Decision Tree Analysis of the Relationship between Defects and Construction Inspection Grades

Chien-Liang Lin and Ching-Lung Fan

Abstract-Defects are an important indicator of project quality; moreover, eliminating defects is a key objective of project management. Therefore, using the appropriate analytical tools and methods, training and testing the defect data, and selecting the best algorithm for the defect feature are important. These steps can directly reveal the decision rules for each defect, and they can assist in determining key approaches to construction site management for effective defect prevention. In this study, a model obtained by using the chi-squared automatic interaction detection (CHAID) algorithm was validated, and its prediction benefits were calculated. A total of 499 defect types were retrieved from the Public Construction Management Information System in Taiwan and used as the foundation of a statistical analysis of 990 construction projects with 17,648 construction defects. First, a cluster analysis of inspection scores and defect frequencies was performed to reclassify and establish a new grade. Next, five rules were established for using the decision tree to classify defects and inspection grades. Finally, results revealed that the prediction accuracy of the CHAID algorithm was 75.45%. The five rules can be used for defect management and prevention strategies.

Index Terms—Decision tree, CHAID, defect, inspection grades.

I. INTRODUCTION

Defects are common phenomena in the construction industry, and may adversely affect the cost, duration, and resources of a project. Failures and defects can result in unnecessary expenditure and delays. Because construction defects generally occur when several associated causes combine, a thorough understanding of defect causality is required to systematically prevent construction defects [1]. Therefore, understanding the impact of defect causality is crucial for the development of strategies to prevent and reduce defects.

Various methods have been proposed to analyze the causes of defects [2]-[5]. Among them, data mining (DM) is a procedure for extracting useful information from data, analogous to digging for an ore in a mine. Such information can reveal unexplained or undiscovered causal relationships [6].

In recent years, the construction industry has used DM techniques to analyze engineering defects to explore the

Chien-Liang Lin is with the Department of Construction Engineering, National Kaohsiung of Science and Technology, 2 Jhuo-yue Rd., Nan-zih, 81164, Kaohsiung, Taiwan (e-mail: ken@nkfust.edu.tw).

Ching-Lung Fan is with Institute of Engineering Science and Technology, National Kaohsiung University of Science and Technology & Department of Civil Engineering, the Republic of China Military Academy, 1 Wei-Wu Rd., Fengshan, 83059, Kaohsiung, Taiwan (e-mail: u0315916@ nkfust.edu.tw). association rules of special value [7], [8]. DM is believed to have excellent potential for defect association analysis or prediction.

The main purpose of DM is to detect, interpret, and predict qualitative and quantitative patterns in the data, thus yielding new information and knowledge. Several methods and algorithms, such as neural networks, support vector machines, and decision tree (DT), have been employed in DM to perform data classification and predictive modeling [9], [10].

Among the analysis methods for classification and prediction, DT have a structure that is easy to understand and explain, and a tree structure diagram is often used to solve a series of classification and decision-making problems for classifying numerous attributes and making predictions. DT can also be used for classification studies in engineering and management [11]-[17]. However, DT techniques have not been widely applied to defect data analysis.

In Taiwan, the Mechanism of Public Construction Inspection has been implemented since 1993. Defect data related to project inspections are posted on the Public Construction Management Information System (PCMIS). These inspection data include the grades and defects related to project quality and management performance. Therefore, the relationship between inspection grades and defects is suitable for in-depth analysis and research.

The construction inspection database of the PCMIS can be used as the analysis object. The DT chi-squared automatic interaction detection (CHAID) algorithm can be used to determine the classification rules by using the defects reported in the data. A construction management unit can develop appropriate management strategies to improve the project quality and performance.

II. DATA SOURCE AND METHODOLOGY

To improve the quality of public projects, Taiwan has established an effective quality management system so that all groups and members involved in construction tasks can ensure the quality of their works. In particular, to meet the quality standards and requirements in the construction process, systematic management, effective control steps, and attention to the construction quality are crucial. In 1993, a three-level quality management structure was formulated for the Public Construction Quality Management System, and the Quality Control System was established for contractors. Moreover, the Quality Assurance System was established by the procuring units, and the competent authorities established the Mechanism of Public Construction Inspection.

