Journal of the Korean Society of Marine Environment & Safety
Research Paper
Vol. 26, No. 7, pp. 798-809, December 31, 2020, ISSN 1229-3431(Print) / ISSN 2287-3341(Online) https://doi.org/10.7837/kosomes.2020.26.7.798

해상안전 통계 항목 다양화를 위한 EDA 기반 통계 속성 도출 및 활용에 관한 연구

강성경* · 이영재**†
* 동국대학교 일반대학원 경영정보학과 박사과정, ** 동국대학교 경영정보학과 교수

Study on the EDA based Statistics Attributes Discovery and Utilization for the Maritime Safety Statistics Items Diversification

Seong Kyung Kang* · Young Jai Lee**†
* PhD Candidate, Department of MIS, Dongguk University, Seoul 04620, Korea
** Professor, Department of MIS, Dongguk University, Seoul 04620, Korea

요 약 : 과학적 행정을 위한 증거 기반 정책 수립과 평가에 대한 요구로 통계(데이터) 활용 중요성이 날로 강조되고 있다. 통계는 사회 전반의 현상을 수치로 제공함으로써 직관적으로 어떤 현상을 설명할 수 있도록 하며, 합리적인 의사결정을 위한 공공자원으로 설명된다. 이러한 특성으로 통계는 정부 정책 결정 및 각종 현상의 연구·분석 등에 기초자료이자 근거자료로 널리 활용되고 있으나 그 중요성에 비해 통계의 역할은 제한적인 수준이다. 이는 현재 개방된 통계가 단순 결과 요약 자료 수준이며 공급자 위주로 생산되어 수요자 관점에서 가치 창출을 위한 수단으로는 부족하다는 의미이며, 본 연구에서는 이러한 문제 보완을 위해 현재 제공되는 통계 항목 외에 정책이나 연구에 다양하게 활용할 수 있는 추가 속성을 탐색했다. 연구에 활용한 기준 통계자료는 해양경찰청에서 발간하는 「해상조난사고 통계 연보」이며, 해양경찰에서 작성하는 선박사고 상황보고서 텍스트 분석을 통해 추가할 수 있는 속성들을 도출했다. 텍스트 분석을 통해 도출된 56개 속성에 대해 데이터를 수집하고 EDA를 수행한 결과, 유의확률(p-value < .05)을 만족하는, 상관계수 0.7 이상의 강한 상관관계가 있는 속성 조합 18개와, 중간 정도의 상관관계(0.4 이상 0.7 미만)를 가지는 속성조합 70개, 총 88개의 조합을 발굴할 수 있었다. 더불어 EDA를 통해 발견된 추가 속성을 정책적으로 활용하기 위해 수난대비기본계획 세부 전략별 키워드 분석을 실시하고, 키워드와 EDA 도출 속성 간 매칭작업을 통해 속성의 활용 가능 여부를 검토했다.

핵심용어 : EDA, 증거기반행정, 통계속성, 정책개선, 해상안전

Abstract : *Evidence-based policymaking and assessments for scientific administration have increased the importance of statistics (data) utilization. Statistics can explain specific phenomena by providing numerical values and are a public resource for national decision making. Due to these inherent attributes, statistics are utilized as baseline and base data for government policy determinations and the analysis of various phenomena. However, compared to the importance, the role of statistics is limited, and statistics are often used as simple abstracts, produced mainly for suppliers, not for consumers' perspectives to create value. This study explores the statistical data and other attributes that can be utilized for policies or research to address the problems mentioned above. The baseline statistical data used in this study is from the Maritime Distress Accident Statistical Yearbook published by the South Korean Coast Guard, and other additional attributes are from text analyses of vessel casualty situation reports from the South Korean Maritime Police. Collecting 56 attributes drawn from the text analysis and executing an EDA resulted in 88 attribute unions: 18 attribute unions had a satisfactory significance probability (p-value < .05) and a strong correlation coefficient above 0.7, and 70 attribute unions had a middle correlation. (over 0.4 and under 0.7). Additionally, to utilize the extra attributes discovered from the EDA politically, a keyword analysis for each detailed strategy of the disaster Preparation basic plan was executed, the utilization availability of the attributes was obtained using a matching process of keywords, and the EDA deducted attributes were examined.*

Key Words : *Exploratory Data Analysis, Evidence-Based Decision-Making, Statistics Attribute, Policy Improvement, Maritime Safety*

* First Author : hshs4123@naver.com, 02-2260-3297
† Corresponding Author : yjlee@dongguk.edu, 02-2260-3297
※ 본 논문은 박사학위 논문 일부에서 발췌하여 작성되었습니다.

