Developing a Pattern Model of Damage Types on Bridge Elements Using Big Data Analytics

S. Lim[a*], S. Chung[a], S. Chi[a]

[a]Department of Civil and Environmental Engineering, Seoul National University, Republic of Korea

E-mail: sorami@snu.ac.kr*, hwani751@snu.ac.kr, shchi@snu.ac.kr

Abstract

The number of over 30–year-old bridge structures has increased rapidly in Korea. Due to the lack of maintenance budget and professional inspectors, the demands for more effective and cost efficient bridge condition monitoring solutions have increased. The primary purpose of this study is to develop a model using big data analytics to recognize bridge damage patterns that show the relationships between bridge-related variables and damage types on different bridge elements. This research covered the total of 6,773 bridges in Korea and analyzed Bridge Management System (BMS) data with weather and contractor-related variables brought from the outside of the BMS database. After preprocessing, key predictors (i.e., independent variables) were selected by the association rule discovery algorithm and then damage patterns were extracted by decision tree. The pilot study results with the data originated from three cities in Korea, Ulju-gun, Inje-gun, and Mungyeong-si, showed that different predictors derived by region, and the extracted patterns implied geographical characteristics such as heavy snow and different construction capacities of contractors. The derived patterns were expected to give bridge inspectors prior information about the primary inspection area.

Keywords

Bridge Management, Infrastructure Maintenance, Damage Pattern, Big Data analytics, O&M

1 Introduction

Bridge plays a significant role in public transportation networks but its damage can threat public safety and hinder economic activity. Timely bridge management is needed not only to keep operating services but also to ensure traffic safety. Since a number of bridges were built with rapid economic growth all over the world, nowadays economically developed countries such as the U.S. and Korea are suffering from the substantial number of over 30–year-old bridge structures, the aged structures. The average age of the U.S.'s 607,380 bridges was 42 years old in 2016 [1], and the aged bridges in Korea is expected to increase up to over three times in 2025 from 3,094 in 2015 [2].

The Korean government legislated on the Special Act on Safety Control for Infrastructure in 1995 after the collapse of Seongsu Bridge in Seoul in 1994. The aim of the Act is to enforce periodic inspection of major infrastructure facilities to provide on-time repairs according to the infrastructures' conditions graded by "A" (excellent) to "E" (poor). The grades of "C", "D" and "E" indicate that the structure needs to get repaired. Such visual periodic inspection is manually performed every six month, and further detailed inspection and precise safety diagnosis including structural analysis and safety assessment are performed every two to six years according to the safety grades. However, due to the limited number of budget, time, and professional inspectors, the quality of such periodic inspection becomes less objective and not guaranteed.

To support the limitation of the manual inspections, researchers have developed condition monitoring approaches of bridge structures. One of traditional real-time monitoring methods is sensor-based structural health monitoring (SHM). This method keeps monitoring sensing data obtained from the attached sensors of the bridge and detects outliers such as displacement captured from strain gage sensors while explaining abnormal conditions. The Golden Gate Bridge in 2006 and Jindo Bridge in Korea are widely known as application examples of such sensor-based approaches [1]. Other researchers investigated a bridge condition analysis model by analyzing bridge inspection data. They normally focused on the development of bridge condition deterioration models by applying regression and Markov models [3-4].

Although the previous research showed potential benefits for preventive maintenance, they failed to explain detailed mechanism for damage causation of individual bridge elements. Moreover, the sensor-based

monitoring approaches have difficulties from detecting damages that do not cause structural movement changes such as concrete delamination and clogged drains. Even though a main element of the bridge is broken, it sometimes cannot be detected since the structural is normally designed with safety margins. Additionally, regression and Markov models are normally hard to handle multidimensional data [4] and thus bridge deterioration models utilize a limited number of both independent and dependent variables.

Thus, this study aims to develop a model to explore bridge damage patterns for specific bridge elements. A range of factors containing both structural (e.g., thickness of bridge deck) and non-structural information (e.g., traffic volume, weather, and constructor) were considered as independent variables in the model, and damages that do not cause abnormal structural movements were also considered as dependent variables. Once inspectors do their inspection, they can recognize the primary inspection area based on the derived patterns by matching independent variables in the targeted bridge information with those in the extracted patterns.