The construction inspection in this study refers to the construction inspection records of the Level 3 quality control

Manuscript received May 8, 2018; revised December 30, 2018.

implemented by the public construction surveillance units of the central and local governments.

The public construction surveillance unit comprises experts and scholars (surveillance committee). The committee inspects the quality and progress of a project according to the Public Construction Quality Management System, relevant laws and regulations and the project contract, and the reference standard of the construction inspection work.

Members of the surveillance committee are dispatched for construction inspection. On the inspection day, the committee visits the construction site to inspect the scene, listen to briefings, review relevant documents, and host a quality review meeting. During the meetings, the surveillance committee notifies both the procuring unit and the contractor regarding the evaluation of any defects and provides the project team with an opportunity to clarify and resolve these defects. Finally, the committee itself hosts a meeting to summarize the inspection findings and determine the score and inspection grades, which are then published in the PCMIS administered by the procuring unit.

The score of a construction inspection is calculated based on the average of the scores of all surveillance committees, and is divided into four grades: S (90–100 points), A (80–89 points), B (70–79 points), and C (less than 70 points). High inspection grades and scores generally indicate that a project has an excellent quality management system, high construction quality, well-managed progress, and sound planning and design. That is, higher scores reflect better performance in project management.

Since the implementation of this construction inspection mechanism in Taiwan, numerous inspection records of public projects have been accumulated. The inspection content is divided into four categories, namely, Quality Management System (A1, 113 defects), Construction Quality (B1, 356 defects), Construction Progress (C1, 10 defects), and Planning and Design (D1, 20 defects).

This study retrieved construction inspection records from the PCMIS. A total of 990 projects were sampled from January 2003 to October 2016; the sample included 499 types of defects and 17,648 defect frequencies. DM was performed to analyze the rules governing the relationship between the defects and inspection grades. The CHAID algorithm, which is suitable for the construction inspection database, was selected for conduct the DM. Thus, the rule sets for defect classification were established.

Before performing DM, the data must be preprocessed according to the characteristics of the algorithm. First, statistical analysis revealed that the numbers of S (0.6%) and C (0.2%) inspection grades were significantly lower than those of A (77.1%) and B (22.1%) grades. The unbalanced nature of the data may be problematic if DT classification is performed directly. That is, no significant difference will be observed, even if the number of cohorts is small [18].

Therefore, a cluster analysis was first conducted to regroup the sample data by inspecting the score and defect frequency to mitigate the variance of sample sizes for various inspection grades, thereby ensuring the accuracy and validity of the results. After obtaining the new grades, the data were randomly divided into training and test groups. The DT model for defects was then established, and the model was evaluated to determine its suitability for the defect database. In this study, the PCIMS construction inspection database was used to divide the data into training and test groups with proportions of 70% and 30%, respectively. The model obtained using the CHAID algorithm was used to verify.

After completing the defect classification, a gain chart was used to assess the effectiveness of the classification model. Finally, the rules yielded by the classification can be used for defect prevention.

A. Cluster Analysis Algorithm

Cluster analysis, also known as affinity grouping, is a kind of unsupervised analysis. It divides data into groups, in which a high degree of affinity exits in the same group and apparent differences are present between different groups. A cluster analysis is employed to create groups and assist decision makings by pointing out common characteristics among groups.

K-means is an algorithm frequently used in cluster analysis. A random selection of k seeds from the data is made in this algorithm according to the expected clusters (k) to be divided. These seeds will be the initial centers for the clusters, and once the k seeds are decided, the rest of the samples (p) will be assigned to the clusters in the nearest proximity. After that, the center of each cluster will be re-computed, the distance between each sample and the new cluster center be compared, and the clusters be regrouped. The whole process will be repeated again and again until the minimum SSE (sum of the squared errors) is reached. The calculation equation is shown as follows (1):

SSE =
$$\sum_{i=1}^{k} \sum_{p \in ci} |p - m_i|^2$$
. (1)

k is the number of cluster; p is the sample in the space group, m_i the mean of the samples in the category of c_i , and SSE is the sum of the squared errors of all the samples.