1. 서 론

1970년대 이후 해양사고 통계가 집계된 이래로 정부는 통계 자료를 활용한 분석을 통해 반복 사고를 예방하고, 발생 시 피해 최소화를 위한 방법을 강구하고 있다(Lee et al., 2020).

통계는 사회 전반의 현상을 수치로 제공함으로써 직관적으로 어떤 현상을 설명할 수 있도록 하며, 합리적인 의사결정을 위한 공공자원으로 설명된다. 이러한 특성으로 통계는 정부 정책 결정 및 각종 현상의 연구·분석 등에 기초자료이자 근거자료로 널리 활용된다.

하지만 이러한 중요성에 비해 여러 분야에서 통계 활용의 역할은 미미한 수준이다. 이는 정책이나 연구 분야에 사용되는 통계가 '단순 집계' 수준의 자료로 매년 동일(유사)항목으로 제공되고 있기 때문이며, 현재 제공되는 통계 항목이 매우 제한적이고 다양성 또한 부족함을 보여주는 대목이기도 하다.

물론 통계 항목이 과거보다 다양화되었지만 연구나 정책에서 제언하는 사고 예방책 등 개선점은 단순히 수치상 부각되는 것을 포괄적인 방향으로 제시하는 경향을 보인다. 대부분의 해상안전 정책수립, 연구에서 해양경찰청이나 해양안전심판원의 통계 및 사례집을 활용하고 있지만 현재의 제공 자료만으로는 구체적인 정책 제안에 한계가 있다.

결국 자료의 실효성을 위해서는 사고 분석 과정에서 데이터 특성을 파악하여 어떤 정보를 활용해 결과를 얻을 것인지 구체적인 요구사항을 정의할 필요가 있다. 즉, 현행 통계가 제한적 정보 수준이 아닌 다양성을 가진 자료로 개선되는 것이 필요한 시점이다(Lee et al., 2020).

정책, 연구, 그 밖의 의사결정에 있어 통계는 단순 요약자료 그 이상의 의미를 가진다. 각종 업무에 필요한 정보와 증거를 모두 제공하고, 정책 평가에 있어 하나의 측정 수단이 되기도 한다. 따라서 이러한 통계의 역할을 강화시키기 위해 기존 공급자 위주의 통계 생산이 아닌, 수요자 가치창출 수단으로 통계의 중요성을 부각시킬 필요가 있다.

따라서 본 연구에서는 해양경찰청에서 발간하는 「해상조난사고」의 통계 항목(속성) 외에 정책 및 연구 분야에 폭 깊게 활용할 수 있는 해상안전(선박사고) 관련 속성들을 추가로 발굴하고자 한다.

본 연구에서는 선행연구(Lee et al., 2019)에서 제시한 선박사고 상황보고서 용어 분석결과 자료를 활용하여 추가 속성에 대한 데이터를 수집하고, 탐색적 데이터 분석(EDA : Exploratory Data Analysis, 이하 EDA)을 실시했다. 또한 EDA 결과로 도출된 상관성 있는 데이터 조합을 실제 정책 과정에서 활용하기 위한 정책 전략별 키워드 분석과 EDA 도출 속성 간 매칭 작업을 제시했다.

2. 문헌연구

2.1 통계 생산 · 활용 패러다임 변화와 빅데이터

사회 환경 변화에 따른 불확실성과 정책에 대한 신뢰성 및 수용성을 높이기 위해 정책 과정에서 '과학적 증거' 활용이 강조되고 있다. 이는 데이터를 정책 결정 근거로 활용하고자 하는 것으로 '증거 기반(evidence-based), 데이터 기반, 통계 기반'등의 용어가 통용되고 있다. 명확하고 객관적인 증거에 기초해 정책을 수립하고, 그 효과가 달성되었는지 평가하면 정책 실패를 방지, 최소화 시킬 수 있다는 관점이다(Howlett, 2009).

