencies to propose quality assurance protocols to ensure that the delivered data is of the correct format and is complete. Therefore, this section shall discuss a systematic approach of evaluating pavement condition data integrity and quality assurance.

#### 3.1 Metrics for Evaluating Data Integrity

A set of quantifiable metrics for data integrity is developed in this paper. It is noted that these
metrics can be used for any data management system but is applied in this paper in the area of pavement condition data collection and processing.

3.1.1 Codd Integrity Constraints for Macroscopic Evaluation of Data Integrity

The Codd integrity constraints can be used to determine the quality of the data of such information system. It involves directly evaluating the entity, referential and domain integrities as described in Table 1 (Codd, 1970). These metrics allow a quick macroscopic view of the integrity of the delivered data without specifically investigating any part of dataset that is in error.

3.1.2 Quality Metrics for Microscopic Evaluation of Data Integrity

The Codd integrity constraints provide a quick evaluation of data integrity, but do not allow troubleshooting when there is indeed a problem with the delivered dataset. In this case there is a need for a microscopic evaluation of data integrity. Wang and Wang (1996) have proposed several additional dimensions to evaluate the completeness of the delivered data (shown in Table 2). These dimensions include: free-of-error, completeness (schema completeness, column completeness, and population completeness) and consistency. These metrics allow a much more detailed interpretation of data integrity compared to the Codd integrity constraints, which is essentially a quick identifier for defects.

Table 2: Definitions for Free-of-Error, Completeness and Consistency Ratings (Wang and Wang, 1996)

<table>
<thead>
<tr>
<th>Aspects of Data Integrity</th>
<th>Explanation</th>
<th>Mathematical Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free-of-Error</td>
<td>Values in a given column are error-free.</td>
<td>Free-of-Error Rating $= 1 - \frac{\text{Number of Data Units in Error}}{\text{Total Number of Data Units}}$ where error occurs when (a) the field type is incorrect; (b) the field format is incorrect, and (c) the value of the data does not satisfy the specified range.</td>
</tr>
<tr>
<td>Schema Completeness</td>
<td>Attributes are not missing from the schema.</td>
<td>Completeness Rating $= 1 - \frac{\text{Number of Incomplete Headers}}{\text{Total Number of Headers}}$ where schema completeness rating is defined for every column in the data table.</td>
</tr>
<tr>
<td>Column Completeness</td>
<td>Degree to which there exist missing values in a column of a table.</td>
<td>Completeness Rating $= 1 - \frac{\text{Number of Incomplete Rows}}{\text{Total Number of Rows}}$ where column completeness rating is defined for every column in the data table.</td>
</tr>
<tr>
<td>Population Completeness</td>
<td>Degree to which members of the population that should not be present.</td>
<td>Completeness Rating $= 1 - \frac{\text{No. of Incomplete Items}}{\text{Total No. of Items}}$ where population completeness rating is defined for every column in the data table.</td>
</tr>
<tr>
<td>Consistency</td>
<td>Degree of occurrence of redundant data in a table. Redundancy means that there are at least two set of data which are exactly the same (i.e. non-unique).</td>
<td>Consistency Rating $= 1 - \frac{\text{No. of Inconsistent Items}}{\text{Total No. of Items}}$ where consistency rating is defined for every column in the data table.</td>
</tr>
</tbody>
</table>
3.2 Framework for Data Integrity Quality Assurance

Given the metrics for evaluating pavement condition data integrity, a systematic framework for quality assurance protocols is developed (Figure 2) for the pavement condition database presented in Figure 1. Similar approach may be adopted and modified for other sub-elements of the pavement management database presented in Figure 1. The steps involved are:

Step 0: Select $n$ random samples from the delivered database.

Step 1: Obtain delivered data from data collector (either by contract or from in-house sources)

Step 2: Given delivered data, perform the two-stage data quality checks for integrity and completeness:

(a) Perform the Codd’s Integrity Constraints checks to test for entity integrity, column integrity and referential integrity defined in Table 1. Data is accepted only if the ratings exceed a set of specified thresholds. If any error found in this stage can be corrected, they should be corrected or deleted at this stage.

