KIGAS Vol. 24, No. 5, pp 47~55, 2020
(Journal of the Korean Institute of Gas)
https://doi.org/10.7842/kigas.2020.24.5.47

Data Analytics를 활용한 위험물 화재사고 분석

신은지 · 고문수* · †신동일**
명지대학교 재난안전학과 석사과정, *안성소방서, **명지대학교 화학공학과 교수
(2020년 2월 17일 접수, 2020년 10월 21일 수정, 2020년 10월 22일 채택)

Fire Accident Analysis of Hazardous Materials Using Data Analytics

Eun-Ji Shin · Moon-Soo Koh* · †Dongil Shin**
Dept. of Disaster and Safety, Myongji University, 116 Myeongji-ro, Cheoin-gu, Yongin, Gyeonggi-do 17058, Korea
**Anseong Fire Station, 855 Miyang-ro, Anseong, Gyeonggi-do 17595, Korea*
***Dept. of Chemical Engineering, Myongji University, 116 Myongji-ro, Cheoin-gu, Yongin, Gyeonggi-do 17058, Korea*
(Received February 17, 2020; Revised October 21, 2020; Accepted October 22, 2020)

요 약

위험물 사고는 해당 물질의 누출에 그치지 않고, 초기대응이 부적합한 경우, 화재, 폭발로 이어져 그 피해규모가 확대될 위험이 크다. 하지만 4차 산업혁명과 빅데이터 시대의 대두가 논의되고 있는 시점에서, 새로운 기법들에 바탕한 위험물 사고의 체계적인 분석은 시도되지 못하고, 단편적인 통계 수집에 그치고 있는 것이 아쉬운 실정이다. 본 연구에서는 지난 11년간(2008~2018) 축적된 소방청 위험물 화재사고 데이터를 대상으로 기계학습에 기반한 분석을 진행하였다. Text mining 분석을 통해 분석한 자료를 시각화하여 나타내었고, 아울러 위험물 화재사고 데이터에 존재하는 주요 인자를 이용해 피해규모 예측모델의 개발 가능성을 회귀분석 방법을 적용하여 탐색하였다.

Abstract - Hazardous materials accidents are not limited to the leakage of the material, but if the early response is not appropriate, it can lead to a fire or an explosion, which increases the scale of the damage. However, as the 4th industrial revolution and the rise of the big data era are being discussed, systematic analysis of hazardous materials accidents based on new techniques has not been attempted, but simple statistics are being collected. In this study, we perform the systematic analysis, using machine learning, on the fire accident data for the past 11 years (2008 ~ 2018), accumulated by the National Fire Service. The analysis results are visualized and presented through text mining analysis, and the possibility of developing a damage-scale prediction model is explored by applying the regression analysis method, using the main factors present in the hazardous materials fire accident data.

Key words : hazardous materials, fire accident, machine learning, text mining, regression

I. 서 론

산업 발전에 따라 위험물을 취급하는 업체는 증가하고 있다. 특히 대량 위험물을 취급하는 제조소는 2008년 256개소에서 2017년 323개소로 증가하였다. 2008년부터 2018년 즉, 지난 11년간 국내 위험물 사고는 총 713건이며 지역별로는 경기도 207건(29%), 울산 80건(11%), 경북 57건(8%) 순으로 발생하였다.[1]

†Corresponding author:dongil@mju.ac.kr

지난 2015년 10월 7일 오후 9시 35분경 용인시 처인구 모현면의 한 유류보관 창고에서 불이 발생했다. 용인소방서에 따르면 이 유류창고에 저장된 전체 위험물은 423.8톤이었다. 화재는 석유화학제품의 연속적인 폭발로 일어난 화재이며, 화재진압에 어려움이 있었다. 용인소방서장은 오후 11시경 대응 2단계를 발령하였고, 다음날 오전 3시 43분경 완전히 진화했다. 이 화재는 약 1억 5000만원의 재산피해를 냈고, 1명이 다치고 200여명이 긴급대피하는 소동이 있었다.[2]

