Graph-based Clustering of Bridge Management System Data for Bridge Maintenance Cost Estimation

Gyueun Lee, Seokho Chi
Department of Civil and Environmental Engineering, Seoul National University, South Korea
gyueun91@snu.ac.kr

Abstract. The number of bridges has increased, and the data acquired from bridge maintenance works have accumulated more in the bridge management systems (BMS). However, the current inflexible BMS schema is limited in dealing with large-scale data with complex relationships. The historical inspection and repair data of the bridges were not able to be considered during maintenance cost estimation in practice because of the complexity of joining the BMS data. Therefore, this study proposed a model for converting BMS data into a graph database by modelling the relationships between graphical nodes. The authors then compared the clustering results between the existing relational database model and the proposed graph-based model. The experimental results of the silhouette coefficient demonstrated that the graph-based model outperformed the existing base model. In addition, a significant difference in historical maintenance cost was identified. This research suggests a promising direction of utilizing graph database to enhance the clustering performance for bridge maintenance cost estimation.

1. Introduction

The number of bridges has been increasing yearly, while many countries have limited budgets and resources for bridge maintenance. South Korea is also experiencing an increased, required maintenance budget because national bridges have been growing in length, but their condition grades are deteriorating. However, the current practice of estimating bridge maintenance cost is based on the number of bridges and their approximate condition grade, without considering the unique structural and operating characteristics of each bridge. Even if the bridges have the same condition grade, the required repair costs can vary significantly depending on their individual characteristics. Thus, the cost estimation for bridge maintenance by grouping bridges with similar characteristics is necessary to ensure optimal allocation of resources per different bridge groups and prepare effective maintenance strategies (Cheng and Leu, 2009).

Several research has been conducted to group similar bridges by using the data collected in the bridge management system (BMS) (Miyamoto et al., 2000; Wu et al., 2021). In general, BMS is designed as a relational database (RDB), which is a highly structured database in a table form. RDB has the advantage of consistent data management with its structured schema, but it also limits the flexibility to modify the data structure and perform complex joins between tables (Nayak et al., 2013). In BMS, the basic specifications of the bridge and the annual inspection and repair records are often stored as separate tables in practice. In this situation, the latent knowledge and the relationships among variables are missed and data analysis becomes limited. To overcome the limited usage and applicability of the traditional RDB, a graph database is introduced to explain the relationships among data.

In this paper, the authors devised an approach to clustering the bridges using a graph database. This approach can consider the latent features which were hard to be considered in the traditional relational BMS. To achieve the goal, several tasks were completed. The first task generated the graph database and transformed the original BMS data into it. Secondly, the research embedded the features in nodes and the relationships as vectors. Third, the embedded data were clustered using the K-means++ algorithm. The clustering performance of the graph-based and base models were compared using a silhouette coefficient value. Lastly, the authors

validated the usability of clustering for the maintenance cost estimation using the historical repair and inspection data.

2. Literature Review

The authors reviewed previous research on the clustering of bridge management data and the effectiveness of graph database techniques.

The expected bridge maintenance cost has been calculated by multiplying the representative maintenance cost of each condition grade by the number of bridges (Sun et al., 2020). This uniform unit cost by the bridge condition grade has limitations to reflect the structural and operating characteristics of each bridge. Many researchers have thus attempted to apply the clustering algorithms to group the bridges based on their similarities, such as their locational proximities, structure type, and material, and assign appropriate maintenance costs per bridge clusters (Cheng and Leu, 2009; Galvan-Nunez and Attoh-Okine, 2017; Radovic et al., 2017). They often used the bridge's basic specifications data from National Bridge Inventory (NBI) database like BMS for clustering. However, due to the complex multidimensionality of the database, the data gathering has been used in a limited way. To overcome the limited usage of the data, additional efforts have been made. For instance, Liu and El-Gohary linked the data extracted from bridge inspection reports for clustering (Liu and El-Gohary, 2022). Despite the efforts, there is still a lot of room to improve bridge clustering results. Recently, there has been a growing interest in various fields to improve data modeling and clustering performance by using a graph database (Tiwary et al., 2022).