우리나라의 경우 박근혜 정부의 공공 데이터 개방, 문재인 정부의 데이터 근거 과학적 행정 구현이 강조되면서 증거기반정책 및 데이터 주도 행정이 자리 잡고 있다(Oh et al., 2017). 특히 정책과정에서 국가통계를 근거자료로 활용하도록 하는 '승인통계제도'의 운영과, 정책 도입부터 평가에 있어 통계지표를 개발·적용하도록 하는 '통계기반 정책제도'만 보아도 통계가 정책 결정에 있어 가장 직접적인 과학적 유형으로 인식되고 있음을 알 수 있다.

또한 통계 생산에 있어 우리나라는 각 기관에서 고유 업무 수행을 위해 필요한 통계를 개별적으로 작성하는 '분산형 통계제도'를 채택하고 있어 각 기관에서 생산되는 '행정 자료'를 기반으로 통계를 작성하면 분야별 전문지식을 효율적으로 반영할 수 있다는 장점이 있다(National Statistical Office, 2020).

각종 제도뿐만 아니라 IT의 발전도 통계 생산 방식을 변화시켰다. 이제는 방대한 데이터 수집은 물론, 정형 데이터뿐만 아니라 텍스트, 사진, 영상, 위치정보 등 다양한 형태의 비정형 데이터까지도 취급 가능해졌으며, 저장, 분석 기술의 발달로 기존에 수집할 수 없었던 자료, 수집 대상이 아니었던 자료, 혹은 수집했더라도 분석이 어려워 버려졌던 자료까지도 업무에 활용할 수 있게 되었다.

즉, 데이터 수집, 저장, 가공, 분석의 발전은 조직이 데이터를 활용해 업무를 효율적으로 수행할 수 있는 기반을 만들었고, '국가통계' 영역에 있어서도 데이터 탐색을 통해 유용한 인사이트를 얻고자 하는 노력이 증대되고 있다.

이처럼 과학적 행정을 위한 증거 기반 정책 수립과 평가에 대한 요구로 (통계)데이터 활용 중요성이 날로 강조되고 있으며, 특히 통계 생산에 있어 빅데이터 환경은 방대하고 다양한 형태의 데이터 탐색을 통해 유의미한 속성(변수)과 패턴을 도출할 수 있는 기반을 제공하고 있다

2.2 탐색적 데이터 분석(EDA)

탐색적 데이터 분석(EDA : Exploratory Data Analysis)은 데이터를 다각도로 관찰하여 이해하고자 하는 방법론이다. 데

이터의 특징과 구조에서 인사이트를 발견하는 귀납적 분석 기법이며, 선입견 없이 데이터를 유연하게 탐색할 수 있다.

EDA는 확증적 데이터 분석[1]전에 가설 생성단계에서 증거를 수집하는 관점으로 주로 가설 형성의 장려와 관련이 있다. 주로 기술통계법을 사용하여 데이터를 요약, 가공해 유용한 정보를 얻으며, 이 과정에서 데이터의 오류 확인 및 변수 간의 대략적인 관계나 크기 확인, 새로운 가설 등을 모색하게 된다.

즉, 데이터에 대한 이해를 바탕으로 서로 다른 질문, 가설 등을 찾아내 결론의 특이성을 도출하는 것이 EDA라고 할 수 있다(Behrens, 1997). 단, EDA에서는 공식적인 통계 모델링이나 추론은 포함되지 않으며, EDA를 통해 얻은 새로운 정보나 지식을 추후 확증적 데이터 분석에 활용하여 추론이나 예측모델을 테스트하기도 한다(Good, 1983).

EDA는 변량분석 수준과 그래픽화 여부의 교차에 따라 단변량 비그래픽(univariate non-graphical), 단변량 그래픽(univariate graphical), 다변량 비그래픽(multivariate non-graphical), 다변량 그래픽(multivariate graphical) 4개 유형으로 분류할 수 있다(Seltman, 2018).