지난 2016년 8월 27일 오전 8시 26분경 웨스트버지니아주 뉴마틴스빌에 있는 아실사 나트리움 공장에서 17만 8400파운드의 액화 압축염소를 적재한 직후 철도 탱크차에 42인치 길이의 균열이 생겼다. 이후 2시간 30분동안 17만 8400파운드의 염소 하중이 모두 방출되어 오하이오 강 계곡을 따라 남쪽으로 이동하는 대형 증기 구름을 형성했다.[3]

지난 11년간 위험물 사고 인명피해는 534명 중 경상자 303명(56.7%), 중상자 163명(30.5%), 사망자 68명(12.7%) 순이다. 시도별 재산피해액은 총 59,796.6백만원 중 경기도 20,159백만원(33.7%), 울산 15,385.7백만원(25.7%), 경북 7,169.8백만원(11.9%) 순이다.[1]

기존 위험물 사고 통계는 연말에 백서 형태로 발간되고 있으나, 체계적인 분석은 진행되지 않고 통계 수집에 그치고 있다. 본 연구에서는 지난 11년(2008년 ~ 2018년)간 발생한 위험물 화재사고 데이터를 바탕으로 Text Mining을 통해 변수들의 특징을 파악하여 시각화하였다. 또한, 회귀분석을 통해 주요 인자를 뽑아 재산피해 규모를 예측하는 예측 모델링을 진행하였다.

Ⅱ. 이론소개

2.1. Text Mining

자연어 처리 기반의 Text Mining은 대규모의 비정형 텍스트 데이터에서 의미 있는 정보를 추출하는 것이다. 즉, 비정형 데이터들을 정형화하고, 새롭고 유용한 정보를 찾는 기계학습 기반의 기술이다. 여기서 비정형 데이터란 숫자 데이터와 달리 구조화되지 않은 데이터를 말한다. 본 연구에서는 Text Mining을 통해 빈도수 등과 같은 변수들의 특징을 파악하였으며, 이 특징을 Word Cloud, 히스토그램 등과 같은 그래프 형태로 나타내 분석 결과를 시각화하였다. 여기서 Word Cloud란 빈도수가 높은 단어가 중앙에 큰 글씨로 나타나며, 그 단어를 중심으로 주변에 글씨 크기나 위치가 빈도수에 따라 결정되어 나타나는 구름 형태의 그림이다.

2.2. 회귀분석(Regression)[4]

회귀분석은 변수들간의 인과관계를 분석하여 이를 함수식으로 표현하고 그 타당성 여부를 검토하는 것이다.

회귀분석을 설명하기 위해서는 결정계수(R-square)를 알아야 한다. 결정계수는 독립변수가 종속변수를 얼마나 잘 설명하는지를 나타내는 값으로 1에 가까울수록 설명력이 높고 0에 가까울수록 설명력이 낮다. 본 연구에서는 독립변수가 2개 이상이므로 다중회귀분석(Multiple Regression Analysis)을 실시하였으며 결정계수는 독립변수의 개수가 증가하면 결정계수 또한 증가하게 된다. 이러한 단점을 보완하기 위해 다중회귀분석에서는 수정된 결정계수가 쓰이며, 수정된 결정계수 또한 1에 가까울수록 설명력이 높다.

다중회귀분석의 표준 가정은 다음 네 가지이다. 첫째, 종속변수와 독립변수 간에 선형 관계가 있다. 둘째, 독립변수간에 상관관계가 없이 독립성을 만족해야한다. 셋째, 잔차(Residuals)는 모든 독립변수 값에 대하여 동일한 분산값을 가진다. 넷째, 잔차는 정규분포를 만족하며 기댓값은 0이다. 다중회귀분석의 기본 식은 다음 식 (1)과 같다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p \qquad (1)$$

여기서 x_i는 독립변수를 나타내며, y는 종속변수를 나타낸다. β_i는 각 독립변수의 계수를 나타낸다.

Ⅲ. 연구방법

본 연구에서는 통계 분석에서 많이 사용되는 오픈소스 통계 프로그램인 R을 사용하여 분석을 진행하였다.