To summarize, many researchers have utilized BMS data which are structured in relational databases (Ghahari et al., 2019) to perform efficient maintenance data analysis. However, due to the limitations of the data structure and inflexible formats, the decision makers are facing practical drawbacks. In this study, the authors propose a method to overcome these limitations by converting the relational database into a graph database.

3. Research Process

Figure 1 shows the research framework. The prepared data were the records of the bridge's decks in the Korean BMS from 2009 to 2022. There were three kinds of data tables related to maintenance. The first table was the specifications of the bridges, which included basic information (e.g., construction year, structural type, length, width, and whether the seismic design was applied). Environmental information (e.g., the annual daily traffic and precipitation) was also set as properties. A total of 4,063 bridges with at least one deck inspection record were used. The second table was the historical inspection records of the bridge decks. This table included information on when the inspection was performed, the extent of the observed damages, and the expected cost of repair and reinforcement works. There were 9,320 records in this table. The third table comprised of the repair and reinforcement project information. This table contained information on when and how much the repair and reinforcement works were performed for each bridge, and there were 1,657 records in total.

The authors created the base model and the proposed model to compare clustering performances. The base model only used the bridge specification data such as length, width, construction date, and environmental variables for clustering. The proposed model used the bridge specification, inspection records, and repair records together to generate the graph database for clustering.

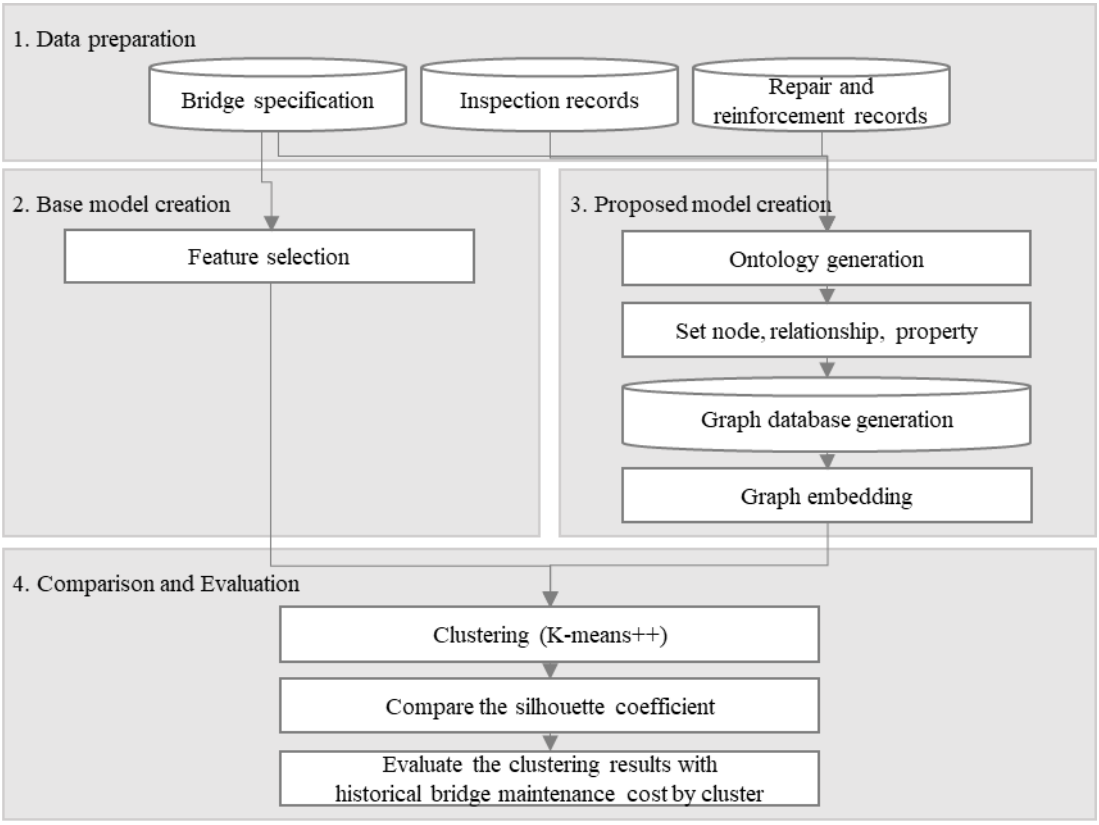

Figure 1: Research Framework

3.1 Nodes, Relationships, Properties Setting for Graph Database

Neo4j, which is one of the best functional non-relational databases (Fernandes and Bernardino, 2018), was used to create the Korean BMS data as the graph database. It stores and manages the data in the form of nodes, relationships, and properties. It is supported by a query language called Cypher (Francis et al., 2018) to retrieve and analyze the data in Neo4j.

