Journal of The Korea Institute of Building Construction
Vol. 24, No. 1, pp. 121-131 / February, 2024
https://doi.org/10.5345/JKIBC.2024.24.1.121

pISSN 1598-2033
eISSN 2233-5706
www.jkibc.org

Research Paper

비정형 공사감리문서 정보와 이항 로지스틱 회귀분석을 이용한 건축 현장 비용성과 평가 프레임워크 개발

Cost Performance Evaluation Framework through Analysis of Unstructured Construction Supervision Documents using Binomial Logistic Regression

김창원[1] · 송태근[2] · 이기석[2] · 유위성[3]*
Kim, Chang-Won[1] · Song, Taegeun[2] · Lee, Kiseok[2] · Yoo, Wi Sung[3]*

[1]Associate Research Fellow, Innovation Procurement Research Center, Korea Institute of Procurement, Gangnam-Gu, Seoul, 06231, Korea
[2]Researcher, Department of Construction Economic & Finance Research, Construction & Economy Research Institute of Korea, Seoul, 06050, Korea
[3]Research Fellow, Department of Construction Economic & Finance Research, Construction & Economy Research Institute of Korea, Seoul, 06050, Korea

*Corresponding author
Yoo, Wi Sung
Tel : 82-2-3441-0860
E-mail : wsyoo@cerik.re.kr

Received : December 27, 2023
Revised : January 23, 2024
Accepted : January 31, 2024

ABSTRACT

This research explores the potential of leveraging unstructured data from construction supervision documents, which contain detailed inspection insights from independent third-party monitors of building construction processes. With the evolution of analytical methodologies, such unstructured data has been recognized as a valuable source of information, offering diverse insights. The study introduces a framework designed to assess cost performance by applying advanced analytical methods to the unstructured data found in final construction supervision reports. Specifically, key phrases were identified using text mining and social network analysis techniques, and these phrases were then analyzed through binomial logistic regression to assess cost performance. The study found that predictions of cost performance based on unstructured data from supervision documents achieved an accuracy rate of approximately 73%. The findings of this research are anticipated to serve as a foundational resource for analyzing various forms of unstructured data generated within the construction sector in future projects.

Keywords : construction supervision document, unstructured data, building construction site, cost performance evaluation framework, binomial logistic regression

1. 서 론

1.1 연구 배경 및 목적

프로젝트의 성공적 완수를 위해 일정, 비용, 품질, 안전 등에 대한 성과관리가 이루어지며, 이 중 비용성과는 모든 이해관계자가 공통적으로 관심을 갖는 주요요인으로 평가된다[1-5]. 이에 선행 연구들에서는 정량 데이터를 기초로 다양한 분석방법론을 이용한 객관적 성과측정 및 평가모형을 연구결과로 제시하고 있다[2-8]. 그러나 이와 같은 정량 평가체계는 단위공

정별 예산 및 소요 비용 등과 같은 데이터의 확보가 전제되어야 하나, 해당 데이터는 시공사의 영업비밀이라는 인식에 따라 정보취득에 현실적 한계가 존재한다[2,9,10]. 또한 데이터 수집 단계에서 설계변경에 따른 예산조정 결과 미반영 등과 같은 오류 발생 시 분석결과의 신뢰성 저하라는 부정적 결과를 야기할 수 있다.

이와 같은 상황에서 시공사와 독립된 제3의 위치에서 생산과정을 관리·감독하여 프로젝트의 성공적 완수 및 고품질 결과물 창출을 지원하는 주요업무인 공사감리 과정에서 생성되는 다양한 문서 내 비정형의 정성적 정보는 비용성과를 평가할 수 있는 주요 데이터로 활용할 수 있다. 공사감리는 「건축법」, 「주택법」 등의 관련 법령에 근거하여 대상사업 및 업무범위를 규정하고 있으며, 설계도서와 시공결과의 정합성, 공정·안전·품질관리 활동 등을 지도·감독하는 업무로 정의된다[11-13]. 실제 업무 수행을 통해 공사감리계획서, 감리일지, 안전 및 환경관리계획서, 비용 관리 대장, 시공 검토서, 최종감리보고서 등과 같은 다양한 문서가 생성되며, 특히 최종감리보고서는 공사 전반에 대한 종합적인 점검의견이라는 주요한 질적 정보를 포함하고 있다[14]. 이와 같은 비정형 정보 기반의 분석은 정량 데이터를 기초로 하는 전통적인 통계기법의 활용으로 유의미한 결과 도출이 불가능하였으나, 텍스트마이닝, 사회연결망분석(Social Network Analysis, SNA)과 같은 고도화된 방법론의 등장을 통해 다양한 시사점의 도출을 지원할 수 있다[15-19].

이에 본 연구는 건축 현장에서 생성된 최종감리보고서를 활용하여 프로젝트의 비용성과 평가 및 분석을 수행할 수 있는 프레임워크 제시를 목적으로 한다. 본 연구에서 제시하는 프레임워크는 최종감리보고서에 제시된 비용성과 관련 내용에 대해 텍스트마이닝 및 SNA를 적용하여 주요 키워드를 도출하고, 이항 로지스틱 회귀분석 기반 비용성과 평가체계로 구성된다. 본 연구의 결과는 프로젝트의 다양한 성과지표에 대한 비정형 데이터 기반 평가 프레임워크 마련을 위한 기초자료로 활용이 가능할 것으로 예상된다.