The model analyzed data stored in Bridge Management System (BMS) developed by the Korea Institute of Civil Engineering and Building Technology (KICT), and big data algorithms including association rules discovery and decision trees in classification were applied to extract patterns of bridge damages.

This paper introduces the framework to develop the pattern recognition model and pilot test results.

2 Research Methodology

2.1 Research Framework

Figure 1 illustrates the model development framework using big data analytics.

After collecting data, data passed through the preprocessing steps including data reduction, data cleaning, data transformation, and data integration. The preprocessed data was then divided by predominant features such as cities. Due to a large number of variables, the data dimensionality was reduced with the execution of feature selection and the damage patterns were finally extracted by the selected features. Cross validation was applied to validate the findings. The framework was developed and implemented by R software version 3.3.2 for statistical computing and graphics.

2.2 Data Collection

In addition to inspection data obtained from the total of 6,773 bridges in Korea, stored in BMS, weather data and contractor-related information were collected from the websites of the Korean Meteorological Administrat-

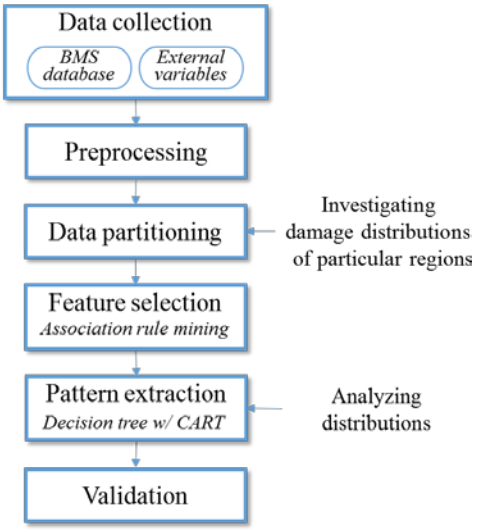

Figure 1. Model Development Framework

ration (KMA) and the Construction Association of Korea (CAK) respectively. KMA published annual weather reports from 1992 to 2015 and the reports covered only 78 regional measurements. Such weather information of the 78 cities was synchronized to the same cities declared in the BMS database. The weather information for the remaining cities in BMS except 78 was filled with the nearest neighbor's information. In the latter case, the distance to determine the nearest was measured by the simple Euclidean method.

The contractor-related information was explained by the level of construction capability of Korean contractors evaluated by CAK. The level of construction capability was calculated by sum of construction cost for recent three years, management performance, engineering capabilities, and company's credit rating. Such company's constructability level was captured from the CAK database and matched to the contractor information stored in BMS as the constructor of the bridge.

2.3 Data Characteristics

The collected data consisted of general, structural, and inspection information of bridges, weather information, and contractor-related variables. The general information bridge information included bridge classes, locations, competent authorities, offsets, detours,

traffic volumes, lengths, widths, the number of lanes, the number of spans, main structure types, substructure types, design live loads, attached facilities, and others. The structural information included span lengths, decks, girders, diaphragms, ribs for spans and support types, abutments, piers, expansion joints, shoes, and stopper factors for supports.

Such general and structural information could not be changed by time since they represent innate characteristics confirmed after construction. However, condition grades in inspection information are updated after the inspection. The weather information including temperature, relative humidity, precipitation (i.e., rainfall), sunshine, wind, and specific atmospheric phenomena (e.g., freezing and snow cover) was selected based on previous studies and time series variables [5-8].

The BMS database covered 6,773 bridges built from 1966 to 2016 in Korea and structural data included 19,625 records for spans and 32,805 rows for supports. The inspection data stored 9,775 detailed inspection records and 900 precise safety diagnoses results from 1994 to 2015, and consequently, the inspection data included 834,815 records in a level of different damage types of bridge elements.

The distribution of condition grade of the whole bridge was different from that of damages on bridge elements. The condition grades of the whole bridge were distributed as grades “A” (29.0%), “B” (67.1%), “C” (3.8%), “D” (0.1%), “E” (0%) (Figure 2). For the bridge damage level, grades “A” (5.6%), “B” (66.5%), “C” (24.8%), “D” (2.6%), “E” (0.4%), were shown in Figure 2. The number of damaged grades from “C” to “E” of the whole bridges took a substantially smaller portion than normal grades from “A” to “B”. The proportion of the damaged grades of the bridge damage level, however, was larger than the other grades.