Before making the graph database, the ontology for bridge maintenance data was generated to set the entities and their relationships clearly. The ontology served as a blueprint for graph database generation. As shown in Figure 2(a), the bridge is the main node in the ontology, and the related nodes are connected by edges. Figure 2(b) is the meta graph of BMS data imported in Neo4j. When importing them in Neo4j, the variables presenting the node were converted to properties in each node. Table 1 shows the sample of main nodes and their properties. The bridge node contained the properties that represented the specifications of the bridge. The inspection node contained the defects-related information observed in the deck. If defects occurred on the bridge, the quantity of defects and the expected maintenance cost were filled as properties. For example, the 'B07101' was the code of the defect which meant 'Crazing in Reinforced Concrete', and the attached number meant the quantity of the defect. The repair work node contained the repair and rehabilitation method with their cost. The 'Cost_100' was the code of the repair method which meant 'Injection,' and the attached number represented the cost of the repair method.

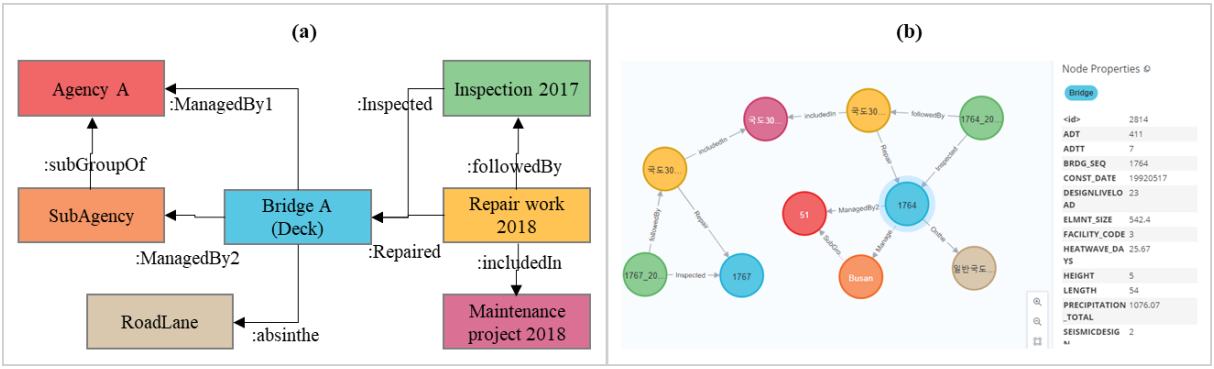

Figure 2: Concept on the Bridge Maintenance Data Graph: (a) Bridge Maintenance Ontology, (b) Meta Graph in Neo4j

Table 1: Properties of each Nodes.

Node Labels	Number of Node	Property (sample)
Bridge	4,036	BRDG_SEQ:1876, ADT:26199.67, ADTT:283.17, CONST_DATE:1998-12-31, DESIGNLIVELOAD:24, ELMNT_SIZE:8745.0, FACILITY_CODE:1, HEATWAVE_DAYS:10.33, HEIGHT:37.3, LENGTH:750.0, PRECIPITATION_TOTAL:1497.97, SEISMICDESIGN:2, SUPERSTRUCTURE:29, TEMPERATURE_WINTER:-0.47, TEMPERATURE_YEAR:12.2, WIDTH:11.0
Inspection	9,320	brdg_inspctdt:2552_20130420, BRDG_SEQ:2552, B07100:71.6, B07101:0.0, B07102:0.0, B07103:0.5, B07105:0.0, B07106:5.95, B07107:0, B07108:40.88, B07111:0.0, B07121:0.0, B07200:0.0, B07201:0.0, B07202:0.0, B07221:0.0, B07222:0.0, B07326:0.0, B07999:1730.62, ELMNT_GRADE:2(B), INSPCT_YEAR:2013, PRIORITY:3, TOTAL_EXPCTCOST:1741426.36, USEYEAR:26.3
Repair Work	1,847	brdg_rprdt:10044_20171222, BRDG_SEQ:10044, CMPLT_DT:2017-12-22, Cost_100:97.0, Cost_200:5714.0, Cost_300:48817.0, Cost_600:0.0, Cost_800:0.0, RPRYEAR:2017, TotalCost:54628.0,