단변량 분석은 한 번에 하나의 변수(데이터 열)를 확인하는 것이고, 다변량 분석은 변수 간 '관계'를 탐색하기 위해 한 번에 두 개 이상의 변수를 확인하는 것을 말한다. 그래픽 방법은 데이터를 도표나 그래프 등으로 시각화하여 표현하는 것으로 전체적인 그림을 볼 수 있다는 장점을 가진다

통상 EDA의 절차는 문제를 정의한 뒤 데이터를 탐색하기 위한 데이터 관찰(파악), 개별 속성 분석, 속성 간 관계분석, 가설모색 과정으로 구성되며, 앞의 EDA 유형에 따라 분석 과정은 다소 차이를 보일 수 있다.

데이터 관찰은 전체 데이터를 살펴 데이터의 오류나 누락 등을 파악하고, 데이터의 형태(범주형, 수치형)와 척도(명목, 서열, 등간, 비율)를 확인하여 데이터 분석 방법이나 수단을 확인하는 단계이다.

다음으로 데이터의 개별 속성을 확인하여 속성 값이 예측한 범위와 분포를 가지는지, 이상치(outlier)가 존재하는지를 찾아낸다. 개별 데이터를 살펴보면서 전반적인 추세와 특이사항을 관찰할 수 있다. 데이터의 중심 파악을 위해 평균(mean), 중앙값(median), 최빈값(mode)을 활용할 수 있고, 분산도 파악을 위해 범위(range)와 분산(variance) 등의 요약통계량을 사용할 수 있다. 이 때 시각화를 이용하면 요약통계량에서 드러나지 않는 패턴을 한 눈에 볼 수 있는 장점이 있다(Fig. 1. 참조).

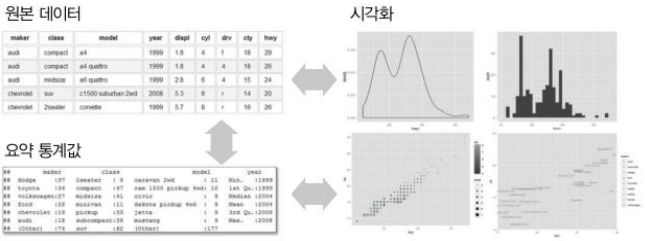

Fig. 1. Exploratory Date Analysis Process Sample (Kim, 2016).

속성 간 관계 분석은 의미 있는 상관관계를 가지는 속성 조합을 찾아내는 것으로 도출한 속성이 많을수록 여러 조합에 대해 다양한 분석을 수행할 수 있다. 이 과정부터 본격적으로 '탐색적' 분석이 수행된다고 할 수 있다. 이 때 조합한 속성 특성에 따라 관계 분석 시 적용할 분석 및 시각화 방법은 Table 1과 같다.

Table 1. Summary statistics way in conformity with data combination

Data Type	Summary Statistics
Categorical-Categorical	Cross-Tabulation
Numerical-Numerical	Correlation Coefficient
Categorical-Numerical	Statistical Value by Category

마지막으로 EDA 결과에서 발견된 내용을 바탕으로 새로운 연구 질문과 가설을 모색하게 된다. 추후 EDA에서 발견된 패턴과 새로 만든 가설을 테스트하기 위해 확증적 데이터 분석을 수행할 수 있다.

2.3. 국가통계 활용현황과 한계

2.3.1 정책 분야에서의 국가통계 활용 현황

'통계기반 정책평가제도'의 존재만 보아도 정책 분야에서 통계가 얼마나 큰 역할을 하는지 알 수 있다. 국가통계는 정책 수립과 평가에 있어 주요 지표가 된다.

'선박사고' 안전과 관련한 주요 정책은 해양경찰청의 '수난대비기본계획'에서 찾아볼 수 있다. 이 계획은 「수상에서의 수색 · 구조 등에 관한 법률」에 따라 5년 단위로 수립되며, 해상에서 자연·인위적 원인으로 발생하는 조난사고로부터 사람과 재산을 보호하기 위한 정책 사업들로 구성된다.

현재 2차 계획('18 ~ '22)이 시행되고 있으며, 수난구호 협

1) 가설 설정 후 수집한 데이터로 가설을 평가하고 추정하는 방법으로, 가설을 테스트 하는 것이 주목적이다. 주로 추론통계를 사용하며, 선행연구에 기반을 두어 가설 유효성 검정에 있어 엄격한 절차와 방법을 가진다.