1.2 연구 범위 및 방법

본 연구는 「건축법」이 적용되는 민간 건축현장에서 생성되는 공사감리 문서를 활용한 비용성과 평가 프레임워크 구축으로 범위를 한정하였다. 또한 프로젝트의 종합적인 비용성과를 평가하기 위해서는 계획, 설계, 시공, 유지관리 등 전 생애주기에 걸친 데이터 수집이 필요하나, 해당 데이터 취득은 현실적으로 한계가 존재한다는 점을 고려하여 실질적인 생산이 이루어지는 동시에 투입비용 규모가 가장 큰 시공단계로 한정하였다. 연구의 수행 방법은 다음과 같다.

첫째, 예비적 고찰을 통해 국내 공사감리업무 대상 및 세부업무 구성을 분석하고, 본 연구의 결과인 프레임워크의 개발에 활용된 방법론인 연관규칙 분석 및 SNA, 이항 로지스틱 회귀분석에 대해 검토하였다.

둘째, 감리보고서 내에 포함된 주요 비정형 정보와 다양한 분석기법을 이용한 프레임워크 개발 절차 및 결과를 제시하였다. 세부적으로는 프레임워크 개발을 위한 기초자료인 비용성과 관련 주요 키워드 도출과 이항 로지스틱 회귀분석을 이용한 성과평가 모형 구축 및 정확도 검증 단계로 구분된다. 세부적으로 살펴보면, 우선 중점 키워드의 도출은 감리보고서 내 포함된 비용성과 관련 텍스트를 추출하고, 텍스트에 포함된 키워드 노출 빈도를 기준으로 설정하였다. 다음으로 도출된 키워드를 대상으로 연관규칙 분석 및 SNA를 활용하여 키워드 간 네트워크 구조를 분석하여 건축 현장의 비용성과에 높은 영향력을 미칠 수 있는 주요 키워드들을 선정하였다. 마지막으로 선정된 주요 키워드를 독립변수로 활용하고, 계획 대비 비용을 종속변수로 활용하는 이항 로지스틱 회귀모형 기반 성과평가 포형을 구축하고, 이에 대한 검증을 수행하였다. 이 때, 종속변수는 계획 대비 실적이 100% 이상이면 1, 100% 미만이면 0으로 더미(dummy)로 변환하여 활용하였다.

셋째, 본 연구에서 제시한 비정형 정보 기반 건축 현장의 비용성과 평가 프레임워크 개발을 통해 도출된 시사점에 대해 논의하였다.

2. 예비적 고찰

2.1 국내 건축공사감리 수행체계

국내 감리업무는 건설공사, 소방공사, 전기공사, 정보통신공사 등을 대상으로 운영되고 있으며, 이 중 본 연구에서 대상으로 설정한 건축 프로젝트는 건설공사의 일환으로 분류된다. 건축 프로젝트에 대한 감리체계를 세부적으로 살펴보면, 생애주기 기준으로는 공사 전, 공사 중, 공사 완료로 구분하고 있으며, 건축물 유형 및 규모에 따른 업무의 수행방식은 비상주감리, 상주감리, 책임감리로 분류된다. 즉, 건축 프로젝트 감리업무 수행체계는 생애주기, 건축물 유형 및 규모에 따라 차별화되어 운영되고 있는 것으로 검토할 수 있다. 예를 들어 건축 프로젝트의 공사 중 단계에서 상주감리는 바닥면적 5,000m^2 이상 건축공사(축사 또는 작물재배사의 건축공사 제외), 아파트 건축공사 등으로 구분되며, 책임감리는 16층 이상 또는 바닥면적 합계 5,000m^2 이상의 다중이용건축물(문화 및 집회시설, 판매시설, 의료시설 중 종합병원 등)로 규정되어 있다.

건축 프로젝트 대상 세부적인 감리업무 내용은 국토교통부 고시 제2020-11호(건축공사감리세부기준)에 제시되어 있으며, 설계도서와의 적합성, 공정·품질·안전관리의 적정성 등과 같이 약 50여개의 업무로 구성되어 있다. 또한 해당 업무들은 기본 및 기본 외 업무로 구분하여 운영되고 있다는 것이 고유한 특성의 하나라 할 수 있다. 즉, 건축공사감리업무는 관련 법령에서 규정한 대상 및 업무 내용을 기준으로 수행된다는 특성을 보유하고 있으며, 전술한 바와 같이 제3의 객관적 입장에서 프로젝트의 수행과정을 모니터링한 결과 문서를 생성한다는 점 고려시 본 연구의 범위인 비용성과를 추정할 수 있는 프레임워크 개발을 위한 유의미한 데이터로서 활용이 가능할 것으로 예상된다.

2.2 연관규칙분석과 SNA를 이용한 주요 키워드 추출

본 연구는 데이터마이닝 기법 중 데이터 속상 간 패턴 등 연관성을 도출하는 연관규칙 분석(ARA, Association Rule Analysis)과 SNA를 이용하여 주요 키워드 간 연관성 및 네트워크 체계를 분석하였다.

ARA는 특정 속성(attribute)이 발생할 때 다른 속성이 발생하는 경향을 나타내는 규칙을 찾는 분석 방법이다[20]. 이 방법론은 대규모 데이터에서 유용한 정보를 추출하고, 연관 패턴을 발견하는 데 큰 장점이 있어, 의료, 금융, 유통 등 다양한 산업에서 활용되고 있으며, 최근 건설산업에서는 텍스트 자료의 분석에 널리 이용되고 있다[21-25]. 이에 본 연구는 비용 성과에 영향을 미칠 수 있는 감리보고서 내 키워드를 선정하는 단계에서 ARA를 활용하였다.