Table 2 shows how the nodes were connected by relationships. The authors focused on what the existing relational database could not express and tried to inherit the knowledge of bridge maintenance in the graph database. The “Inspected” and “Repaired” relationships were modeled to imply which maintenance work was done on the bridges. The “followedBy” relationships served as a link connecting the preceding inspection and the subsequent repair work. The “includedIn” relationships connected the repair work and the parent projects to link the bridges that had been repaired together by the same project. The “ManagedBy” relationships connected the bridges managed by the same agency.

Table 2: Node and Relationship of the BMS data modeled in Neo4j *(Continue)*.

Relationship	(Node)-[Relationship]->(Node)	Meaning
Inspected	(:Inspection)-[:Inspected]→(:Bridge)	Bridges are connected to their inspection records.
Repaired	(:RepairPjt)-[:Repaired]→(:Bridge)	Bridges are linked to repair work performed.
includedIn	(:RepairPjt)-[:includedIn]→(:Mproject)	Each repair work is connected to a parent project.

followedBy	(:RepairPjt)-[followedBy]→(:Inspection)	The repair work performed after the inspection is linked.
ManagedBy	(:Bridge)-[:ManagedBy]→(:Agency) (:Bridge)-[:ManagedBy]→(:SubAgency)	Bridges are managed by agency and their subagency.
Subgrouped	(:SubAgency)-[:subGroupOf]→(:Agency)	Subagency is the subgroup of the agency.
isOnthe	(: Bridge)-[:isOnthe]→(:RoadLine)	Bridges on the same road line are connected.

3.2 Graph Embedding

Graph data have a triple form, and are expressed as the nodes and the edge connecting them (Angles, 2012). It means the graph database often has complex relationships and structures that can be expressed in high-dimensional representation. Thus, a process of capturing typological information from the graph database is required. This process is called graph embedding, which transforms nodes of the graph into a low-dimensional vector (Grover and Leskovec, 2016). Graph Data Science (GDS) library in Neo4j supports several graph embedding techniques. The authors chose the GraphSAGE algorithm, which can preserve the typology, connectivity, and attributes of the neighboring nodes and relationships as vectors. It is an inductive representation learning algorithm by sampling and aggregating features from a node's local neighborhood (Ahmed et al., 2017). The nodes of the bridges, their properties, and their relationships were projected onto an in-memory graph using the GDS library to use GraphSAGE. After projection, the GraphSAGE algorithm trained the feature properties and returned the embedded vectors. The embedded vector representing the bridges was then imported to Python as a form of the data frame to be fed into the K-means++ clustering algorithm. The authors defined these embedded values as a dataset for the proposed model.

3.3 Clustering

Two datasets were prepared for clustering. One dataset was in the form of a relational database from the original BMS for the base model. Twenty-four variables representing the basic specifications of the bridge were selected and scaled for data standardization. The other dataset was the graph embedding data that historical maintenance information was inherent for the proposed model. Before clustering, the two datasets were analyzed by principal component analysis (PCA), which is the technique to reduce the dimensionality of the dataset while retaining important information (Abdi and Williams, 2010). The authors chose the K-means++ algorithm for clustering, which is a method to minimize the average squared distance between the data points inside the same cluster with a randomized seeding technique (Arthur and Vassilvitskii, 2007). It is known to show more consistent results than K-means by locating the initial centroids far from each other. The number of clusters which is K, was set using the elbow detection algorithm (Kodinariya and Makwana, 2013). The authors set the optimum number of clusters as three, which was the elbow point and also an appropriate number to interpret the result. In the final stage, the bridges assigned at the same cluster can be considered as bridges having similar characteristics.