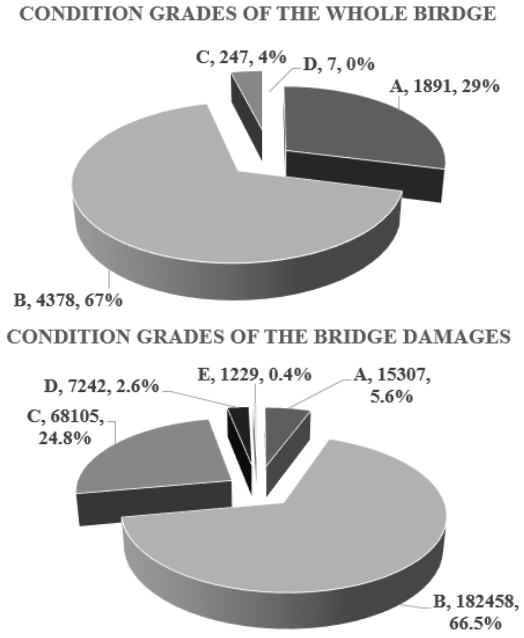

Figure 2 Distributions of condition grades: (a) Condition grades of the whole bridge; (b) Condition grades of the bridge elements

2.4 Data Preprocessing

Data preprocessing including data reduction, data cleaning, data transformation, and data integration was conducted to improve the quality of the data and thus to obtain more accurate analysis results consequently [9]. At first, data reduction was performed to reduce the number of variables as follows: text variables including descriptive variables (e.g., bridge name) which can be replaced with the codes (e.g., bridge code), data entry variables (e.g., input date), and high-rate missing and redundant values.

Next data cleaning was conducted to remove noise and inconsistent data and to fill in missing values. A specific type of bridge data was deleted since the number of data was too small to find a pattern: nine cable-stayed bridges and two suspension bridges. Randomly and frequently distributed blank cells were filled in a global constant, “9999”, which could still indicate missing value by support running the process. The reason why a global constant was used was to maintain the current distribution of data.

After downsizing the data, data transformation was performed to generate hierarchy of variables and to discretize the data. Nominal variables had the relationship between subordinates and superiors, for example, cities and provinces. By investigating such hierarchical relationships, subordinate variables were remained and superiors were removed. Data discretization was processed to replace numeric variables (e.g., bridge length) by interval categories (e.g., 11.7–23.6, 23.6-35.2, and others).

Before categorizing the continuous numerical values, the number of bins (k) was decided by Sturge’s rule (see Equation (1)) using size of data (n) [10].

$$k = 1 + 3.322(log_{10}\mathrm{n}) \quad (1)$$

The computed k values were 18 for 170,609 records of the span data and 17 for 69,858 rows of the support data. K-means clustering was then utilized to distribute one continuous variable into k clusters. This technique was used for data discretization [5, 9] based on a rule that minimizes intra-cluster distances and maximizes inter-cluster distances [11]. More specifically, k points, or *centroids*, were first initiated to the target data space. Next, each data point was assigned to the closest

centroid to make k clusters. Euclidean distance was used to calculate the closeness between the data point and the centroid. The positions of the k centroids were then recalculated with newly assigned data points to the centroids. The second and third steps were repeated until all the centroid positions did not change further [5, 9]. The "discretize" function in the "arules" package in R software was utilized.

The last preprocessing step was data integration to generate data subsets for analyses. The general and structural variables were combined according to the representative bridge number and the inspection variables were added based on the span or support numbers of the bridge having the same bridge number. The weather variables were included to the datasets according to the regional information and the contractor-related variables were added by contractor names. The final dataset consisted of 70 variables from the BMS database, 20 variables from the annual weather data, and four contractor-related variables. Since the structural variables of spans and supports were in different data formats, two datasets were eventually developed separately: the span dataset with 170,609 records and the support dataset with 69,858 records.

2.5 Data Partitioning

The region (i.e., cities) was chosen as a data partitioning criterion. Among 137 cities, three cities were selected for a pilot test based on three criteria: comparatively large amount of data, exposure to harsh weather condition (e.g., heat wave and heavy snow), and geographical conditions of the locations (e.g., mountain regions, seaboards, islands, and inland areas).