(b) Perform the detailed quality checks to test for errors, completeness and consistency defined in Table 2. The delivered data is accepted only if the ratings exceed the set of minimum thresholds. If any error found in this stage, they should be checked to determine if they can be corrected or deleted.

Step 3: Delivered data has passed quality assurance checks and are imported into the PMS database. If the quality assurance checks are not passed, vendors will have to recollect additional data to complete the database. Further quality checks for accuracy and relationships with other elements of the PMS database could be implemented from this point.

The above-stated steps provided agencies with a tool to check if the delivered data is of sufficient quality for pavement management applications. It is noted that two elements are important in the proposed QA procedure. They are namely: sample size selection and determination of threshold for acceptance. The following discussion details how these two variables are determined for practical applications.

3.3 Selection of Sample Size and Acceptance Threshold

3.3.1 Determination of Sample Size

Typically, the standard random sampling technique is employed in data quality management (Lee et al., 2006). Using a random number generator, a random sample of the required size can be drawn. If the database table contains $N$ primary keys (or rows), and the required sample size is $n$, we can make use of Eq. (2) to determine the sample size:

$$n = \frac{z_{\alpha/2}^2 p (1 - p)}{\varepsilon^2}$$

(2)

where $z_{\alpha/2}$ is the two-tailed value of a standard normal variable, $\alpha$ is the confidence level, $p$ is the proportion of defects, and $\varepsilon$ is the desired precision or acceptable error. Eq. (2) serves well only if the true proportion of defects $p$ is between 0.3 and 0.7 (which obviously is undesirable for any database) and is not recommended for most practical database management purposes. Note that in this case $q = (1 - p)$ is analogous to the various ratings shown in Tables 1 and 2.

In cases where we expect the number of defective records to be fairly low (i.e. $p$ is typically below 0.01), it becomes more desirable to set the expected number of defective records at a specified number. Gitlow et al. (1989) suggested setting the expected number of defective records in the sample to be at least 2. The minimum sample size should satisfy the inequality...
shown in Eq. (3) for a given check $j$, a column attribute $i$ and a defective rate $p_{ij}$:

$$n_j = \frac{2}{p_{ij}}$$

where $(n_{ij})_{\text{min}}$ is the estimated minimum sample size and $p_{ij}$, the defective rate, can be estimated from historical data or past experience. In this case, we can determine the overall sample size $n$ required for the entire database using Eq. (4).

$$n = \max \left\{ \left( n_j \right)_{\text{min}} \right\} \quad \forall \, i, j$$

---

**Figure 2: Framework for data integrity quality assurance**

---
3.3.2 Determination of Acceptance Threshold

Knowing that it is desirable to obtain \( n \) sample rows for quality assurance, it is then necessary to determine the threshold rating to accept the set of data during the quality assurance process. In cases of entity and reference-related ratings, it is necessary for the acceptance threshold to be set to one. This is necessary to prevent any logic error or computational error in the entire database that can complicate the pavement management process. For the other parameters, the following hypothesis test can be performed:

\[
\begin{align*}
H_0 & : q = q_0 \\
H_1 & : q > q_0
\end{align*}
\]  

(5)

where \( q = (1 - p) \) is the proportion of non-defects (and is equal to rating for a given check) and \( q_0 \) is the desired proportion of non-defects (i.e. desired rating for a given check). The \( z \)-statistic of the hypothesis test is:

\[
z_\alpha = \frac{q_0 - q}{\sqrt{\frac{q(1-q)}{n}}}
\]  

(6)

where \( \alpha \) is the level of significance and \( n \) is the sample size. Given the sample size \( n \), a selected significance level, and \( a-priori \) knowledge of the ratings, the minimum threshold ratings can be determined:

\[
q_0 = q - \frac{z_\alpha q(1-q)}{n}
\]  

(7)

The derived set of threshold criteria can then be used for acceptance during quality assurance.

4. APPLICATION OF DEVELOPED FRAMEWORK TO INDIANA PAVEMENT MANAGEMENT SYSTEM

The proposed framework for pavement condition data integrity evaluation and quality assurance is applied in the context of the Indiana pavement management system. A framework similar to that illustrated in Figure 1 is adopted in Indiana. A vendor is contracted to perform the automated pavement condition data collection and collects information on the roughness, rutting, faulting and other pavement surface distresses. In addition, the agency also performs pavement roughness testing, friction testing and falling weight deflection tests on the network. An enormous amount of data is collected on more than 17,000 km of highways managed by the Indiana Department of Transportation (INDOT) annually. The agency also has the onus to continually ensure that the data is complete and is of high integrity before importing them into the PMS database.