First of all, in order to bring as close as possible the sample sizes in the inspection grades to create new grades, the K-means cluster analysis in SPSS was implemented on the 990 construction project scores and the defect frequencies (the number of times is 17,648). Ultimately, the sample sizes were produced from the four new grades: A = 458 projects, B = 135 projects, C = 303 projects and D = 94 projects.

B. DT Algorithm

A DT has a tree-like structure comprising a root node, branches, and leaf nodes. Each node represents an attribute, leaf nodes represent classification categories, and branches represent value in the test attributes.

Each DT begins with the development of the root node, and nodes are added according to the problems and attributes of the classification. Branches are used to determine whether a data entry should be applied to a child node in the next layer. This process continues until all data entries reach the leaf nodes. Leaf nodes represent the obtained result, which can be converted into the corresponding "if... then..." rule.

DT are based on the input variables used to obtain the

target variables for distribution of the attributes and the root nodes for verification and data cutting. According to the DT rules and methods for data classification, each branch represents a test result. The leaf nodes represent the distribution of the target variables and are represented by the shape of the branches. Each path from the root node to the leaf nodes can be extracted using a DT rule.

DT algorithms can determine tree structures that strongly influence the attributes of the target variable distribution. In addition, they can classify the data through the selection of variables and designation of the target, and they can present the classification system or prediction model with a hierarchical structure.

The DT algorithm commonly used in academia and industry can be classified into four types: C5.0, CART, CHAID, and QUEST [19]. The main differences between these algorithms are the methods of splitting the attribute criteria of the root and child nodes and the numbers of split child nodes.

In the CHAID algorithm proposed by Kass [20], the chi-squared test is used to determine the splitting condition, and the probability value is used to determine whether to continue splitting the CHAID to measure all the possible prediction variables. For each variable, the significances of the differences between the various categories of dependent variables are tested, the non-significant categories are merged into a homogeneous group, and the remaining categories are analyzed repeatedly until the differences are no longer significant.

The CHAID calculation terminates when sufficient p values have been obtained. This method of splitting is a common alternative, known as the exhaustive method. Merging continues until a binary split, after which the split with the most favorable p value among all the splits is adopted. Thus, the CHAID algorithm does not require the prune back operation employed in the CART algorithm; however, the CHAID algorithm cannot process continuous data.

III. ANALYTICAL RESULTS AND DISCUSSION

In this study, a DT was used to analyze construction defects and grades. A total of 990 construction inspection cases (17,648 defect frequencies) were divided into training and test groups with proportions of 70% and 30%, respectively. The CHAID algorithm employs data from the test group to validate the model and calculate its prediction accuracy. The DT in this study gained defect frequencies (see Table I).

The DT constructed by the CHAID algorithm is shown in Fig. 1. The target attribute in the root node represents the inspection grades. The first-level attribute is "failure to inspect construction progress, building material, or equipment; failure to fill or compile checklists; misjudgment, or failure of implementation" (A47). The second-level attributes are "debris on concrete surface" (B4) and "substandard concrete pouring or ramming, resulting in cold joints, honeycombs, or pores" (B1). The third-level attributes are "failure to log the construction journal, failure to log in the predetermined format, or incomplete logging" (A75),

"steel bars exposed in the workplace are prone to stabbed and bruised disasters, and did not take a curved tip, stamping or installation of protective sheathing and other facilities" (B307), "failure to meet the garbage and waste cleaning requirements and generated environmental impacts" (B59), and "failure to install required fall protection facilities such fences, covers, safety nets, and seat belts on jobsite fringes and openings with height gaps of at least 2 m" (B285). The fourth-level attributes are B307, B1, B293, A48, A75, B58, A78, and B282.