2.6 Feature Selection

When a dataset is composed of numerous predictors (i.e., independent variables), feature selection, also known as dimensionality reduction, is normally conducted to improve the learning performance of a classification model such as decision trees [12]. The feature selection increases the accuracy of results and the speed of learning by excluding less correlated features and giving higher weights for more correlated features [13], and it can consequently prevent overfitting [9].

The feature selection mainly focuses on choosing the best subset of attributes [6], so that association rule mining was implemented in this study [13]. The "association rule" is extracted based on co-occurrence of attributes, and therefore, the rule does not mean causality. This method was originated from basket analysis which aimed to find frequent buying patterns by analyzing market transaction data [13]. In this study, attributes (i.e., columns) and inspections (i.e., rows) of the span and the support datasets corresponded to items and transactions respectively.

Let $I = \{i_1, i_2, i_3, \cdots, i_d\}$ be the set of all items in market basket data and $T = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3, \cdots, \boldsymbol{t}_n\}$ be the set of all transactions. Each transaction $\boldsymbol{t}_i$ has a combination of items from I. An association rule is a set of items which contains an antecedent (i.e., "if") and a consequent (i.e., "then") with the implication expression of the form $A \rightarrow B$, where A and B are disjoint (i.e., $A \cap B \neq \emptyset$) [9, 13, 14].

To find frequent item sets, two measurements, *support* and *confidence*, act as determinants. Above all, *support count*, which is distinguished from just *support*, is defined as the number of transactions that contain a particular item set X. *Support count*, σ(X), can be mathematically expressed as follows:

$$\sigma(X) = n(\{t_i | X \subseteq t_i, t_i \in T\}) \quad (2)$$

Minimum support and *confidence* to detect the interesting rules are predefined by an analyst. *Support* counts the number of transactions which covers all items in A and B together. *Confidence* indicates how frequently items in B appear in transactions that include A. The concept of *confidence* resembles conditional probability, so that the expressions of both formulas are similar. The metrics are shown below [13]:

$$Support(A \rightarrow B) = \sigma(A \; and \; B) \quad (3)$$

$$Confidence(A \rightarrow B) = \sigma(A \; and \; B)/\sigma(A) \quad (4)$$

Apriori algorithm, proposed by Agrawal and Srikant in 1994 [15], was utilized in this research due to its efficiency on big data analysis [9]. The "apriori" function in the "arules" package on R software was used in this study.

2.7 Pattern Extraction

From the sets of selected features, damage patterns were able to be determined by decision tree analysis. A path of the tree from the top to the bottom can be a pattern.

CART (Classification And Regression Tree) algorithm was utilized in this study. Introduced by Breiman, CART was named with the emphasis on the two types of trees: the classification tree for the continuous target variables and the regression tree for discrete variables [16]. CART generates binary trees which have a branch with only two leaves, dissimilar to other decision tree algorithms such as ID3, C4.5, and CHAID [9, 17]. The Gini index is used for the attribute selection of discrete attributes and for the decrement of variance of continuous attributes. This study applied

CART because it can handle both continuous and discrete attributes simultaneously and has benefits on the treatment of missing values [18-19]. The "rpart" function in the "rpart" package on R software was utilized for the model building.

2.8 Validation

Once damage patterns were extracted, the developed model was validated by cross validation: K-fold validation. The whole dataset was divided into k subsets, and then the algorithm selected one subset as a testing set and remainders as a training set. This step was repeated k times and the model parameters were updated until the fluctuation of performance is reduced [9].

3 Pilot Results and Discussions

In this study, a pilot test using the span dataset was conducted to show the validity of the developed research framework. Through data partitioning, Ulju-gun, Inje-gun, and Mungyeong-si in Korea were selected (Figure 3). The primary criterion was the size of data and the second was specific geographical conditions of the locations: seaboards, mountain regions, and inland areas. 59 (Ulju-gun), 64 (Inje-gun), 118 (Mungyeong-si) rules were found by the association rule mining, and approximately 30 attributes appeared within the extracted rules. Table 1 summarizes data properties and the results of the association rule mining for three cities. Using the extracted rules, three decision trees were developed and patterns were produced when they met the target class of the condition level "C". A pattern for the Ulju-gun dataset explained a regional characteristic, such as heavy snow in winter, but the derived patterns were difficult to represent general regional characteristics. Nevertheless, the fact that extracted attributes were differentiated from each city represents the significance of regional partitioning.