2008년부터 2018년까지의 소방청 위험물 화재사고 데이터를 사용하였다. 위험물 화재사고 데이터 중 연도, 화재발생(월), 요일, 지역, 발화요인, 최초착화물 등을 이용하여 다음과 같은 세 가지의 Text Mining을 진행하였다.

첫 번째로 지역별, 시간별, 원인별, 물질종류별 발생건수 분석을 위해 위 데이터는 변형하지 않고 원본 데이터를 이용하여 분석하였다. 두 번째로 오류 방지를 위해 원본 데이터를 약간 변형하였다. 사업장의 개수가 많거나 지역이 넓거나 사용량이 많거나 하는 곳은 오류가 크게 발생할 위험이 있다.

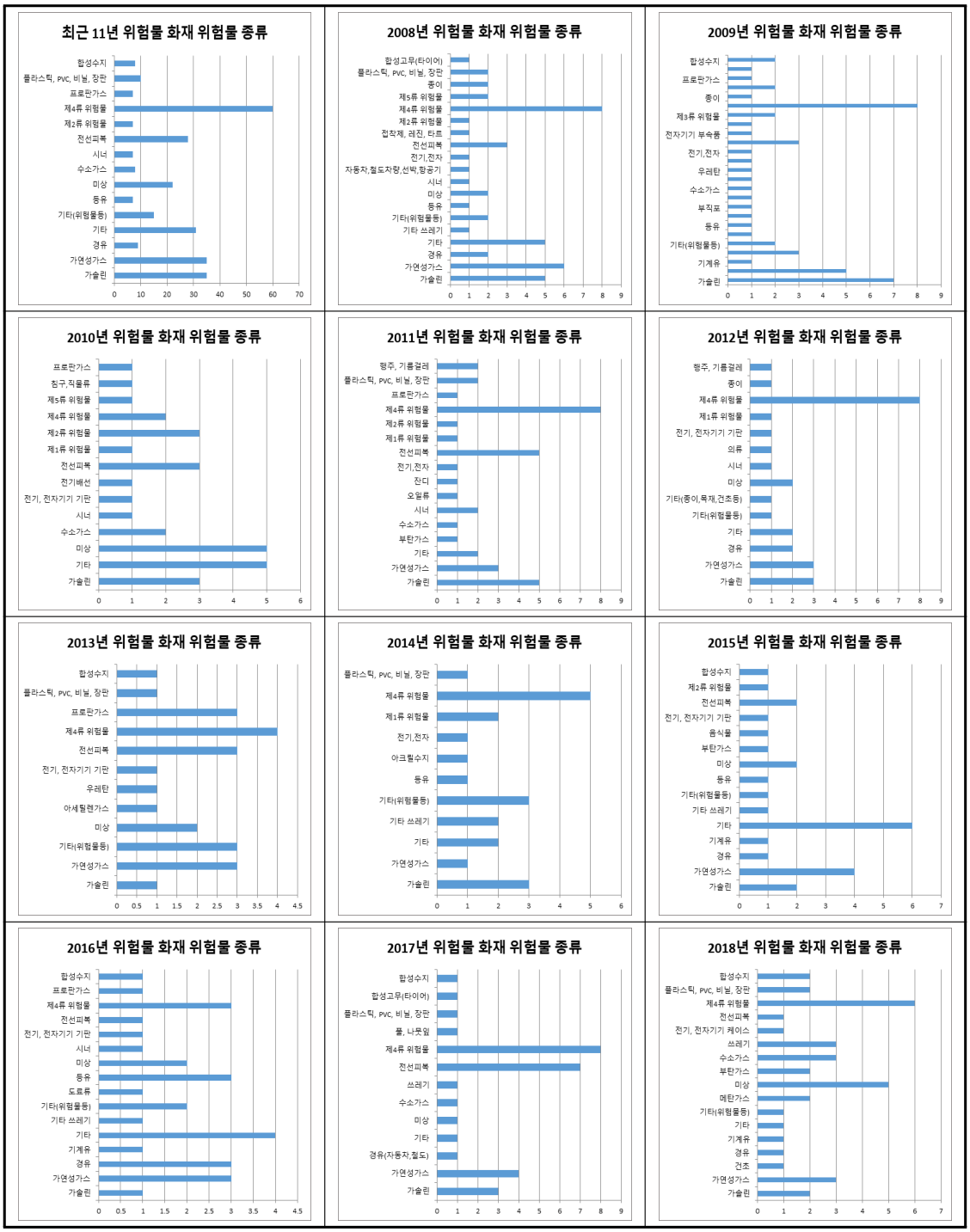

Fig. 1. Frequency of hazardous materials accidents by year.