TABLE I: DEFECT TYPES AND FREQUENCIES

Code	Defect type	Defect
Code	Delect type	frequencies
A4	Lack of, unimplemented, or incomplete quality	186
	supervision and inspection records	
A47	Failure to inspect construction progress,	293
	building material, or equipment; failure to fill or	
	compile checklists; misjudgment, or failure of	
	implementation	
A48	Failure to notify the supervision unit or	184
	contractor in written form to mitigate known	
	construction defects within the deadline	
A75	Failure to log the construction journal, failure to	406
	log in the predetermined format, or incomplete	
	logging	
A76	Failure to implement quality control checklist,	657
	failure to quantify inspection standards or	
	errors, or failure to record the inspection	
170	accurately	212
A/8	Environmental protection, construction safety	213
	and sanitation compliance matters without	
4.01	Ne results content statistical exclusion encoder	170
A81	no quality control statistical analysis, corrective	170
R 1	Substandard concrete pouring or ramming	275
DI	resulting in cold joints, honeycombs, or pores	215
B/	Debris on concrete surface	264
B58	Construction site groundwater is not treated	66
B 50	affecting environmental sanitation and safety	00
B59	Failure to meet the garbage and waste cleaning	157
207	requirements and generated environmental	107
	impacts	
B282	Other recording errors in building material and	457
	equipment reviews	
B285	Failure to install required fall protection	167
	facilities such fences, covers, safety nets, and	
	seat belts on jobsite fringes and openings with	
	height gaps of at least 2 m	
B293	The construction frame is not properly	83
	connected with the stable structure or does not	
	meet the requirements	
B307	Steel bars exposed in the workplace are prone to	96
	stabbed and bruised disasters, and did not take a	
	curved tip, stamping or installation of protective	
	sheathing and other facilities	

In this study, five rules were obtained using the CHAID algorithm (Table II). Each rule was mainly focused on the defect attribute of grade D for determining the conditions that are likely to result in poor inspection grades. The relationship between the defects can be determined from these five rules and the percentages of the four grades (A, B, C, and D).

For example, in the case of CHAID Rule 1, if A47 is not satisfied, but B4, B307, and A48 are satisfied, then the probability obtaining grade D is 80%, whereas those for both

A and B are 10%. Table II shows that CHAID Rule 5 yielded the highest probability of predicting grade D (93.75%),

whereas CHAID Rule 3 had the lowest probability of predicting grade D (57.14%).



Fig. 1. Defect decision tree: CHAID algorithm.

TABLE II: CHAID RULES AND INSPECTION GRAD

Rule	Decision Rules	Inspection grades			
		A(%)	B (%)	C (%)	D (%)
1	if A47=0 & B4=1 & B307=1 & A48=1 then D	10.00	10.00	0	80.00
2	if A47=1 & B1=0 & B59=1 & B58=1 then D	0	0	12.50	87.50
3	if A47=1 & B1=1 & B285=0 & A78=1 then D	0	28.57	14.29	57.14
4	if A47=1 & B1=1 & B285=1 & B282 = 0 then D	20.00	0	20.00	60.00
5	if A47=1 & B1=1 & B285=1 & B282 = 1 then D	0	6.25	0	93.75

In addition, the importance value was used to represent the relative importance of the DT attributes. The higher the importance value of the attribute, the strong its predictive ability. Attribute importance was evaluated by computing the total reduction of impurity achieved by each attribute. The summation of the importance values for all selected attributes in a given algorithm must be 1. In this study, the attribute importance values were between 0 and 1.

Fig. 2 shows the importance values of the defect attributes according to the CHAID algorithm. Evidently, the most

important attribute of the CHAID model was A47, which indicated that this attribute has the highest predictive power among all the defect attributes. In order of importance, "failure to inspect construction progress, building material, or equipment; failure to fill or compile checklists; misjudgment, or failure of implementation" (A47) was followed by "debris on concrete surface" (B4; 0.183), "substandard concrete pouring or ramming, resulting in cold joints, honeycombs, or pores"(B1; 0.138), and "failure to log the construction journal, failure to log in the predetermined format, or incomplete logging" (A75; 0.117).



Fig. 2. Defect importance values produced by the CHAID algorithm.

In this study, we used a gain chart to assess the classification effectiveness of the DT and a confusion matrix to determine the correctness of classification predictions.

The gain chart in Fig. 3 compares the most commonly used DM methods, with the horizontal and vertical axes both representing percentages. The horizontal axis percentages are sorted according to probability from high to low, representing percentages of the test data set; the vertical axis represents the percentages of correctly predicted values (gain).



The upward curve of gain chart indicates that the better the model is, the greater is the area under the curve. If the

model's cumulative gain curve displays a 45° angle to the horizontal axis (e.g. green line), then the model is a random model, and the predicted benefits cannot be classified.

The classification of predictive ability of the DT in this study reveals that 80% of the predicted values are correct when more than 50% of the test group is used (Fig. 3).