In this paper, the pavement condition summary table in the PMS database is used (see Figure 1). This database consists of the roughness data (in terms of international roughness index IRI), rut depth data, fault depth data, location references and other general information related to the pavement segment collected by the vendor. Table 3 provides the definitions of each variable, the field data type and format associated with that variable in the pavement condition summary table. In particular, the variable “Element ID” serves as a primary key (or identifier) to a pavement segment in a given year and serves as the connecting variable relating the table to other tables in pavement management database.
Table 3: Delivered Data for Use in Pavement Condition Summary Table

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Field Type (Size/Format)</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Element ID</td>
<td>String</td>
<td>ID of Pavement Segment (Primary Key)</td>
</tr>
<tr>
<td>D</td>
<td>Integer (3)</td>
<td>District</td>
</tr>
<tr>
<td>POST</td>
<td>Floating (3.1)</td>
<td>Reference Post</td>
</tr>
<tr>
<td>T</td>
<td>Alphanumeric (1)</td>
<td>Route Type (Interstate, US, State)</td>
</tr>
<tr>
<td>RTE</td>
<td>Alphanumeric (5)</td>
<td>Route Number</td>
</tr>
<tr>
<td>DIR</td>
<td>Alphanumeric (1)</td>
<td>Direction Of Rating</td>
</tr>
<tr>
<td>PVMT</td>
<td>Alphanumeric (5)</td>
<td>Pavement Type</td>
</tr>
<tr>
<td>DATE (Month/Day/Year)</td>
<td>Alphanumeric (5)</td>
<td>Date Rating Made</td>
</tr>
<tr>
<td>YEAR</td>
<td>Alphanumeric (4)</td>
<td>Year Rating Made</td>
</tr>
<tr>
<td><strong>Roughness Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRI_LWP</td>
<td>Floating (3.1)</td>
<td>IRI value for the left wheel path (inch/mile)</td>
</tr>
<tr>
<td>IRI_RWP</td>
<td>Floating (3.1)</td>
<td>IRI value for the right wheel path (inch/mile)</td>
</tr>
<tr>
<td>IRI_AVE</td>
<td>Floating (3.1)</td>
<td>Average IRI value for both wheel paths (inch/mile)</td>
</tr>
<tr>
<td><strong>Pavement Surface Distress Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEDCT</td>
<td>Floating (3.1)</td>
<td>Total Number Of Deductions</td>
</tr>
<tr>
<td>PCR</td>
<td>Floating (3.1)</td>
<td>Pavement Condition Rating</td>
</tr>
<tr>
<td><strong>Rut Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUT_LWP</td>
<td>Floating (1.2)</td>
<td>Rut depth for the left wheel path (inch)</td>
</tr>
<tr>
<td>RUT_RWP</td>
<td>Floating (1.2)</td>
<td>Rut depth for the right wheel path (inch)</td>
</tr>
<tr>
<td>RUT_AVE</td>
<td>Floating (1.2)</td>
<td>Average rut depth for both wheel paths (inch)</td>
</tr>
<tr>
<td><strong>Fault Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAULT</td>
<td>Floating (1.2)</td>
<td>Fault depth (inch)</td>
</tr>
</tbody>
</table>

4.1 Use of Metrics to Evaluate Pavement Condition Data Integrity

The potential of the proposed metrics in evaluating integrity is demonstrated in the following analyses. The Codd integrity constraints are first evaluated for the historical pavement condition data collected in Indiana for a period between 1998 and 2005. Figure 3(a) shows the historical Codd’s integrity ratings for selected parameters (international roughness index IRI, rut depth, fault depth and pavement condition rating PCR) while Figure 3(b) shows the historical free-of-error ratings between 1998 and 2005. The set of data is collected by the same vendor engaged by INDOT since 1998. For the Codd integrity constraint (macroscopic) checks, it is observed that:
- The average historical entity integrity rating 1.0, indicating that there is no primary key redundancy.
- The average historical referential integrity rating is 1.0, indicating that the attributes (foreign key items) are properly referenced to a primary key in the table.
- The average historical column integrity rating is 0.648. This is interesting as it means that 
  \[ 1 - 0.648 = 0.352 \] (or 35.2%) of the table contains null items. This can be further investigated in the detailed quality checks.