또한 SNA는 비용성과 평가 및 추정에 활용되는 키워드 관계를 노드와 링크를 통해 네트워크 형태로 구축하기 위한 목적으로 활용하였다. 네트워크의 특성을 확인할 수 있는 지표들은 평균연결정도(degree), 밀도(density), 포괄성(inclusiveness) 등이 있으며, 네트워크의 구조적인 특징을 파악할 수 있는 대표 지표는 중심성(centrality)으로 정의된다. 중심성을 평가할 수 있는 지표로는 매개 중심성(betweenness centrality), 아이젠 벡터 중심성(eigen vector centrality), 연결 중심성(degree centrality), 페이지 랭크 중심성(page rank centrality)등이 있으며, 본 연구에서는 SNA에서 보편적으로 활용되는 네트워크 내 노드와 연결된 링크의 수로 산정되는 연결 중심성을 활용하였다.

2.3 이항 로지스틱 회귀모형

다양한 산업에서 데이터의 분류 기법으로 널리 활용되고 있는 이항 로지스틱 회귀(Binominal Logistic Regression, BLR) 모형은 일반적으로 2개의 분류 그룹으로 구성된다. 로지스틱 회귀 모형은 목적 값 분류에 적용될 수 있는 기법으로, 선형 회귀와 유사한 결과를 해석하는 데 유용하다.

이항 로지스틱 회귀모형은 회귀분석과 클래스 분류에 적용할 수 있는 모형으로, 종속변수가 2개(0과 1) 클래스를 갖는 분포를 띠고 그 모수가 독립변수에 의존한다. 종속변수(y)가 0 또는 1인 분류에서 독립변수(x)값을 이용하여 $\mu(y=1|x1, x2,$

x3...)를 구한 후 식 (1)에 따라 y값의 클래스를 분류한다. 여기서 모수 μ가 x의 함수라고 가정하여 0과 1 사이의 값을 생성하는 로지스틱 함수(logistic function)로 변환하며, 그 과정은 다음과 같다. 먼저 1이 나올 확률 μ와 0이 나올 확률 1-μ의 비율인 오즈(odds)=μ / (1-μ)를 구하고, 오즈(odds)를 로그변환하여 로짓(z)=log(μ / (1-μ))을 구한다. 마지막으로 로짓(z)의 역함수가 로지스틱함수이다. 로짓(z)은 가중치(β)의 선형결합으로 이루어진다.

$$y_{predict} = \begin{cases} 1 & \text{if } \mu \geq 0.5 \\ 0 & \text{if } \mu < 0.5 \end{cases} \quad (1)$$

$$\mu(z) = \frac{1}{1 + \exp(-z)}$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots \beta_n x_n$$

$y_{predict}$: 이항 로지스틱 회귀모형 예측 결과(클래스)
$\mu(z)$: 클래스 1이 나올 확률(로지스틱함수)
z : 독립변수의 가중합(로짓)

3. 공사감리문서 기반 건축현장 비용성과 평가 프레임워크

3.1 프레임워크 개발 절차

본 연구에서 제시하는 건축현장 비용성과 평가 프레임워크는 건축법 기반의 공사감리업무에서 생성되는 최종감리보고서를 기반으로 구축되었다. 이는 해당 보고서에 기술된 종합의견 및 감리업무 평가내용은 공사에 대한 전반적인 점검 결과와 감리원의 의견이 함축되어 있어, 건축현장의 비용성과를 추정할 수 있는 효율적인 질적 정보를 제공한다는 점을 고려한 것이다. 본 연구는 43개 건축현장을 대상으로 수집한 공사감리보고서 중 포괄적인 점검 결과 및 종합의견을 포함하고 있는 39개의 건축현장의 보고서를 최종 분석 대상으로 설정하였다.

세부적인 프레임워크 구축 절차는 다음과 같다. 우선 최종감리보고서 내 점검 결과 및 종합 의견에 등장하는 단어(명사)의 빈도를 구하고, 이를 통해 건축현장의 성과와 관련된 124개의 키워드를 선정하였다. 다음으로, 1개 문장에 등장하는 단어 집합이 1개 행으로 구성된 데이터셋을 구축하여 ARA 및 SNA를 수행하였다. 이와 같은 분석을 통해 비용성과와 관련이 높을 것으로 판단되는 11개의 키워드 선정하고, 키워드 간 비중(%) 자료를 생성하여 BLR에 활용하였다. BLR 수행 시 완전 공선성(perfect multicollinearity)에 관한 문제를 피하고자 11개 키워드 중 '시공' 키워드를 제외하였다.

3.2 ARA 및 SNA 기반 주요 키워드 추출

주요 키워드들의 도출을 위해 본 연구는 파이썬의 TIKA 라이브러리를 이용하여 공사감리 종합의견 텍스트를 추출하였다. 또한 KoNLPy 패키지의 OKT 모듈을 이용해서 단어(명사)를 추출한 후 문장 단위로 묶은 자료 형태로 구성하였는데, 이는 구축 자료를 이용하여 ARA를 수행하기 위한 전처리 과정에 효과적이다[26]. 낮은 빈도를 보이는 단어의 경우 비용성과와 무관하거나 의미 파악이 힘들어 빈도를 기준으로 상위 124개 키워드를 선정하였다. Table 1은 전체 빈도의 약 70%를 차지하는 37개 키워드를 내림차순 정렬하여 나타낸 것으로, 빈도는 '관리', '품질', '계획', '안전', '시공', '자재', '공정', '점검' 등의 순위로 나타났다.