력체계 구축, 협력기관 간 장비 및 인력 지원, 수난구호에 관한 정보 수집 및 교환, 교육/훈련, 수산/수난 구호 장비 및 시설 확충·관리, 응급의료 및 수색/수조 체계화, 연안사고 대응체계에 대한 세부 계획을 담고 있다.

이 계획에서 사용되는 주요 국가통계는 사고 관련 항목이며, 성과지표로는 '사고발생 현황(선박/인명)', '구조현황(선박/인명)', '구조불능 현황(선박/사망/실종)' 등 사고 결과와 관련된 속성을 사용한다. 즉 사고 결과에 대한 (발생)빈도가 전년도 대비 얼마나 증감되었는지를 토대로 다음 정책 목표를 설정하고 계획을 수립하고 있다.

문제는 이러한 정책 및 여러 의사결정에서 사용되는 현행 통계항목이 어떤 현상을 설명하기에 완전한 지표가 되지는 못한다는 점이다. 앞서 성과지표로 사용한 통계항목은 사고 '결과'를 나타내는 가장 뚜렷한 지표이지만, 실제로 선박사고를 줄이기 위해 시행한 세부 과제의 효과인지를 측정하기 위한 적정지표인지는 알 수 없다.

해양경찰청 수난구호 관련 세부 과제는 위에서 언급한 내용을 토대로 총 7개 추진전략에 대해 18개의 추진과제로 나뉘며, 18개 과제는 다시 44개의 세부과제로 계획되어있다. 전략별 과제는 과거부터 같은 맥락으로 수행되었던 부분도 있지만, 시대적 변화에 따라 더욱 세분화되고 맞춤화된 상세 대책들이 수립되고 있다. 또한 기술 발전에 따라 능동적 사고 예방과 대응을 위한 첨단 기술 활용 전략 및 사고 관리를 위한 각종 협업 전략들이 강구되고 있다.

이처럼 정책은 과거보다 구체적여지고, 새로워지고 있기 때문에 이러한 변화를 반영할 수 있는 통계항목(지표) 또한 함께 개발되어야 한다. 또한 과거부터 생산된 통계가 현실을 반영하기 어렵다면 이 또한 개선되어야 정책 분야에 있어 통계가 제대로 된 역할을 할 수 있다(Kwon, 2017).

2.3.2 연구 분야에서의 국가통계 활용 현황

통계는 정책 과정에서 뿐만 아니라 연구 분야에 있어서도 중요한 역할을 한다. 통계법에서의 통계 목적만 보아도 정부 정책 수립이나 평가 이외에 '경제·사회현상의 연구, 분석' 등에 활용할 목적으로 수집한 수량적 정보라고 명시되어 있다. 이러한 연구 결과물은 결국 정책 수립 과정에 활용되기 때문에 정책과 긴밀하게 연결된다고 할 수 있다.

선행연구에서 선박(해상) 안전과 관련해 사고 분석 및 개선 방안 모색 과정에서 국가통계가 어떻게 활용되었는지 살펴보면 20여 년 전과 현대의 연구 결과가 유사한 측면을 보임을 알 수 있다(Fig. 2 참조). 이는 현재의 통계가 과거보다 다양화되고는 있으나 여전히 개선방안 모색 과정에서 통계 정보가 가지는 한계가 있음을 보여주는 대목이기도 하다.

이러한 한계로 최근에는 단순 통계 자료에서 찾기 어려운 다양한 정보를 탐색하기 위해 정부의 사고보고서나 재결서 등의 행정자료를 분석하기도 하며, 현장조사나 그 외 관련 통계들을 접목하여 분석의 범위를 넓혀가고 있는 추세이다.

2.3.3 통계(속성) 다양화를 위한 개선 연구

통계의 활용 범위를 넓히기 위해 기존 대비 다양한 속성을 발굴하고자 하는 시도가 계속되고 있다. 특히 이러한 개선 작업에 있어 빅데이터와 행정 자료를 활용한 연구가 활발히 진행되고 있다.