3.4 Evaluation

The clustering performance was hard to be quantified because it is an unsupervised learning algorithm. The silhouette coefficient of the base model and the proposed model from clustering results can be used to compare the tightness and separation of the cluster (Rousseeuw, 1987).

A silhouette score presents a measure of the proximity of each point in a cluster to points in neighbouring clusters, ranging from -1 to 1. The score calculates the cohesion of a cluster by averaging the distances between one random data point and all other data points within the same cluster, as shown in the equation below. $a^{(i)}$ is the mean intra-cluster distance, and $b^{(i)}$ is the mean nearest-cluster distance for each data point.

$$Silhouette\ coefficient^{(i)} = \frac{b^{(i)} - a^{(i)}}{\max\{b^{(i)}, a^{(i)}\}} \quad (1)$$

However, it is not possible to confirm that the purpose of clustering has been achieved only with this score. Thus, empirical evaluation should be done to validate whether the proposed graph-based clustering model for bridge maintenance cost estimation is acceptable or not. The authors statistically compared each cluster's historical maintenance cost and the characteristics to confirm the clustering performance.

4. Experimental Results

4.1 Bridge Clustering Results

Figure 3 represents the distribution of bridge data transformed in the two-dimensional space. The shape of the cluster did not determine whether the clustering had been successfully achieved as intended. However, some observations were able to be attained. Firstly, in the base model, ‘cluster 1 (Orange color dot)’ exhibited distinct characteristics compared to the other two clusters. Similarly, in the proposed model, ‘cluster 1 (Orange color dot)’ showed clear differentiation from ‘cluster 0 (Blue color dot)’ and ‘cluster 2 (Green color dot)’ in terms of its distinct features.

The average silhouette coefficient supported the superiority of the proposed graph-based model with a score of 0.802 compared to the base model with a score of 0.534. The author interpreted that the reason for the good clustering results of the proposed model was that the bridge information in the graph data was linked to the inspection and maintenance project information.

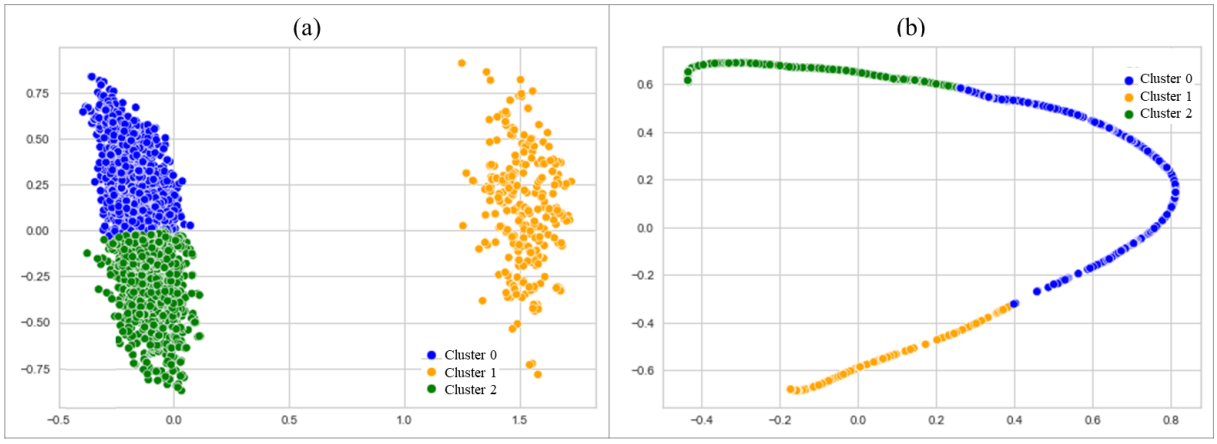

Figure 3: Data Points of Cluster in Two-dimensional Space
: (a) Base Model using Relational Database, (b) Proposed Model using Graph Database

4.2 Evaluation of the Clustering Results with Historical Maintenance Records

The main purpose of this research was to cluster the similar bridges to allocate the limited budget effectively according to their characteristics. 384 bridge data which were merged with their historical repair costs and the previous condition grades from inspection records were used for evaluation of the clustering results. The authors summarized the average unit cost for maintenance work in each bridge in the cluster, as shown in Table 3 and Figure 4. Unit cost meant the total deck bridge repair cost divided by the size. In order to confirm the difference in characteristics except for the size of the bridge, unit cost values were compared.