Table 1. Extraction results of high-frequency keywords

Frequency ranking	Keyword	Frequency ranking	Keyword	Frequency ranking	Keyword	Frequency ranking	Keyword
1	Management	11	Test	21	Secure	31	Perfect
2	Quality	12	Progress	22	Complete	32	Fine
3	Plan	13	Occurrence	23	Completion	33	Start
4	Safety	14	Matters	24	Whether	34	Situation
5	Construction	15	In advance	25	Enforcement	35	Detail
6	Materials	16	Change	26	Improvement	36	Inspection
7	Process	17	Design	27	Complains	37	Defects
8	Check	18	Approval	28	Environment		
9	Review	19	Use	29	Effort		
10	Confirmation	20	Prevention	30	Out		

Table 2. Analysis results for high-frequency keywords using advanced ranking algorithms

No. of rules	Antecedent event	Subsequent event	Support	Confidence	Lift
1	Quality	Management	0.16	0.66	1.36
2	Process	Management	0.12	0.67	1.38
3	Plan	Management	0.12	0.58	1.20
4	Safety	Management	0.12	0.68	1.41
5	Construction	Management	0.11	0.56	1.15
6	Materials	Management	0.09	0.61	1.25
7	Plan	Quality	0.09	0.41	1.71
8	Materials	Quality	0.08	0.52	2.15
9	Review	Plan	0.07	0.49	2.38
10	Test	Quality	0.07	0.79	3.26
11	Confirmation	Quality	0.07	0.45	1.84
12	Secure	Quality	0.06	0.84	3.47
13	Construction	Quality	0.06	0.32	1.30
14	Review	Construction	0.06	0.39	2.03
15	Inspection	Safety	0.06	0.44	2.56
16	Occurrence	Management	0.05	0.53	1.10
17	In advance	Plan	0.05	0.48	2.30
18	Process	Quality	0.05	0.28	1.15
19	Check	Construction	0.05	0.32	1.65
20	Test	Management	0.05	0.57	1.17
21	Process	Plan	0.05	0.26	1.28
22	Construction	Plan	0.05	0.24	1.18
23	Review	Quality	0.04	0.31	1.29
24	Completion	Management	0.04	0.53	1.10
25	Progress	Process	0.04	0.40	2.20
26	Review	Check	0.04	0.29	1.94
27	Check	Plan	0.04	0.28	1.34
28	Complete	Management	0.04	0.52	1.06
29	Inspection	Plan	0.04	0.33	1.61
30	Use	Materials	0.04	0.56	3.76

이와 같이 선정된 124개 키워드만 포함해서 문장 단위로 구축한 데이터 이용하여 ARA를 수행한 결과에서 지지도를 기준으로 상위 30개 규칙의 경우 높은 신뢰도를 보이는 결과는 Table 2와 같다. 세부적으로 살펴보면, 선행사건으로 '품질', 후행사건으로 '관리'의 신뢰도가 선행사건으로 '관리', 후행사건으로 '품질'의 경우보다 더 높은 경우에는 전자의 조합만 표에 제시하였다. 높은 지지도 값을 갖는 규칙 중 '관리', '품질', '계획', '자재' 등과 같은 키워드가 포함된 연관규칙이 빈번하게 생성되었다. '관리', '품질', '계획' 키워드의 경우 후행사건으로 생성되는 규칙이 빈번하였으며, '자재' 키워드는 주로 선행사건으로 규칙이 생성되었다. 이때, 규칙 생성 기준으로 지지도, 신뢰도, 향상도를 각각 0.01, 0.1, 1 이상으로 설정하였다. 이와 같은 분석 결과를 이용한 SNA 수행 결과는 Figure 1과 같다. 그 결과, '관리', '품질', '계획' 등의 중심성이 가장 높았으며, 다음으로 '시공', '자재', '확인', '검토', '공정' 등 키워드에서 중심성이 높게 분석되었다.