따라서 이를 방지하기 위해 앞서 분석한 내용 중 일부를 천 개소로 나눠 분석하였다. 세 번째로 11년에 걸쳐 어떻게 연도별로 변화했는지 추가적으로 분석하였다.

사람들은 단변수 분석을 선호한다. 하지만 Data Analytics를 사용하면 다변수 분석이 가능해 종합적인 시각에서 검토가 가능하다. 본 연구에서는 모든 사고에 대해 공통적으로 기록된 20여개의 input 값에서 재산피해액을 output으로 예측하는 다중회귀분석을 실시하였다.

Ⅳ. 결 과

다음 결과는 지난 11년간 위험물 화재사고 자료에서 연도, 시간대, 발화요인, 최초 착화물 등 여러 변수를 이용한 Text Mining 분석과 주요 변수를 추출하여 재산피해를 예측하기 위한 회귀분석을 진행한 결과이다.

4.1. 위험물 종류별 발생건수 분석

위험물 화재사고 자료 중 '최초 착화물'을 이용해 지난 11년간 어떤 종류의 위험물에서 화재가 많이 발생하였는지 분석하였다. 11년간 위험물 화재사고는 총 361건으로 Fig. 1에 첫 번째 그림을 보면 제4류 위험물(시너, 등유, 경유, 가솔린 포함)에서 발생한 화재는 총 118건으로 가장 많은 화재가 발생한 것을 알 수 있다. 가연성가스가 총 35건, 약 9%로 그 뒤를 잇고 있다. 실제로 소방청 통계 자료에 따르면 제4류 위험물을 취급하는 위험물 업체는 2017년 기준 전체의 97%를 차지하고 있을 만큼 많다. Fig. 1에서와 같이 제4류 위험물에서 발생한 화재는 2010년, 2015년, 2016년을 제외한 나머지 연도에서 가장 높은 빈도수를 차지했다.

4.2. 지자체별 발생건수 분석

본 절에서는 지역별(시도별)로 화재 빈도를 분석한 후 경기도 내 지역에서 화재 빈도를 분석하였다. Fig. 2를 보면 경기도에서 위험물 화재 건수가 가장 많은 것을 알 수 있다. 경기도는 위험물 취급업체가 가장 많이 분포되어있는 지역이다. 하지만 위험물 제조소 천 개당 화재 발생 건수를 비교해보면 서울에서 위험물 화재사고가 가장 많이 발생하였으며, 경기도는 8번째로 많이 발생하였다.(Fig. 3)

경기도에는 위험물 제조소가 많이 분포되어 있는데 그 중 평택시에 위험물 제조소가 밀집되어있다. Fig. 4와 Fig. 5는 빈도수가 높은 단어가 중앙에

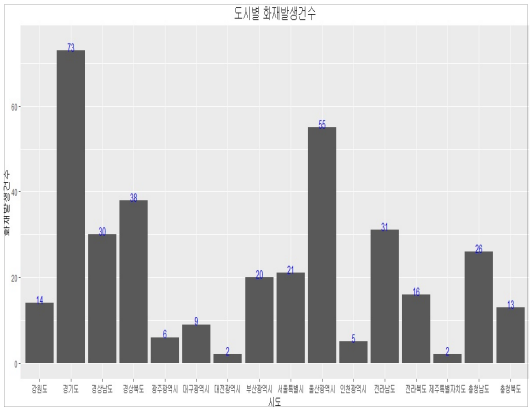

Fig. 2. Number of hazardous materials fires by local governments.