Table 3: Average Unit Cost ($Unit$: ₩$1{,}000/m^2$) by Condition Grade of Bridge Deck in each Cluster.

Condition grade from previous inspection	Not clustered: (a) in Fig. 4 (Data count)	Base model: (b) in Fig. 4			Proposed model: (c) in Fig. 4		
		Cluster 0 (Data count)	Cluster 1 (Data count)	Cluster 2 (Data count)	Cluster 0 (Data count)	Cluster 1 (Data count)	Cluster 2 (Data count)
A	15.77 (11)	0.74 (4)	-	24.35 (7)	-	20.18 (6)	10.47 (5)
B	18.39 (200)	18.12 (93)	12.29 (2)	18.75 (105)	12.68 (37)	21.38 (110)	16.16 (53)
C	43.84 (151)	41.12 (67)	134.75 (5)	40.40 (79)	44.18 (24)	31.15 (91)	75.71 (36)
D	67.65 (21)	37.53 (7)	-	82.70 (14)	-	50.69 (12)	90.26 (9)
E	93.37 (1)	93.37 (1)	-	-	-	93.37 (1)	-

The clustering results of the proposed model were better for the following reasons. First, in all clusters of the proposed model, a gradually increasing trend of unit cost from 'A (Good)' to 'E (Worst)' condition grades was shown. However, the base model did not show a clear trend by condition grades. Second, the proposed model showed the differences in the unit costs of maintenance works between different clusters. This meant that unnecessary or insufficient cost distribution can be reduced by distributing appropriate maintenance costs according to the characteristics of each bridge. In other words, it was possible to allocate maintenance costs more appropriate to the bridge when the representative value by group was used as shown in (c) than when the average value without clustering was used as the representative value as shown in (a) in Figure 4.

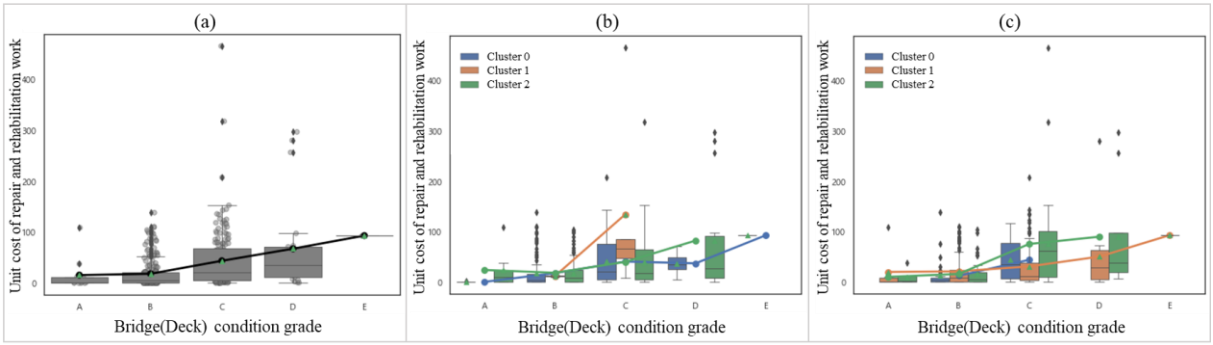

Figure 4: Unit Cost of the Deck Maintenance Work by the Cluster:
(a) Not Clustered, (b) Base Model, (c) Proposed Model

In order to distinguish the representative characteristics of each cluster, the differences were compared by variables, as shown in Figure 5. In the case of the base model, each group did not show any remarkable characteristics between the clusters. However, in the proposed model, the

following interpretation was able to be derived. “The average period of bridge usage in ‘cluster 1’ was less than that of ‘cluster 2’, but it had a higher traffic volume. Therefore, it can be inferred that the unit cost of ‘A’ and ‘B’ grade bridges in ‘cluster 1’ is higher than that of ‘cluster 2’ because traffic control costs were higher.” This kind of knowledge can be inferred from the graph database and can be communicated to decision makers for efficient maintenance.