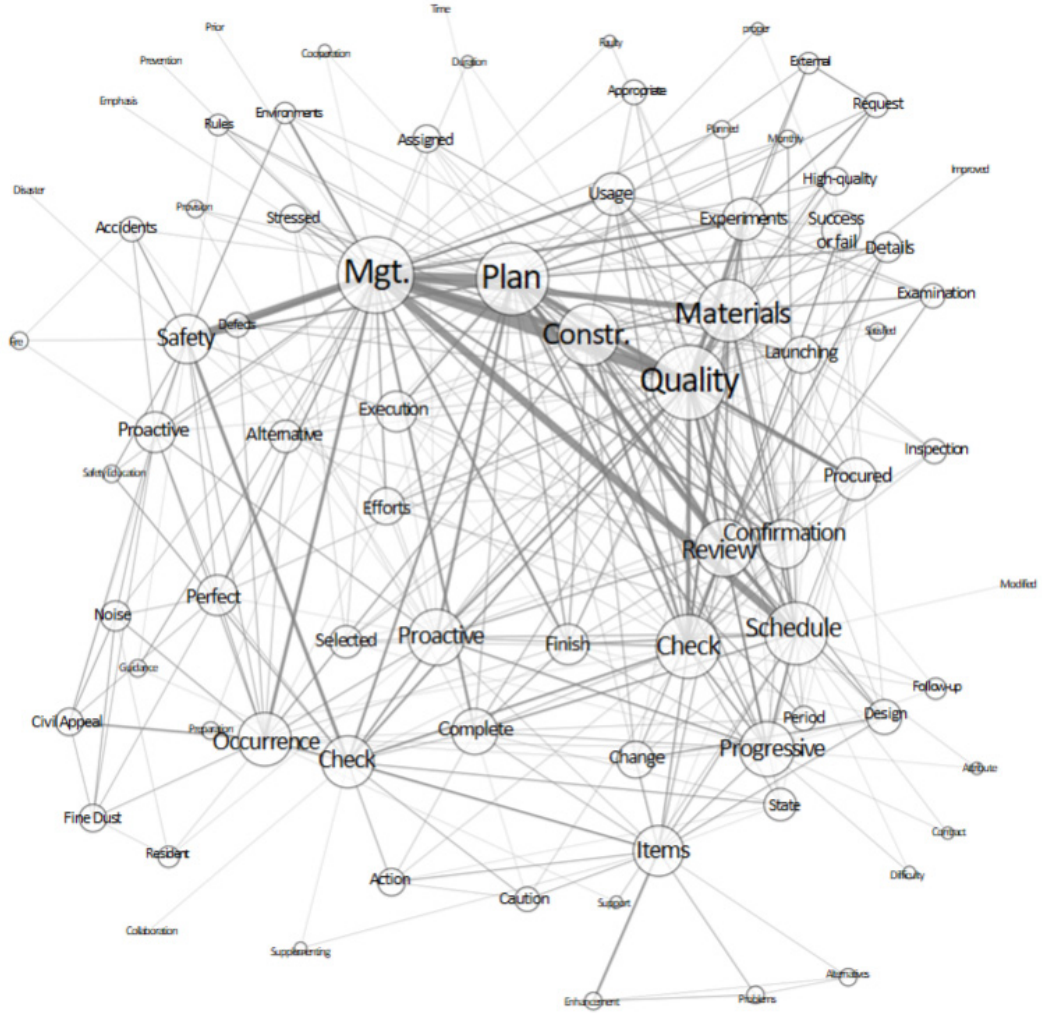

Figure 1. Social network analysis(SNA) depicting interconnections among main keywords

ARA 및 SNA 결과를 활용하여 본 연구는 '검토', '계획', '공정', '관리', '변경', '사전', '안전', '자재', '점검', '품질', '시공' 등 11개 단어(명사)를 비용성과 평가를 추정할 수 있는 모델의 입력값으로 선정하였다. 여기서 '시공'을 제외한 10개 키워드 간 비중(%)을 독립변수로, 계획 대비 실제 비용을 비용성과를 평가하는 종속변수로 활용하여 이항 로지스틱 회귀모형을 구축하였다. 이때 종속변수는 100% 미만이면 0, 100% 이상이면 1로 이진 변환하여 활용하였다.

3.3 BLR 기반 분석 결과

본 연구에서 독립변수로 활용되는 10개 키워드의 비중에 대한 기술통계량은 Table 3과 같다. 분석 결과, 39개 건축현장에 대한 비용성과 평균은 101.5이고 중위수는 103.3으로, 왼쪽으로 꼬리가 긴 형태(왜도 −2.3)의 분포를 보였다. 키워드의 평균 비중은 '관리', '품질', '공정', '안전', '계획', '자재' 등의 순으로 나타났다. '변경'의 첨도는 21.4로 매우 높게 나타났으며, 1사분위수 값이 0인 것은 데이터의 상당수가 0인 값을 가지며, 0 부근에 분포된 것으로 해석할 수 있다. '변경'을 제외하면 다른 키워드의 비중은 상대적으로 값이 고르게 분포하는 것을 알 수 있다.

본 연구는 위와 같이 선정된 키워드의 비중을 독립변수(x)로 두고 식 (2)와 같이 이항 로지스틱 회귀모형을 기반으로 비용성과를 예측하였으며, 이 때, 회귀계수 추정은 L2 규제가 적용된 모형을 활용하였다. 또한, 데이터를 학습 및 테스트 데이터셋으로 분리한 후, 학습데이터를 이용해 LOOCV(Leave-One-Out Cross-Validation) 샘플링 방법으로 회귀계수를 추정하고 평균값을 산정하였다. 도출된 회귀 모델을 테스트 샘플에 적용하는 것을 100회 반복하여 최종 예측 정확도를 평가하였다. 식 (2)에서 z는 로그 오즈(log-odds) 또는 로짓(logit)으로 정의되는 값으로, 로지스틱 회귀계수의 선형결합으로 이루어진다.

Table 3. Descriptive statistics of main keywords' distribution

Variable	Average	First quartile	Median	Third quartile	Max	Standard deviation	Kurtosis	Skewness
Cost performance	101.5	95.8	103.3	108.5	137.9	16.5	11.8	-2.3
Review	6.4	3	5.3	9.3	16.7	4.8	-0.3	0.7
Plan	7.8	5.2	7.8	9.8	26.1	5	3.6	1
Process	9.4	5.6	8.1	10.8	30.2	6.1	4.5	1.9
Management	26.5	21.6	23.3	29.4	60	8.5	5.9	2
Change	4.2	0	1.5	5.2	48	8.3	21.4	4.2
In advance	3.4	1.4	3.2	5.7	7.2	2.6	-1.4	0.1
Safety	8.4	5.4	9.3	10.5	20.6	4.1	1	0.2
Materials	7.6	4.9	7.4	10.8	22.4	5.2	0.4	0.4
Inspection	5.1	0	3.9	10.4	15.8	4.9	-1.3	0.4
Quality	11.5	8.2	11.7	13.9	29.2	5.4	2.4	0.6

$$y_{predict} = \begin{cases} 1 & \text{if } 1/(1+e^{-z}) \geq 0.5 \\ 0 & \text{if } 1/(1+e^{-z}) < 0.5 \end{cases} \quad (2)$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10}$$