The extracted attributes from the Ulju-gun dataset contained thickness of decks and deck elements which were distinct from the other two cities. The only one pattern was extracted as below:

- {Thickness of rib bottom flange < 24mm, thickness of web < 16mm, thickness of deck < 27.5cm, height of rail = 108cm, span length = 50m}.

The major damages found in Ulju-gun were cracks in reinforced concrete (RC) decks and clogged drainages.

Inje-gun was represented with the attributes related to decks and adjacent elements to decks. The selected patterns are found as below:

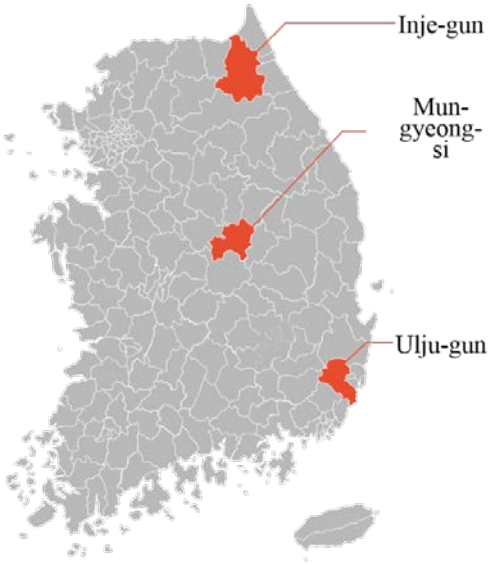

Figure 3. The selected regions in the pilot study

Table 1. The properties and the results of association rule mining for three cities

Region	Data properties			Results of association rule mining		
	No. of bridges	No. of inspec-tions	No. of dama-ges	No. of rules	Sup-port	No. of attribu-tes
Ulju-gun	45	187	3324	59	.05	29
Inje-gun	71	196	2853	64	.10	27
Mun-gyeong-si	22	116	2806	118	.12	26

* confidence = 0.5, the maximum length of rule = 4

1) {Thickness of deck pavement >= 6.5cm, depth of water < 2.15m, constructor = (L company, D1 company, D2 company, I company), height of the median >= 39cm, number of the up lines ≠ (1, 2, 3), depth of water < 0.75m},
2) {Thickness of deck pavement < 6.5cm, type of the median = (concrete, the others), rail material = (steel, aluminum alloy, the others), constructor = (D3 company, H company)},
3) {Thickness of deck pavement < 6.5cm, type of the median = (concrete, the others), rail material = concrete}.

Contractor information, thickness of decks, and the structural information of medial and rail appeared frequently from the patterns.

Damages which took large proportions of the whole Inje-gun data were (1) cracks in RC decks, (2) cracks in concrete rails and curbs, (3) breakages of concrete rails

and curbs, (4) leakages and efflorescence in RC decks, and (5) clogged drainages.

The second pattern indicated that efflorescence and corrosion in RC decks frequently occurred due to chlorides [20]. Ulju-gun has a lot of snow in winter so that de-icing chemicals containing chlorides are widely used.

In addition, water-related damages including clogged drainages, leakages, efflorescence, and corrosion in RC decks took relatively large proportion of damages.

In the Mungyeong-si dataset, plane shape and pavement related attributes were found to be used as predictors for the decision tree. Three patterns were determined as below:

1) {Pavement area < 5872m^2, constructor = (G company, D4 company), plane shape = (straight bridge with skew, curved bridge without skew), number of the down lines = 0, maximum span length >= 42.5m},
2) {Pavement area < 5872m^2, constructor = (G company, D4 company), plane shape = (straight bridge with skew, curved bridge without skew), number of the down lines = 2, depth of water >= 0.35m, pavement area >= 4620m^2},
3) {Pavement area >= 5872m^2}.

The main damage types were breakages in RC decks and breakages in steel rails and curbs. The first pattern appeared to be linked with the characteristics of breakages in RC decks, streel rails, and curbs. The first and second pattern indicated breakages in steel rails and curbs were critical. However, bridges built by Contractor G in other regions also showed 68% of breakages in steel rails and curbs.

4 Conclusions

This study developed a model to find patterns of damages on bridge elements through big data analyses. Weather and contractor-related variables were added on the BMS database, and the preprocessed dataset was divided into three cities for the pilot study. Association rule mining performed feature selection and the patterns were extracted by decision trees.