통계 개선을 위한 시도는 분야를 막론하고 진행되고 있다. 몇몇 선행연구들의 특징을 요약해보면 크게 3가지 관점으로 구분할 수 있다. 첫 번째는 '조사통계'에 대한 비용, 응답자 부담을 줄이기 위해 행정 자료에서 '대체 가능한' 조사항목을 발굴하는 연구(Ahn, 2015; Lee, 2017)가 진행되었고 두 번째는 기존의 국가통계에는 없던 항목을 관련 행정 자료에서 찾아내려는 연구들이다(Choi, 2016; Lee et al., 2012). 마지막으로 기존 통계와 행정 자료에서 관련성 높은 변수들을 결합하여 새로운 변수(파생변수)를 만드는 연구가 있다(Hong, 2015; Park, 2018).

유의미한 (통계) 속성 발굴을 위한 연구는 해상안전 분야에서도 지속적으로 수행되고 있다. Chae et al.(2019)은 선박에서의 비상대응능력 향상 방안 제언을 위해 해양안전심판원 재결서와 선사 비상대응 매뉴얼을 검토하였으며, 사고 대응과정에서의 문제점 발견을 위해 분석 문서에서 대응 관련 속성을 사고 종류별로 제시했다.

Lee et al.(2019; 2020)은 해상안전(선박)과 관련한 통계 자료 한계점 개선을 위해 해양경찰청 사고보고서를 분석하여 선박사고 관련 정보 분류체계를 수정·보완하였으며, 개선된 분류체계 속성 간 상호연관성이 있음을 언급했다. 또한 도출된 속성들을 바탕으로 기존 통계만을 활용했던 연구에서는 볼 수 없던 선박사고 '초동대응'에 대한 분석을 수행했다. 사고 발생부터 신고접수, 상황전파, 현장출동에 대한 속성들을 바탕으로 선박사고 유형별 특성 및 초동대응 관련 속성에 따른 인명피해 양상을 제시했다.

빅데이터 시대에 생성되는 수많은 정형, 비정형 데이터로부터 유의미한 속성을 발굴하고자 하는 노력이 계속되고 있다. 기술 발전으로 다양한 자료로부터 인사이트를 도출하는 것이 가능해졌고, 복잡해지는 사회 현상을 설명하기 위해 새로운 설명변수(속성)들이 필요해지고 있다. 이제는 전통적으로 인과관계를 검정하고자 하는 연역적 방법이 아닌 데이터로부터 새로운 현상들을 발견하고자 하는 귀납적 방법으로의 전환이 필요한 시점이다(Anderson, 2008).