Table 4는 규제상수(L2)에 따른 이항 로지스틱 회귀모형의 비용성과 예측 정확도와 로그로스(log-loss)를 분석한 결과이다. 로그로스는 모델의 예측확률(p)과 실제 클래스 y사이의 로그 개연성(log likelihood)를 측정한 것으로 예측확률이 실제 클래스에 얼마나 가까운지를 나타낸다. 규제 상수 0.01~0.1일 때 예측 정확도는 0.73~0.74로 가장 높게 나타났고, 0.001~0.1에서 낮은 로그로스 값을 보였다. 파이썬에 로지스틱 회귀분석 모듈에 규제의 역수를 입력하여, Table 4에 제시된 규제상수 0.001~1000은 실제 모형에서 규제 1000~0.001에 해당한다. 따라서 규제상수 0.001을 제외하면, 실제 모형에서 규제 값이 커질수록 예측 정확도와 로그로스가 향상되었다.

Table 4. Comparison of log loss metrics versus binomial logistic regression model accuracy

	0.001	0.01	0.1	1	10	100	1000
Prediction accuracy	0.60	0.73	0.74	0.70	0.68	0.66	0.66
Log loss	0.64	0.61	0.65	0.84	1.25	1.61	1.89

위와 같은 단계를 거쳐 구축할 수 있는 비정형 데이터 기반 건축현장 비용성과 평가 프레임워크의 개념은 Figure 2와 같이 도식화할 수 있다. 프레임워크의 구성은 데이터 수집, 분석 데이터 설정 및 모델 구축 단계로 정리할 수 있으며, 이에 대한 세부적인 내용은 다음과 같다. 첫째, 데이터 수집 단계에서는 공사감리보고서를 수집하고, 해당 보고서 내 비용성과와 관련된 텍스트 및 키워드 추출, 노출 빈도에 따른 주요 키워드 선정 및 DB화가 수행된다. 이와 같은 단계에서 추출된 관련 문장에서 도출되는 키워드는 명사 형태로 수행되어야 한다는 것이 주요 특징이라 할 수 있다. 둘째, 분석을 위한 데이터 설정 단계는 첫 번째 단계에서 수립된 DB를 근거로 ARA 및 SNA를 이용한 키워드별 네트워크 구조 분석, 네트워크 분석 결과에 기반하여 키워드별 선행 및 후행 구조 검토, 검토 결과에 기반한 최종 분석 데이터 설정으로 구성된다. 마지막으로 모델 구축 단계는 두 번째 단계에서 도출된 최종 분석 키워드들 대상 구성 비중으로 구성된 독립 변수의 설정, 계획 대비 실적을 이진수로 변환하여 더미변수로 설정한 종속변수의 설정, 이항 로지스틱 회귀모형을 적용한 모델 구축 및 해당 모델을 이용한 건축현장 비용 성과 평가 및 진단의 단계로 구성된다. 특히 모델 대상 정확도 등에 대한 진단은 전술한 바와 같이 규제상수(L2)에 따른 확률적 측면에서의 분석이 이루어진다는 것이 특징이라 할 수 있다.

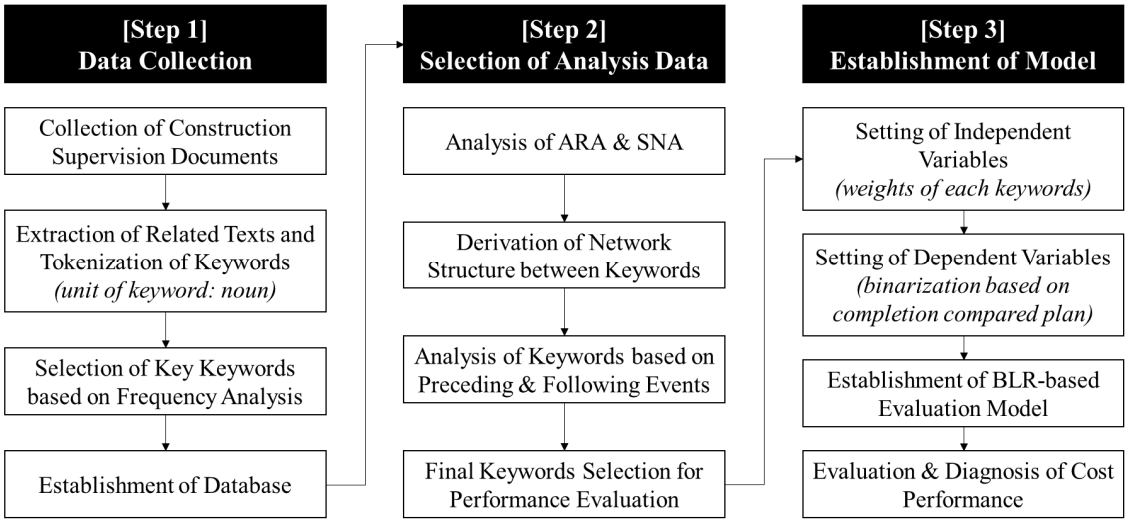

Figure 2. Conceptual framework for evaluating cost performance

4. 결 론

본 연구는 건축현장에서 수행되는 감리업무를 통해 생성되는 최종보고서 내 비정형 데이터를 대상으로 다양한 분석기법을 적용하여 비용성과를 평가할 수 있는 프레임워크를 제시하였다. 프레임워크를 이용한 분석 결과, 약 74% 수준에서의 정확도가 도출되었으며, 이는 제한된 표본 수를 고려한다면 상대적으로 높은 수준으로 해석할 수 있다. 단, 본 연구에서 제시한 프레임워크의 효용성 및 분석결과의 현실성 확보 등을 위해 다음과 같은 추가적인 연구가 필요할 것으로 예상된다.