More specifically, the pilot test was conducted with the span data subsets of Ulju-gun, Inje-gun, and Mungyeong-si. The extracted patterns were different by different cities with the implication of geographical characteristics such as heavy snow and different construction capacities of contractors. The research findings showed potential to support decision makers for strategic preventive bridge maintenance.

In further study, data partitioning with various criteria will be performed using data clustering and other association rule mining and decision tree algorithms will be applied and compared to develop the best fit model.

Acknowledgments

This work was funded by the Promising-Pioneering Researcher Program through Seoul National University (SNU) in 2015.

References

[1] Liang Y., Wu D., Liu G., Li Y., Gao C., Ma Z. J., and Wu W. Big data-enabled multiscale serviceability analysis for aging bridges. *Digital Communications and Networks*, 2(3): 97-107, 2016.

[2] Ministry of Land, Infrastructure, and Transport of Korea *Yearbook of road bridges and tunnels*, 2015.

[3] Adarkwa O. and Attoh-Okine N. Prediction of Structural Deficiency Ratio of Bridges Based on Multiway Data Factorization. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, F4016002, 2016.

[4] Bu G., Lee J., Guan H., and Loo Y. Prediction of Long-Term Bridge Performance: Integrated Deterioration Approach with Case Studies. *Journal of Performance of Constructed Facilities*, 29(3): 04014089, 2015.

[5] Huang R., Mao I. S., and Lee H. Exploring the deterioration factors of RC bridge decks: a rough set approach. *Computer-Aided Civil and Infrastructure Engineering*, 25(7):517-529, 2010.

[6] Melhem H. G., Cheng Y., Kossler D., and Scherschligt D. Wrapper Methods for Inductive Learning: Example Application to Bridge Decks. *Journal of Computing in Civil Engineering*, 17(1):46-57, 2003.

[7] Huang R. and Hsu W. Reliability-based component deterioration model for bridge life-cycle cost analysis. *Journal of Chinese Institute of Civil and Hydraulic Engineering*, 17(4):679-691, 2005.

[8] Creary P. A. and Fang F. C. The Data Mining Approach for Analyzing Infrastructure Operating Conditions. *Procedia - Social and Behavioral Sciences*, 96:2835-2845, 2013.

[9] Han J., Kamber M., and Pei J. *Data mining: concepts and techniques*, Third Edition. Elsevier, Waltham, United States, 2012.

[10] Nie N. H., Hull C. H., Jenkins J. G., Steinbrenner K., and Bent D. H. *SPSS statistical package for the social sciences*, Second Edition. McGraw-Hill, New York, United States, 1975.

[11] Morcous G., Rivard H., and Hanna A. Modeling bridge deterioration using case-based reasoning. *Journal of Infrastructure Systems*, 8(3):86-95, 2002.
[12] Kohavi R. and John G. H. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273-324, 1997.
[13] Rajeswari K. Feature Selection by Mining Optimized Association Rules based on Apriori Algorithm. *International Journal of Computer Applications*, 119(20), 2015.
[14] Slimani T. and Lazzez A. Efficient analysis of pattern and association rule mining approaches. *ArXiv preprint arXiv*: *1402.2892*, 2014.
[15] Agrawal R. and Srikant R. Fast algorithms for mining association rules. Proceedings of 20th International Conference on Very Large Data Bases, 1215:487-499, 1994.
[16] Breiman L., Friedman J. H., Olshen R. A., and Stone C. J. *Classification and regression trees*. Wadsworth & Brooks. Monterey, CA, 1984.
[17] Huang R. and Chen P. F. Analysis of influential factors and association rules for bridge deck deterioration with utilization of national bridge inventory. *Journal of Marine Science and Technology*, 20(3):336-344, 2012.
[18] Patidar P. and Tiwari A. Handling Missing Value in Decision Tree Algorithm. *International Journal of Computer Applications*, 70(13):31-36, 2013.
[19] Ture M., Tokatli F., and Kurt I. Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4. 5 and ID3) in determining recurrence-free survival of breast cancer patients. *Expert Systems with Applications*, 36(2):2017-2026, 2009.
[20] Cusson D., Lounis Z., and Daigle L. Durability Monitoring for Improved Service Life Predictions of Concrete Bridge Decks in Corrosive Environments. *Computer-Aided Civil and Infrastructure Engineering*, 26(7):524-541, 2011.