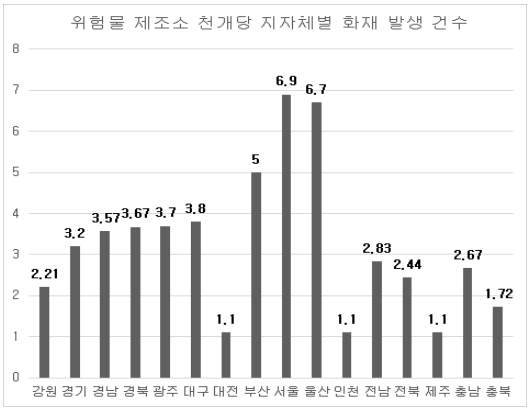

Fig. 3. Number of fires by local governments per thousand hazardous materials manufacturing stations.

큰 글씨로 나타나는 Word Cloud이다. Fig. 4를 보면 평택시에서 위험물 화재사고가 많이 발생했다는 것을 알 수 있다. 또한, 전체 소방서 출동건수를 비교해 보면 남부(울산광역시), 여수, 온산소방서 순이다. 여기서 전남과 울산은 대량 위험물을 취급하는 업체가 많이 분포해 있다.(Fig. 5)

4.3. 시간에 따른 발생건수 분석

Fig. 6를 보면 지난 11년간 2009년에 가장 많은 위험물 화재가 발생하였으나, 2010년에 다소 감소한 것을 확인할 수 있다. **Fig. 7**를 보면 온도가 높은 달이나 습도가 낮은 건조한 달에 화재가 많이

Fig. 4. Ratio of hazardous materials fire occurrences in Gyeonggi-do.

Fig. 5. Ratio of hazardous materials fire occurrences by fire station.

발생한다는 것을 알 수 있다. 계절별로 보면 제4류 위험물이 가장 화재가 많은 화재의 원인이었고, 가을에만 가연성가스에 의한 화재가 가장 많이 발생했다.(**Fig. 8**) 건조한 달인 11월부터 3월달에 발화원인은 부주의가 가장 많았다. 건조한 달에 건초 등에서 발생한 화재가 많을 것이라 예상하였으나, 건초에 의한 화재는 한 건도 발생하지 않았다. 또한, 화재가 많이 발생한 연도나 월에 피해가 높을

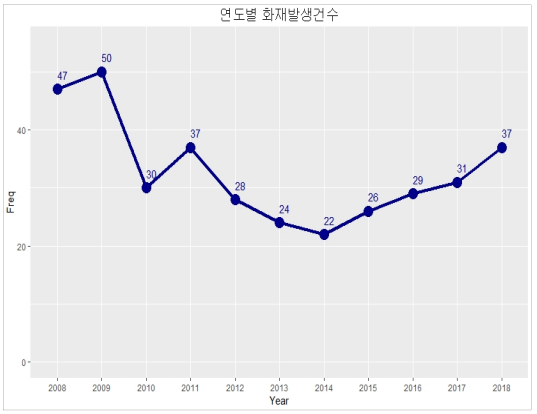

Fig. 6. Number of hazardous materials fire occurrences by year.

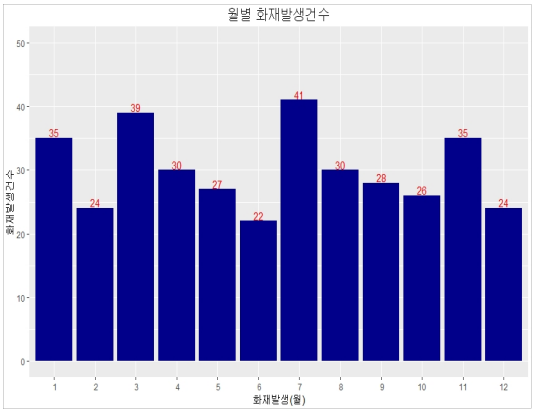

Fig. 7. Number of hazardous materials fire by month.