첫째, 향후 본 연구에서 제시된 프레임워크의 범용성 및 효용성 제고를 위해서는 데이터의 규모가 확대되어야 할 것으로

판단된다. 데이터의 규모는 통계적 모형으로 도출한 모수 추정 결과에 영향을 미칠 수 밖에 없으며, 따라서 수집되는 데이터 규모의 확대는 프레임워크의 성능과 직결된다고 할 수 있다. 따라서 본 연구는 39개 현장의 감리보고서를 대상으로 하였으나, 향후 이를 확장할 시 다양한 건축물 유형 등을 기준으로 비용성과를 추정할 수 있는 프레임워크의 개발이 가능할 것으로 예상된다.

둘째, 데이터 수집시 감리보고서는 각기 다른 양식체계로 기술되어 있어 자료의 신뢰성 부족이라는 잠재적 한계가 있는 것으로 검토되었다. 감리 대상 사업 및 업무 내용은 관련 법령에서 규정하고 있으나, 해당 업무의 최종 성과품은 각기 다른 양식체계로 작성된다는 것은 정책 측면에서의 개선이 필요한 사항으로 예상되며, 이와 같은 개선방안 마련 시 정량 데이터 기반 평가모형이라는 전통적 성과평가 방법에서 비정형 데이터의 활용을 통한 다양한 시사점 도출을 지원할 수 있을 것으로 예상된다.

마지막으로 본 연구에서 제시한 프레임워크와 기존의 설계단계 비용산정 및 예측 중심의 성과체계가 통합되면 단일 건축현장의 착수부터 완공까지 전 생애주기에 걸친 비용성과를 연속적으로 평가할 수 있을 것으로 예상된다. 또한 이와 같은 체계의 마련은 비용성과 뿐만 아니라 공기, 안전 등의 성과를 측정할 수 있는 효율적인 대안으로 확장될 수 있으며, 궁극적으로 비정형 데이터를 이용한 건설산업의 성과 및 경쟁력 추이를 분석할 수 있는 모델로 활용이 가능할 것으로 사료된다.

요 약

공사감리문서는 프로젝트의 수행과정을 제3의 독립적인 위치에서 모니터링한 종합적인 점검의견이라는 주요한 비정형 정보를 제공할 수 있다. 이와 같은 비정형 정보는 최근 분석방법론의 고도화에 따라 다양한 시사점을 제공할 수 있는 유의미한 자료로 평가받고 있다. 이에 본 연구는 건축공사의 최종 감리보고서 내 비정형 데이터를 대상으로 다양한 방법론을 활용하여 비용성과를 평가할 수 있는 프레임워크를 제시하였다. 세부적으로는 텍스트마이닝과 사회연결망분석을 통해 감리보고서 내 주요 키워드들을 도출하고, 해당 데이터들을 이항 로지스틱 회귀분석을 통해 분석하여 비용성과를 평가하였다. 그 결과, 감리보고서 내 비정형 데이터를 이용하여 추정된 비용성과 예측 정확도는 약 73% 수준으로 높게 도출되었다. 본 연구의 결과는 향후 건설산업에서 발생되는 다양한 비정형 데이터의 분석을 위한 기초자료로 활용이 가능할 것으로 예상된다.

키워드 : 공사감리문서, 비정형 데이터, 건축공사 현장, 비용성과 평가 프레임워크, 이항 로지스틱 회귀분석

Funding

This research was supported by a grant(RS-2022-00143493, project number:1615012983) from Digital-Based Building Construction and Safety Supervision Technology Research Program funded by Ministry of Land, Infrastructure and Transport of Korean Government.

Acknowledgement

This research was developed and submitted to a proceeding paper presented at the 2023 Spring Conference of the Korean Institute of Building Construction.

ORCID

Chang-Won Kim, https://orcid.org/0000-0002-0002-1421
Taegeun Song, https://orcid.org//0009-0003-7941-5128
Kiseok Lee, https://orcid.org/0000-0003-3207-4058
Wi Sung Yoo, https://orcid.org/0000-0001-9284-3918