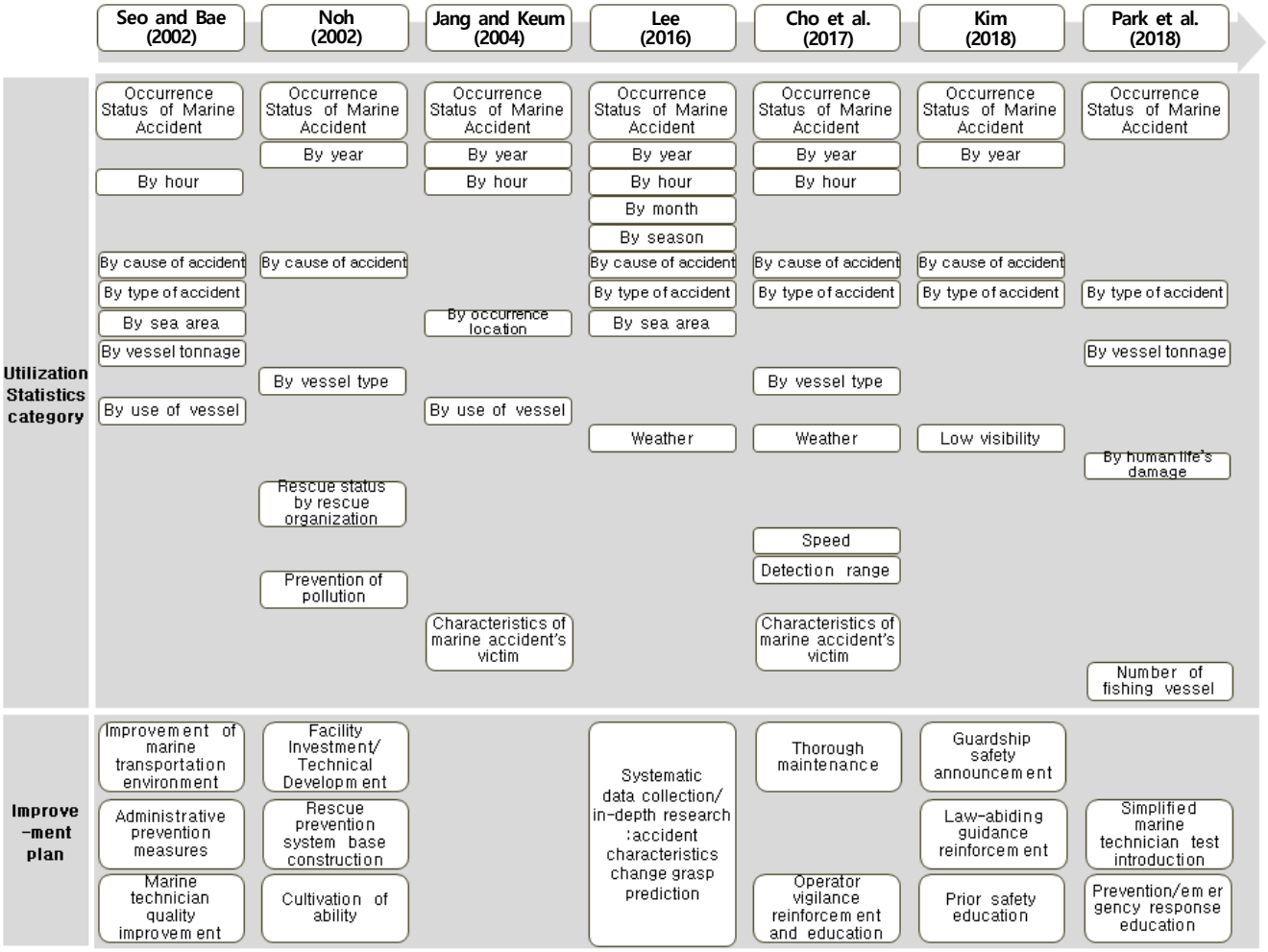

Fig. 2. Preceding research comparison with maritime safety statistics data.

3. 연구 설계

3.1 연구 프레임워크

본 연구는 크게 3개 부문으로 구성된다. 가장 먼저 문헌연구를 통한 현황분석 부문으로, 해상안전(선박사고) 관련 자료에서 현행 통계(속성) 활용 문제점을 도출하고 다양화 가능성을 모색한다.

다음으로 탐색적 데이터 분석(EDA)을 실시한다. 해양경찰서에서 작성하는 행정자료인 선박사고 상황보고서에서 사용하는 용어들을 분석하여(텍스트 마이닝) 특성에 따라 범주화 하고(분류체계 정립), 해당 용어(속성)에 대한 데이터들을 수집, 전처리한다. 데이터가 정리되면 속성 별 특성(데이터 형태 등)을 확인하여 속성 간 관계분석을 실시한다.

마지막으로 정책 분야에 활용하기 위해 정책 전략 별 키워드를 분석하고, EDA 도출 속성과 매칭 한다. 매칭작업에는 행정 자료 분석을 통해 추가로 도출한 속성(개별 속성)과 관계 분석을 통해 도출한 유의미한 속성 조합을 모두 사용할 수 있다.

다음 Fig 3은 본 연구의 프레임워크로, EDA에 대해서는 '다변량 비그래픽' 유형을 중점으로 제시한다.

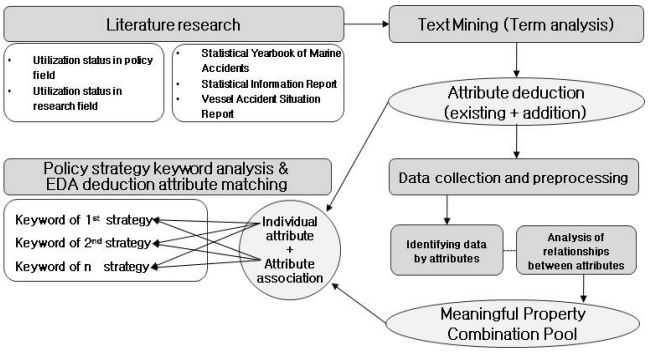

Fig. 3. Research Framework.