Table 4 shows the results of the detailed data integrity checks. It can be seen from Table 4 that:
- The average historical free-of-error, column completeness and population completeness ratings are found to be one for the primary key. This is because of the non-redundancy in primary keys and the fact that all segments have some form of pavement condition data collected in a single year.
The average historical free-of-error and column completeness ratings are 0.998 for IRI, rut depth and fault depth while the population completeness ratings are one for these attributes. This indicates that all segments have some form of automated data being collected but there is a 0.002 defect rate where the data does not satisfy the error criteria.

The average historical free-of-error and column completeness ratings are 0.648 for PCR while its population completion rating is 0.649. The significant difference between surface distress data and the other data is due to the sampling procedures for automated pavement surface distress data collection. Approximately two-thirds of the network is assessed for surface distress annually, whose value is similar to the historical population completeness rating. The small difference between the free-of-error, column completeness and the population completeness ratings are due to a 0.002 defect rate where the PCR data does not satisfy the error criteria.

The historical schema completeness ratings are 1.0, indicating that there is no column or data variables missing in the databases.
The historical consistency ratings are 1.0. This is consistent with the historical entity and referential integrity ratings in the earlier analysis which found that there is no redundancy in the data.

Based on the above discussions, it was found that the proposed set of metrics is capable to analyze data integrity from different perspectives. These metrics provide a rational basis to develop a quality assurance procedure for ensuring data integrity.

Table 4: Determination of Required Sample Size and Acceptance Threshold for Quality Assurance

<table>
<thead>
<tr>
<th>Data Integrity Rating</th>
<th>Average Historical Value</th>
<th>Required Sample Size</th>
<th>Acceptance Threshold b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity Integrity Rating</td>
<td>1</td>
<td>Not Applicable</td>
<td>1</td>
</tr>
<tr>
<td>Referential Integrity Rating</td>
<td>1</td>
<td>Not Applicable</td>
<td>1</td>
</tr>
<tr>
<td>Column Integrity Rating a</td>
<td>0.648</td>
<td>250</td>
<td>0.647</td>
</tr>
<tr>
<td>Free-of-Error Rating for Element ID (Primary Key)</td>
<td>1</td>
<td>Not Applicable</td>
<td>1</td>
</tr>
<tr>
<td>Free-of-Error Rating for Average IRI</td>
<td>0.998</td>
<td>2000</td>
<td>0.997</td>
</tr>
<tr>
<td>Free-of-Error Rating for Average Rut Depth</td>
<td>0.998</td>
<td>2000</td>
<td>0.997</td>
</tr>
<tr>
<td>Free-of-Error Rating for Average Fault Depth</td>
<td>0.998</td>
<td>2000</td>
<td>0.997</td>
</tr>
<tr>
<td>Free-of-Error Rating for Pavement Condition Rating a</td>
<td>0.648</td>
<td>250</td>
<td>0.647</td>
</tr>
<tr>
<td>Schema Completeness Rating</td>
<td>1</td>
<td>Not Applicable</td>
<td>1</td>
</tr>
<tr>
<td>Column Completeness Rating for Element ID (Primary Key)</td>
<td>1</td>
<td>Not Applicable</td>
<td>1</td>
</tr>
<tr>
<td>Column Completeness Rating for Average IRI</td>
<td>0.998</td>
<td>2000</td>
<td>0.997</td>
</tr>
<tr>
<td>Column Completeness Rating for Average Rut Depth</td>
<td>0.998</td>
<td>2000</td>
<td>0.997</td>
</tr>
<tr>
<td>Column Completeness Rating for Average Fault Depth</td>
<td>0.998</td>
<td>2000</td>
<td>0.997</td>
</tr>
<tr>
<td>Column Completeness Rating for Pavement Condition Rating a</td>
<td>0.648</td>
<td>250</td>
<td>0.647</td>
</tr>
<tr>
<td>Population Completeness Rating for Element ID (Primary Key)</td>
<td>1</td>
<td>Not Applicable</td>
<td>1</td>
</tr>
<tr>
<td>Population Completeness Rating for Average IRI</td>
<td>1</td>
<td>Not Applicable</td>
<td>1</td>
</tr>
<tr>
<td>Population Completeness Rating for Average Rut Depth</td>
<td>1</td>
<td>Not Applicable</td>
<td>1</td>
</tr>
<tr>
<td>Population Completeness Rating for Average Fault Depth</td>
<td>1</td>
<td>Not Applicable</td>
<td>1</td>
</tr>
<tr>
<td>Population Completeness Rating for Pavement Condition Rating a</td>
<td>0.649</td>
<td>250</td>
<td>0.648</td>
</tr>
<tr>
<td>Consistency Rating for Redundant Data</td>
<td>1</td>
<td>Not Applicable</td>
<td>1</td>
</tr>
</tbody>
</table>