것이라 예상하였으나 **Table 1**과 **Table 2**를 보면 가장 화재가 많은 연도나, 월에도 인명피해나 재산피해가 가장 높지는 않은 것을 알 수 있다.

Fig. 9를 보면 화요일에 발생한 화재는 69건으로 가장 많이 발생하고, 일요일에 발생한 화재는 35건으로 가장 적게 발생하였다.

시간대별 화재발생건수를 보면 업무가 시작하는 시간대인 9~10시에 화재 빈도가 높았다.(**Fig. 10**) 취약시간대(1~6시)에 주된 발화요인이 부주의일 것으로 예상하였으나, 용접, 절단, 연마, 그리고 화학적 요인에 의한 화재가 많이 발생하였다.

Fig. 8. Popularity of hazardous materials fire types by season.

Table 1. Annual damage by hazardous materials fire [1]

year	Total number of casualties	Total property damage [1,000 Won]
2008	18	203,839
2009	**15**	**693,741**
2010	16	777,147
2011	16	440,371
2012	10	450,742
2013	28	1,846,440
2014	15	146,044
2015	12	2,660,434
2016	24	428,371
2017	6	1,699,560
2018	11	8,131,367
total	171	17,478,056

Table 2. 11-year monthly average of damage by hazardous materials fire

month	Total number of casualties	Total property damage
1	10	1,791,318
2	9	313,603
3	38	472,485
4	12	572,965
5	12	368,241
6	17	439,980
7	9	467,698
8	19	1,630,825
9	14	463,457
10	8	10,086,721
11	13	413,993
12	10	456,770
total	171	17,478,056

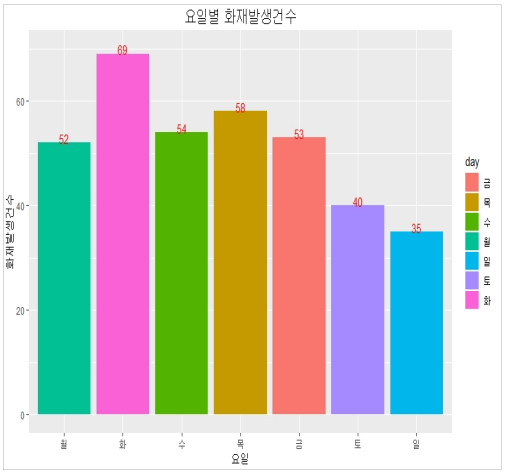

Fig. 9. Number of hazardous materials fire by weekday.

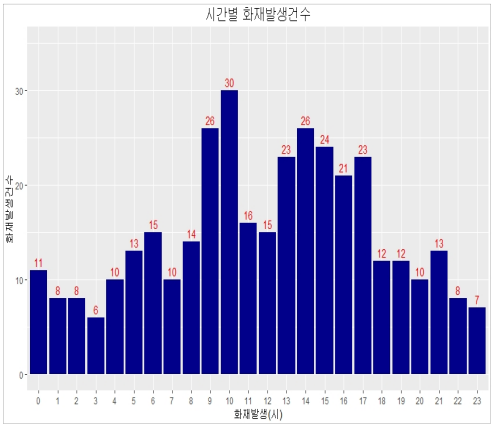

Fig. 10. Number of hazardous materials fire by time.

4.4. 회귀분석 결과

(1) 다중공선성 및 수정된 결정계수

다중공선성이란 회귀분석에서 사용된 모형의 일부 예측 변수가 다른 예측 변수와 상관 정도가 높아, 데이터 분석 시 부정적인 영향을 미치는 현상을 말한다. Fig. 11를 보면 본 연구에 사용한 변수들의 다중공선성은 모두 4 미만으로 변수 간 상관관계가 거의 없는 것을 확인할 수 있다.