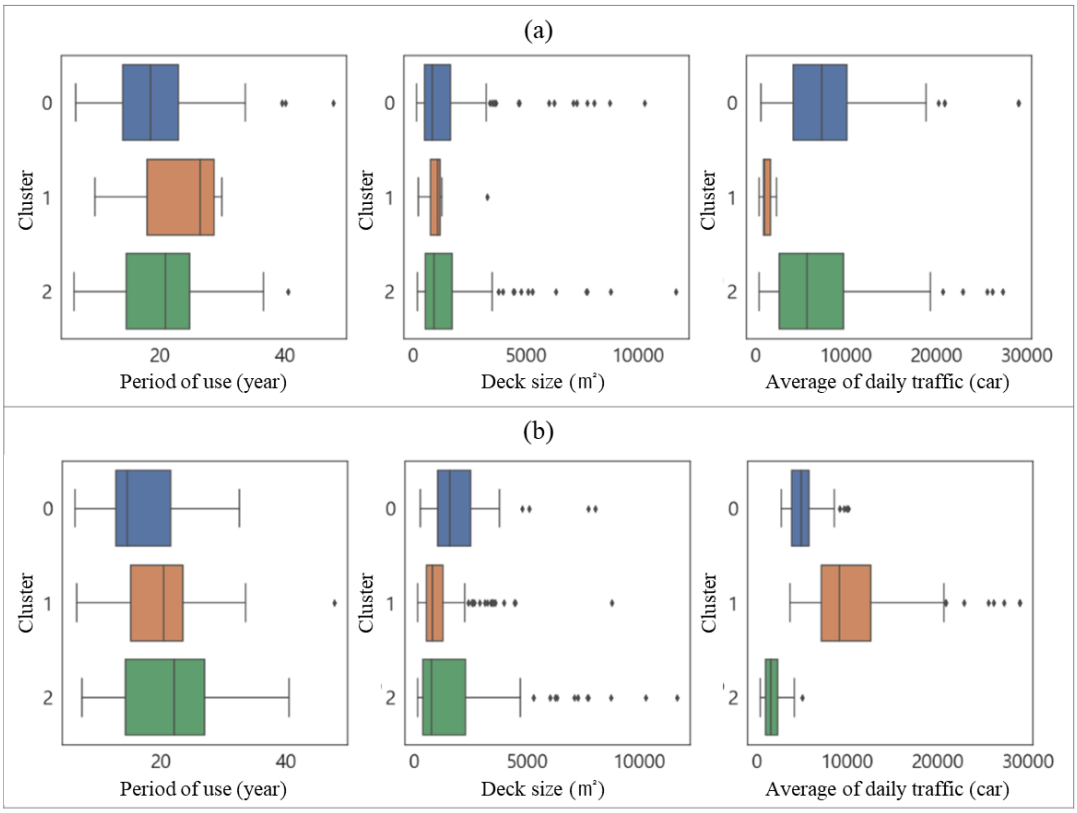

Figure 5: Comparison of between the Cluster: (a) Characteristics of the Bridge by Cluster from Base Model, (b) Characteristics of the Bridge by Cluster from Proposed Model

5. Conclusion

This study proposed a model to improve the clustering performance of bridge maintenance cost estimation by utilizing the graph database. The authors employed the GraphSAGE algorithm to embed nodes and relationships from bridge maintenance records into latent vectors. By clustering bridges with similar characteristics using these embedded vectors, efficient maintenance cost estimation could be supported. The clustering results demonstrated that the significance of converting a relational database into a graph database to leverage knowledge in bridge maintenance cost estimation. For future research, the authors will explore hierarchical clustering methodologies in addition to the K-means clustering methods used in this study. This will allow us to model the hierarchical structural information present in bridge maintenance data. Furthermore, the authors plan to improve maintenance cost estimation by integrating unstructured data, such as bridge inspection reports. The inclusion of such detailed information from disparate data sources will support informed decision-making in practice, enhancing the efficiency and safety of bridge maintenance.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C2003696) and "BK21 PLUS research program" of the National Research Foundation of Korea.