The confusion matrix in Table III compares the model predictions with the test data to measure the effectiveness of the model classification and thus evaluate the supervised data classification. The confusion matrix uses a matrix for data visualization.

In an ideal model, the classification predictions match the actual values. However, most model predictions are not 100% accurate. For example, a set of attributes resulting in grade A, were predicted by the DT model to yield grade B. Therefore, a confusion matrix is required to analyze the recall and correctness of the model predictions. The confusion matrix presents all the classification information in a complete matrix.

Table III and IV show the recall and correctness of the CHAID algorithm. The recall for a category is the number of predictions in this category divided by the actual number of members of this category. For example, the CHAID algorithm correctly predicted 351 entries as grade A; dividing this by the total number of actual grade A entries (458) yields a recall of 76.64%. The CHAID algorithm yielded similar recalls for grades A, B, C, and D grades between 68% and 78%.

Correctness is the number of correct predictions across all categories divided by the total number of entries. For example, the CHAID algorithm correctly predicted 351, 93, 239, and 64 entries for grades A, B, C, and D, respectively (Total 747); the summation of these values divided by the total number (990) yields the correctness of 75.45% (Table IV).

TABLE III: CONFUSION MATRIX FOR CHAID ALGORITHM

TABLE III. CONFUSION MATRIX FOR CHAID ALGORITHM						
N=990	Predicated		Total	Recall		
Actual	А	В	С	D	(N)	(%)
А	351	39	65	3	458	76.64
В	35	93	3	4	135	68.89
С	53	4	239	7	303	78.88
D	10	4	16	64	94	68.09

TABLE IV: CORRECTNESS FOR CHAID ALGORITHM				
Partition	CHAID			
Classification correct (N, %)	747	75.45		
Classification incorrect (N, %)	243	24.55		

IV. CONCLUSION

Construction projects have varying levels of complexity. When 499 defect types are found during a construction inspection, all resources cannot be devoted to comprehensive defect management. Therefore, project managers must choose the most important defects for the implementation of control mechanisms. In particular, because defects directly affect inspection grades, special attention is required to avoid producing them.

DT analysis is a powerful technique because of its binary "if... then..." rule based structure. The rules applied for DM through DT analysis are easy to understand and explain. Therefore, this study employed DT techniques to classify defect rules from the PCMIS construction inspection database. The rules can enable the project management unit to choose the optimal classification algorithm and develop appropriate strategies for defect management.

DM techniques can be used to acquire information hidden in decision rules. By selecting and interpreting implicitly useful rules, DM can yield reliable information regarding construction defects. In addition, cluster analysis of scoring and defect frequencies obtained from the construction inspection database can reduce the number of samples in inspection grade groups, thereby ensuring more accurate results.

This study employed the CHAID algorithm to determine the results of individual defect attributes and effectively generate the defect importance values. The results revealed that "failure to inspect construction progress, building material, or equipment; failure to fill or compile checklists; and misjudgment or failure of implementation" (A47) was the most important attribute.

In addition, the CHAID algorithm yielded five rules that accurately described the association between attributes (grades) and defect predictions. This study demonstrated that DT techniques are useful for analyzing the defect data of construction inspections.

REFERENCES

- Y. Cheng, W. D. Yu, and Q. Li, "GA-based multi-level association rule mining approach for defect analysis in the construction industry," *Automation in Construction*, vol. 51, pp. 78-91, March 2015.
- [2] P. E. Josephson and Y. Hammarlund, "The causes and costs of defects in construction: A study of seven building projects," *Automation in Construction*, vol. 8, no. 6, pp. 681-687, August 1999.
- [3] P. E. D. Love and D. J. Edwards, "Forensic project management: The underlying causes of rework in construction projects," *Civil Engineering and Environmental Systems*, vol. 21, no. 3, pp. 207-228, 2004.
- [4] N. Forcada, M. Macarulla, M. Gangolells, M. Casals, A. Fuertes, and X. Roca, "Post-handover housing defects: Sources and origins," *Journal* of *Performance of Constructed Facilities*, vol. 27, no. 6, pp. 756-762, 2013.
- [5] H. Aljassmi, S. Han, and S. Davis, "Analysis of the complex mechanisms of defect generation in construction projects," *Journal of Construction Engineering and Management*, vol. 142, no. 2, 2016.
- [6] W. T. Lin, S. T. Wang, T. C. Chiang, Y. X. Shi, W. Y. Chen, and H. M. Chen, "Abnormal diagnosis of emergency department triage explored with data mining technology: An emergency department at a medical center in Taiwan taken as an example," *Expert Systems with Applications*, vol. 37, no. 4, pp. 2733-2741, April 2010.
- [7] Y. M. Cheng and S. S. Leu, "Integrating data mining with KJ method to classify bridge construction defects," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7143-7150, June 2011.