a Sample size is determined using Eq. (1) with an error of 5% and confidence level of 95%.
b Acceptance threshold is determined using Eq. (10) with a 95% confidence level and a sample size of 2000.

4.2 Application of Metrics in Data Integrity Quality Assurance

Using the framework described in Figure 2, the set of developed metrics is next applied in data integrity quality assurance.

Step 0: Determination of Sample Size and Quality Acceptance Threshold

Before the actual quality assurance process, it is important to pre-determine the required sample size and quality acceptance threshold. From the previous section, it is noted that some form of prior knowledge of the past ratings is required. In this paper, data collected over a period of nine years (1998 to 2005) are used to obtain some form of prior knowledge of the historical ratings.

Using the average historical rating evaluated in the previous section, the minimum sample size required can be calculated for each rating and is shown in Table 4. It can be observed that
at least 2000 samples are required to accurately predict the proportion (or the ratings) for each check. Using Eq. (4), it is clear that 2000 rows (or primary keys) are required. Considering that on average, an approximate number of 14,000 primary keys are collected annually, the sampling rate is about 15% of the entire population.

Given that the selected sample size is 2000, the acceptance threshold for each rating can be evaluated. Table 4 also computes the thresholds for each rating. It is necessary for entity integrity, referential integrity, schema completeness and consistency ratings to be one to ensure that (i) there is no missing highway segment in the pavement management database, (ii) there is no redundancy in primary key (i.e. there is no “double-counting” of the same highway segment within the same table) and (iii) all foreign-key references to other tables in the pavement management database are error-free. For the same reasons, any primary-key-related free-of-error and population completeness ratings must be one. For the other ratings, the acceptance threshold rating can be determined using Eq. (7).

Table 5: Application of Framework on Pavement Condition Data Collected in the 2006 Data Collection Cycle