References

1. Baloi D, Price AD. Modeling global risk factors affecting construction cost performance. International Journal of Project Management. 2003 May;21(4):261-9. https://doi.org/10.1016/S0263-7863(02)00017-0
2. Kim CW, Yoo WS, Lim H, Yu I, Cho H, Kang KI. Early-warning performance monitoring system (EPMS) using the business information of a project. International Journal of Project Management. 2018 Jul;36(5):730-43. https://doi.org/10.1016/j.ijproman.2018.03.010
3. Luong DL, Tran DH, Nguyen PT. Optimizing multi-mode time-cost-quality trade-off of construction project using opposition multiple objective difference evolution. International Journal of Construction Management. 2021;21(3):271-83. https://doi.org/10.1080/15623599.2018.1526630
4. Nguyen DT, Le-Hoai L, Tarigan PB, Tran DH. Tradeoff time cost quality in repetitive construction project using fuzzy logic approach and symbiotic organism search algorithm. Alexandria Engineering Journal. 2022 Feb;61(2):1499-518. https://doi.org/10.1016/j.aej.2021.06.058
5. Lim HS, Seo JH, Yoo WS, Kim CW. Critical impact factors affecting the performance of domestic construction projects through megatrend analysis. Journal of the Korean Institute of Building Construction. 2022 Apr;22(2):207-18. http://dx.doi.org/10.5345/JKIBC.2022.22.2.207
6. Kim SG, Park KY, Yu YJ. A development of risk performance index for mega-project performance measurement in view of the integrated cost/schedule/risk. Korean Journal of Construction Engineering and Management. 2009 Jan;10(1):69-78.
7. Moon H, Lee HS, Park M, Lee B, Joo S, Son B. Cost performance comparison of project delivery methods in public sector - focusing on mediator effect of biddropping on change orders -. Korean Journal of Construction Engineering and Management. 2015 Sep;16(5):86-96. https://doi.org/10.6106/KJCEM.2015.16.5.086
8. Kang N, Choi J. Construction cost-schedule integration management methodology by using progress integration unit. Korean Journal of Construction Engineering and Management. 2017 May;18(3):42-51. https://doi.org/10.6106/KJCEM.2017.18.3.042
9. Kim CW, Kim T, Yoo WS, Cho H, Kang KI. Optimized growth curve for estimating performance measurement baseline depended on domestic construction facility type. KSCE Journal of Civil Engineering. 2017 Nov;22:2691-701. https://doi.org/10.1007/s12205-017-0180-2
10. Kim CW. Development of diagnostic performance index for domestic construction projects [Ph.D thesis]. [Seoul (Korea)]: Korea University; 2017. 2 p.
11. Building Act [Internet]. Sejong (Korea): Ministry of Land, Infrastructure and Transport; 2022 Apr 20. Available from: https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EA%B1%B4%EC%B6%95%EB%B2%95
12. Housing Act [Internet]. Sejong (Korea): Ministry of Land, Infrastructure and Transport; 2022 Aug 04. Available from: https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EC%A3%BC%ED%83%9D%EB%B2%95
13. Public Notice on Detailed Standards for Building Supervision [Internet]. Sejong (Korea): Ministry of Land, Infrastructure and Transport; 2020 Dec 24. Available from: https://www.law.go.kr/%ED%96%89%EC%A0%95%EA%B7%9C%EC%B9%99/

%EA%B1%B4%EC%B6%95%EA%B3%B5%EC%82%AC%EA%B0%90%EB%A6%AC%EC%84%B8%EB%B6%80%EA%B8%B0%EC%A4%80

14. Sung Y, Hur YK, Lee SW, Yoo WS. Development of performance indicators on private building construction sites using supervisory report. Korean Journal of Construction Engineering and Management. 2022 Nov;23(6):65-75. https://doi.org/10.6106/KJCEM.2022.23.6.065
15. Jackson J. Data mining; a conceptual overview. Communications of the Association for Information Systems. 2002 Mar;8:267-96. https://doi.org/10.17705/1CAIS.00819
16. Michael Chung H, Paul G. Special Section: Data Mining. Journal of Management Information Systems. 1999;16(1):11-6. https://doi.org/10.1080/07421222.1999.11518231
17. Yan H, Yang N, Peng Y, Ren Y. Data mining in the construction industry: Present status, opportunities, and future trends. Automation in Construction. 2020 Nov;119:103331. https://doi.org/10.1016/j.autcon.2020.103331
18. Ahn J, Ji SH, Ahn SJ, Park M, Lee HS, Kwon N, Kim Y. Performance evaluation of normalization-based CBR models for improving construction cost estimation. Automation in Construction. 2020 Nov;119:103329. https://doi.org/10.1016/j.autcon.2020.103329
19. Kim CW, Yoo WS, Seo J, Kim B, Lim H. A roadmap for applying digital technology to improve the efficiency of construction supervision in building projects: focusing on korean cases. Buildings. 2023 Dec;14(1):75. https://doi.org/10.3390/buildings14010075
20. Jang SR, Kim HS. Association rules analysis between the types and causes of disputes in construction projects. Korean Journal of Construction Engineering and Management. 2022 Sep;23(5):3-14. https://doi.org/10.6106/KJCEM.2022.23.5.003
21. Park H, Lee M, Hwang S, Oh S. TF-IDF based association rule analysis system for medical data. KIPS Transactions on Software and Data Engineering. 2016 Mar;5(3):145-54. https://doi.org/10.3745/KTSDE.2016.5.3.145
22. Ryu JH, You YY. The fourth industrial revolution core technology association analysis using text mining. Journal of Digital Convergence. 2018 Aug;16(8):129-36. https://doi.org/10.14400/JDC.2018.16.8.129
23. Tan PN, Steinbach M, Kumar V. Introduction to data mining. New Delhi (India): Pearson; 2005. 769 p.
24. Son KY, Ryu HG. Association rules analysis of safe accidents caused by falling objects. Journal of the Korea Institute of Building Construction. 2019 Aug;19(4):341-50. https://doi.org/10.5345/JKIBC.2019.19.4.341
25. Kim Y, Kim J, Kim C, Kim K. Cryptocurrency recommendation model using the similarity and association rule mining. Journal of Intelligence and Information Systems. 2022 Dec;28(4):287-308. https://doi.org/10.13088/JIIS.2022.28.4.287
26. Park E, Cho S. KoNLPy: Korean natural language processing in Python. Proceedings of the 26th Annual Conference on Human and Cognitive Language Technology; 2014 Oct 1; Chuncheon, Korea. Seoul (Korea): Korean Institute of Information Scientists and Engineers; 2014. p. 133-6.