- [8] S. Lee, S. Han, and C. Hyun, "Analysis of causality between defect causes using association rule mining," *International Journal of Civil, Environmental, Structural, Construction and Architectural Engineering*, vol. 10, no. 5, pp. 659-662, April 2016.
- [9] Y. U. Ryua, R. Chandrasekaranb, and V. S. Jacobc, "Breast cancer prediction using the isotonic separation technique," *European Journal* of Operational Research, vol. 181, no. 2, pp. 842-854, 2017.
- [10] Y. S. Kim, "Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size," *Expert Systems with Applications*, vol. 34, no. 2, pp. 1227-1234, 2008.
- [11] M. J. Lee, A. S. Hanna, and W. Y. Loh, "Decision tree approach to classify and quantify cumulative impact of change orders on productivity," *Journal of Computing in Civil Engineering*, vol. 18, no. 2, pp 132-144, April 2004.
- [12] D. Arditi and T. Pulket, "Predicting the outcome of construction litigation using boosted decision trees," *Journal of Computing in Civil Engineering*, vol. 19, no. 4, pp. 387-393, 2005.
- [13] Y. Shin, T. Kim, H. Cho, and K. I. Kang, "A formwork method selection model based on boosted decision trees in tall building construction," *Automation in Construction*, vol. 23, pp. 47-54, May 2012.
- [14] S. Syachrani, H. S. Jeong, and C. S. Chung, "Decision tree-based deterioration model for buried wastewater pipelines," *Journal of Performance of Constructed Facilities*, vol. 27, no. 5, pp. 633-645, 2013.
- [15] G. Mistikoglu, I. H. Gerek, E. Erdis, P. E. M. Usmen, H. Cakan, and E. E. Kazan, "Decision tree analysis of construction fall accidents involving roofers," *Expert Systems with Applications*, vol. 42, no. 4, pp. 2256-2263, March 2015.
- [16] J. S. Chou, S. C. Hsu, C. W. Lin, and Y. C. Chang, "Classifying influential information to discover rule sets for project disputes and possible resolutions," *International Journal of Project Management*, vol. 34, no. 8, pp. 1706-1716, 2016.
- [17] M. Alipour, D. K. Harris, L. E. Barnes, and O. E. Ozbulut, "Load-capacity rating of bridge populations through machine learning: Application of decision trees and random forests," *Journal of Bridge Engineering*, vol. 22, no. 10, 2017.
- [18] J. Pallant, PSS Survival Manual: A Step by Step Guide to Data Analysis Using SPSS, 3th ed., New York: McGraw-Hill, 2007, ch. 18, pp. 242-256.
- [19] J. Han and M. Kamber, *Data Mining Concept and Technology*, 3th ed., San Francisco: Morgan Kaufmann, 2012, ch. 8, pp. 330-348.
- [20] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Applied Statistics*, vol. 29, no. 2, pp. 119-127, 1980.



Chien-Liang Lin is an associate professor of construction engineering at National Kaohsiung University of Science and Technology. He received his M.S. and Ph.D. degree from the Penn State University in 1991-1996. His works focus specifically on system dynamics, operations research, system thinking and learning organization, and productivity analysis & control.



Ching-Lung Fan is a lecturer in the Department of Civil Engineering at the Republic of China Military Academy. He received his M.S. degree from the National Taiwan University in 2006. Right now, he is a Ph.D. candidate in the Institute of Engineering Science and Technology at National Kaohsiung University of Science and Technology. His primary research interests include construction management, performance evaluation of construction, construction

risk management, and construction productivity.