<table>
<thead>
<tr>
<th>Data Integrity Rating</th>
<th>2006 Value</th>
<th>Acceptance Threshold</th>
<th>Quality Assurance Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity Integrity Rating</td>
<td>1</td>
<td>1</td>
<td>Passed</td>
</tr>
<tr>
<td>Referential Integrity Rating</td>
<td>1</td>
<td>1</td>
<td>Passed</td>
</tr>
<tr>
<td>Column Integrity Rating</td>
<td>0.662</td>
<td>0.647</td>
<td>Passed</td>
</tr>
<tr>
<td>Free-of-Error Rating for Element ID (Primary Key)</td>
<td>1</td>
<td>1</td>
<td>Passed</td>
</tr>
<tr>
<td>Free-of-Error Rating for Average IRI</td>
<td>1</td>
<td>0.997</td>
<td>Passed</td>
</tr>
<tr>
<td>Free-of-Error Rating for Average Rut Depth</td>
<td>1</td>
<td>0.997</td>
<td>Passed</td>
</tr>
<tr>
<td>Free-of-Error Rating for Average Fault Depth</td>
<td>1</td>
<td>0.997</td>
<td>Passed</td>
</tr>
<tr>
<td>Free-of-Error Rating for Pavement Condition Rating</td>
<td>0.662</td>
<td>0.647</td>
<td>Passed</td>
</tr>
<tr>
<td>Schema Completeness Rating</td>
<td>1</td>
<td>1</td>
<td>Passed</td>
</tr>
<tr>
<td>Column Completeness Rating for Element ID (Primary Key)</td>
<td>1</td>
<td>1</td>
<td>Passed</td>
</tr>
<tr>
<td>Column Completeness Rating for Average IRI</td>
<td>1</td>
<td>0.997</td>
<td>Passed</td>
</tr>
<tr>
<td>Column Completeness Rating for Average Rut Depth</td>
<td>1</td>
<td>0.997</td>
<td>Passed</td>
</tr>
<tr>
<td>Column Completeness Rating for Average Fault Depth</td>
<td>1</td>
<td>0.997</td>
<td>Passed</td>
</tr>
<tr>
<td>Column Completeness Rating for Pavement Condition Rating</td>
<td>1</td>
<td>0.647</td>
<td>Passed</td>
</tr>
<tr>
<td>Population Completeness Rating for Element ID (Primary Key)</td>
<td>1</td>
<td>1</td>
<td>Passed</td>
</tr>
<tr>
<td>Population Completeness Rating for Average IRI</td>
<td>1</td>
<td>1</td>
<td>Passed</td>
</tr>
<tr>
<td>Population Completeness Rating for Average Rut Depth</td>
<td>1</td>
<td>1</td>
<td>Passed</td>
</tr>
<tr>
<td>Population Completeness Rating for Average Fault Depth</td>
<td>1</td>
<td>1</td>
<td>Passed</td>
</tr>
<tr>
<td>Population Completeness Rating for Pavement Condition Rating</td>
<td>0.662</td>
<td>0.648</td>
<td>Passed</td>
</tr>
<tr>
<td>Consistency Rating for Redundant Data</td>
<td>1</td>
<td>1</td>
<td>Passed</td>
</tr>
</tbody>
</table>

Step 1: Data Delivery from Data Collector

Once the required sample size and acceptance thresholds are obtained, the delivered pavement condition data in any given year can be evaluated and is subjected to the proposed quality
assurance checks. In this paper the pavement condition data collected in the 2006 data collection cycle is used to demonstrate the applicability of the quality assurance checks.

Step 2: Quality Assurance Checks

The two-step quality assurance checks are performed to test for Codd integrity constraints and for error, completeness and consistency. Table 5 show the test results for the data collected in the 2006 data collection cycle. It can be observed that the delivered data from the vendor has passed the required quality assurance checks from all aspects of data integrity. In cases where the rating is not equal to one, further examination of the data is required. In the illustration shown in Table 5, it is noted that the column integrity rating is 0.662. Upon further evaluating the free of error ratings and population completeness ratings, it is found that this value of 0.662 is attributed to the sampling procedures of the pavement surface distress where only two-third of the network is sampled in that year. It was found that there are no further errors (e.g. wrong format or field type) that warrant additional data-cleansing.

Step 3: Import to PMS Database

The delivered data that has passed the quality assurance checks can be accepted by the agency and can be either (i) further tested for accuracy, precision and relationships with other data tables in the pavement management database, or (ii) import to the pavement management system for use.

5. CONCLUSIONS

This paper has presented a systematic and rational approach to evaluate data integrity and to perform data integrity quality assurance. The proposed methodology is applied to automated pavement condition data collection within a pavement management framework. A set of metrics is developed to measure data integrity and it is found to be capable of evaluating data integrity from both macroscopic and microscopic perspectives. The set of metrics is then incorporated in a quality assurance framework developed in the paper to test for data integrity. The applicability of the developed quality assurance procedures is demonstrated in the paper through a case study. It is found that the proposed quality assurance framework can ensure that the pavement condition data is of high integrity, free-of-error, complete and consistent before importing data into the pavement management system. It is foreseen that the implementation of this framework will improve data quality and enhance the efficiency and effectiveness of pavement management decisions.

ACKNOWLEDGMENTS AND DISCLAIMER

This study was conducted under the Joint Transportation Research Program with the support of the Federal Highway Administration and the Indiana Department of Transportation. The contents of this paper, however, reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented. The contents do not necessarily reflect the official views or policies of the Indiana Department of Transportation or the Federal Highway Administration. This paper does not constitute a standard, specification, or regulation.
REFERENCES