결정계수는 총 변동 중 회귀 직선에 의해 설명되는 변동의 비율을 의미한다. 다중회귀분석의 경우

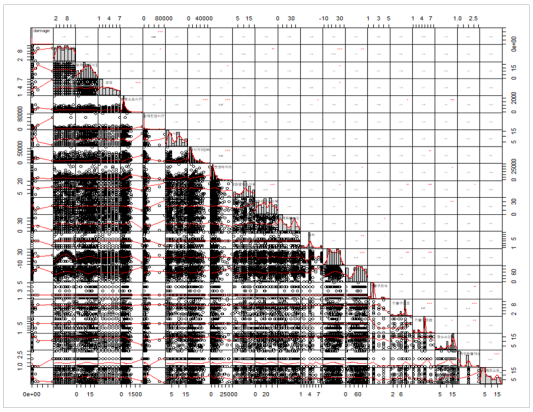

Fig. 11. Multicollinearity.

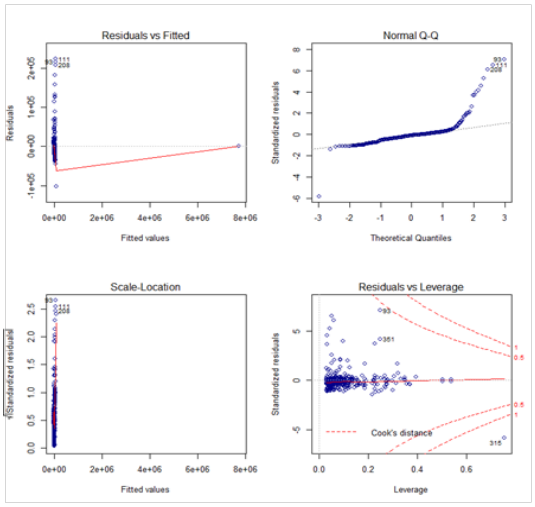

Fig. 12. Regression diagnostics plot.

결정계수는 독립변수의 개수가 증가하면 결정계수 또한 증가하므로 이러한 단점을 보완하기 위해 수정된 결정계수가 쓰인다. 이 수정된 결정계수는 0부터 1 사이의 수를 가지며 숫자가 1에 가까울수록 좋다. 본 연구에서는 Stepwise의 변수선택 방법으로 0.8745라는 수정된 결정계수를 얻었으며, 발화열원 소분류, 화재진압시간, 장소소분류, 건축위험물대상, 습도, 출동 소요시간의 변수를 선택하였다. 눈에 띄는 이상치를 제거하면 0.9926라는 수정된 결정계수를 얻으며 회귀 이상진단을 위해 다음과 같은 Fig. 12를 얻을 수 있다.

Table 3. Example of property damage prediction results

prediction	lower	upper	actual value	difference from actual value	predictive success
7715334.0000	0	7813493.75	7715334	5.587935e-09	TRUE
1696.4364	0	80295.50	1658	3.843637e+01	TRUE
3252.4782	0	79993.97	3157	9.547815e+01	TRUE
1073.0464	0	75601.12	884	1.890464e+02	TRUE
12467.9897	0	84779.45	12191	2.769897e+02	TRUE

첫 번째 그래프는 실제 잔차가 어떻게 분포되어 있는지 확인할 수 있으며, 기울기 0이 이상적이다. 기울기가 0에 가까우므로 이상적이라고 할 수 있다. 두 번째 그래프는 데이터의 분포가 정규성을 만족하는지 파악하기 위한 그래프로 점이 직선에 가까울수록 정규성을 만족한다는 것이다. 점들이 직선에 가깝게 분포하고 있으므로 잔차는 정규성을 만족한다. 세 번째 그래프는 일반 잔차에서 개선된 표준화된 잔차를 제곱근하여 표현한 것으로 잔차의 정도를 파악하는데 용이하다. 이 그래프 또한 기울기 0이 이상적이다. 마지막 그래프는 설명변수가 얼마나 극단에 치우쳐 있는지 확인할 수 있다. 대체적으로 잘 분포하고 있는 것으로 보아 극단에 치우져치지 않은 것을 알 수 있다.

(2) 재산피해 예측 결과

본 연구에서 재산피해 예측에 사용된 변수는 총 6가지로 '발화열원 소분류', '화재진압시간', '장소소분류', '건축위험물대상', '습도', '출동 소요시간'이다.