References

Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. 2, 433–459. https://doi.org/10.1002/wics.101

Ahmed, N.K., Rossi, R.A., Zhou, R., Lee, J.B., Kong, X., Willke, T.L., Eldardiry, H., 2017. Inductive Representation Learning in Large Attributed Graphs 1–11.

Angles, R., 2012. A comparison of current graph database models. Proc. - 2012 IEEE 28th Int. Conf. Data Eng. Work. ICDEW 2012 171–177. https://doi.org/10.1109/ICDEW.2012.31

Arthur, D., Vassilvitskii, S., 2007. K-means++: The advantages of careful seeding. Proc. Annu. ACM-SIAM Symp. Discret. Algorithms 07-09-Janu, 1027–1035.

Cheng, Y.M., Leu, S. Sen, 2009. Constraint-based clustering and its applications in construction management. Expert Syst. Appl. 36, 5761–5767. https://doi.org/10.1016/j.eswa.2008.06.100

Fernandes, D., Bernardino, J., 2018. Graph databases comparison: Allegrograph, arangoDB, infinitegraph, Neo4J, and orientDB. DATA 2018 - Proc. 7th Int. Conf. Data Sci. Technol. Appl. 373–380. https://doi.org/10.5220/0006910203730380

Francis, N., Green, A., Guagliardo, P., Libkin, L., Lindaaker, T., Marsault, V., Plantikow, S., Rydberg, M., Selmer, P., Taylor, A., 2018. Cypher: An evolving query language for property graphs. Proc. ACM SIGMOD Int. Conf. Manag. Data 1433–1445. https://doi.org/10.1145/3183713.3190657

Galvan-Nunez, S., Attoh-Okine, N., 2017. Hybrid Particle Swarm Optimization and K-Means Analysis for Bridge Clustering Based on National Bridge Inventory Data. ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A Civ. Eng. 3, 1–6. https://doi.org/10.1061/ajrua6.0000864

Ghahari, S.A., Volovski, M., Alqadhi, S., Alinizzi, M., 2019. Estimation of annual repair expenditure for interstate highway bridges. Infrastruct. Asset Manag. 6, 40–47. https://doi.org/10.1680/jinam.17.00021

Grover, A., Leskovec, J., 2016. Node2vec: Scalable feature learning for networks. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 13-17-Augu, 855–864. https://doi.org/10.1145/2939672.2939754

Kodinariya, T.M., Makwana, P.R., 2013. Review on determining of cluster in K-means. Int. J. Adv. Res. Comput. Sci. Manag. Stud. 1, 90–95.

Liu, K., El-Gohary, N., 2022. Improved similarity assessment and spectral clustering for unsupervised linking of data extracted from bridge inspection reports. Adv. Eng. Informatics 51, 101496. https://doi.org/10.1016/j.aei.2021.101496

Miyamoto, A., Kawamura, K., Nakamura, H., 2000. Bridge management system and maintenance optimization for existing bridges. Comput. Civ. Infrastruct. Eng. 15, 45–55. https://doi.org/10.1111/0885-9507.00170

Nayak, A., Poriya, A., Poojary, D., 2013. Type of NOSQL databases and its comparison with relational databases. Int. J. Appl. Inf. Syst. 5, 16–19.

Radovic, M., Ghonima, O., Schumacher, T., 2017. Data Mining of Bridge Concrete Deck Parameters in the National Bridge Inventory by Two-Step Cluster Analysis. ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A Civ. Eng. 3, 1–9. https://doi.org/10.1061/ajrua6.0000889

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sun, J.W., Park, K.H., Lee, Y.J., 2020. Analysis of Unit Maintenance Cost according to Bridge Safety Grade. Proc. Korea Concr. Inst. Conf. 32(1), 251–252.

Tiwary, K., Patro, S., Sahoo, B., 2022. Bridgebase: A Knowledge Graph Framework for Monitoring and Analysis of Bridges, in: Proceedings of the Canadian Society of Civil Engineering Annual Conference 2021. https://doi.org/https://doi.org/10.1007/978-981-19-0656-5_34

Wu, C., Wu, P., Wang, J., Jiang, R., Chen, M., Wang, X., 2021. Critical review of data-driven decision-making in bridge operation and maintenance. Struct. Infrastruct. Eng. 18, 47–70. https://doi.org/10.1080/15732479.2020.1833946