Table 3은 재산피해 예측 결과의 일부이다. 예측값의 95% 신뢰구간을 최솟값(lower)과 최댓값(upper)으로 나타내었다. 재산피해액은 음수값이 될 수 없으므로 예측 신뢰구간의 최솟값을 0으로 설정하였다. 재산피해액의 실제값을 actual value로 나타내었으며, 예측 성공 여부는 prediction 값이 신뢰구간 [lower, upper] 사이에 있는 값이라면 TRUE로 표시하였다. 재산피해액 예측 분석 결과 본 회귀모형의 예측 정확도는 0.794이다.

V. 결 론

본 연구에서는 2008년부터 2018년 위험물 화재사고 데이터를 이용하여 Text Mining을 통한 분석과 재산피해 예측을 위한 회귀분석을 진행하였다.

Text Mining을 통한 분석으로 각 변수별 화재 발생건수 확인하여 특징을 추출하였다. 위험물 화재사고는 제4류 위험물에서 가장 많이 발생하였다. 특히, 제4류 위험물을 다루는 위험물 업체은 2017년 기준 전체의 97%를 차지하고 있을 만큼 많이 사용 되는 위험물인 만큼 철저한 관리가 필요한 것으로 보인다. 지역별 위험물 화재 발생 건수는 경기도가 71건으로 가장 높았으나, 위험물 시설 천개소 당 발생건수는 서울 6.9건, 울산 6.7건, 부산 5건 순서로 높은 것을 확인하였으며, 경기도는 3.2건에 불과했다. 이는 경기도에 위험물 시설이 많이 밀집되어 있지만 그에 비해 화재는 서울, 울산 등과 비교하여 적게 발생한다는 것을 알 수 있다. 위험물 화재는 7월, 3월, 1월, 11월 순으로 많이 발생하였다. 계절별 화재 빈도가 높은 위험물은 제4류 위험물 있었으나 가을엔 가연성가스의 화재 빈도가 가장 높았다. 또한, 화요일에 화재가 가장 많이 발생하였으며, 주요 업무 시간인 10시 ~ 15시 사이에 화재가 많이 발생하였다.

지난 11년간 위험물 화재사고 건수는 총 361건으로 절대 건수가 작기 때문에 사고의 건수가 재산피해와 인명피해에 비례하지 않고 오히려 대형사고가 영향이 크다는 것을 확인할 수 있었다.

다변수 예측모델은 재산피해를 크게 하는 중요인자와 중요인자 간의 민감도 분석이다. 본 데이터를 이용하여 회귀분석을 통한 예측을 진행하였다. 재산피해에 영향을 미치는 변수는 '발화열원 소분류', '화재진압시간', '장소소분류', '건축위험물대상', '습도', '출동 소요시간'으로 이 변수를 이용한 회귀분석을 통해 0.794이라는 예측 정확도를 얻었다. 재산피해를 줄이기 위해 발화열원과 건축위험물대상 등 관리 및 개선이 필요할 것으로 보인다. 인명피해의 경우 사망과 부상의 통합문제, 정수값 문제 등이 있어 이는 향후 연구를 통해 개선할 것이다.

기존 연간 화재 통계에는 위험물 화재사고가 361건에 그치고 있어 화재 예방 및 예측연구를 하

는 데에 있어 어려움이 있었다. 이후 통계가 추가된다면 위험물 화재에 예방 및 예측에 도움이 될 수 있어 화재 통계의 추가가 요구된다.

감사의 글

본 연구는 산업통상자원부 스마트디지털엔지니어링전문인력양성사업의 연구비지원(P0008475-G02 P04570001901)에 의해 수행되었습니다. 이에 감사드립니다.

REFERENCES

[1] National Fire Data System(NFDS)
[2] Kang S. M., 'Yongin Oil Warehouse Fire', Respond to Yongin Fire Station for quick response, http://www.gukjenews.com/news/articleView.html?idxno=349328, (2015)
[3] Hazardous Materials Accident Report, National Transportation Safety Board, (2016)
[4] Milton J., Arnold C., "*Introduction to probability and statistics*", McGraw-Hill Education, (2002)