3.2 문제 및 데이터 정의

앞서 문헌연구 고찰을 통해 통계의 중요성 및 활용한계를 알아보았다. 선박사고(안전) 관련 정책이나 선행연구에서 주요하게 사용되는 항목은 사고발생과 관련한 결과변수들이며, 수난대비기본계획에서도 목표나 전략수립, 평가에 있어 사고 발생건수와 인명피해자 수 등을 사용한다.

하지만 위의 결과변수로만 어떤 현상을 설명하기는 충분하지 않다. 위 변수들은 목표 달성 여부를 판단할 수 있는 뚜렷한 결과 지표이기는 하지만, 실제로 그 목표를 달성하기 위한 전략별 세부과제의 효과인지를 파악하기 위한 적정 지표인지는 알 수 없다. 사고와 인명피해 감소를 위해 수행하는 전략 과제들은 무수히 많고 세부 내용 또한 그 특징이 다르다.

연구 분야에 사용되는 항목도 극히 제한적인 정보를 담는 통계항목을 활용하기 때문에 새로운 패턴이나 구체적인 개선방안을 모색하는데 한계가 있다.

따라서 이러한 통계 정보의 한계를 개선하기 위해 정책 및 연구 분야에 사용하는 기존 개방 통계(속성) 외에 세부 정책별 성과지표로 활용 가능하며, 연구 과정의 각종 분석에도 적용 가능한 새로운 속성들을 탐색해보고자 한다.

연구 전반에 사용한 데이터의 특성 및 수집, 전처리, 분석 방법 및 수단은 다음과 같다.

[데이터 속성] - 수집 한계 항목은 제외
- 현행 해상조난사고 국가통계 항목(해양경찰청 승인통계)
- 선박안전(해상안전) 관련 정책에서 사용되는 통계 및 관련 데이터
- 해양경찰청 상황보고서 텍스트 분석을 통해 추가적으로 통계화(지표화) 가능할 것으로 판단되는 속성

[데이터 수집]
- 해양경찰서에서 작성한 상황보고서(행정 자료)로부터 위에서 언급한 3개 관련 속성 값을 추출하여 Excel 데이터 셋 구성
- 추가로 탐색된 속성이더라도 사고 사례별(case) 결측치 문제로 데이터 수집 불가능 항목은 제외

[데이터 처리, 분석 방법 및 도구]
- Excel에서 수집된 속성 별 결측치, 이상치, 기타 오류 수정
- 기술통계량을 확인하여 필요시 변수 가공 (Feature Engineering)
- 전처리 데이터를 활용해 EDA 수행 (다변량 분석을 중점으로 진행하여 상관성 높은 변수 조합 탐색)
- 분석 도구 : Excel과 R을 사용해 분석 진행

4. 탐색적 데이터 분석

4.1 텍스트 마이닝 기반 추가 속성 도출 및 데이터 수집

해상조난사고 통계는 해양경찰서에서 작성하는 사고 상황보고서로부터 집계된다. 상황보고서는 사고 발생부터 종료까지의 모든 대응사항을 담고 있는 행정 자료로, 내용을 직접 서술하는 비구조적인 서식이다.

비구조적인 서식은 구조화된(정형화된) 서식에 비해 정보 오류나 누락 등의 문제가 발생할 수 있지만, 정해진 틀의 정보 외에도 다양한 정보를 담을 수 있다는 장점이 있다. 따라서 해상조난사고와 관련한 추가 속성(통계) 발굴을 위해 상황보고서에서 표현되는 텍스트(용어)를 분석하면 기존 통계에서 수집되지 않는 속성을 발견할 수 있다.

본 연구에서는 선행연구(Lee et al., 2019)에서 상황보고서 텍스트 분석을 통해 도출한 분류체계를 차용해서 상관성이 있는 속성 조합을 발굴해보고자 한다. 선행연구에서 제시한 상황보고서 속성은 Table 2와 같이 10개 범주로 요약된다.

Table 2. Situation report word attribute classification

Attribute Classification	Specifics
Basic Information	serial number, document number, case title, number of reports
Initial Accident Information and Radio Wave Information	date & time of the accident, time of the report, time of the shipment, means of filing the report, the reporter, the current status of the ship (incident situation),
Accident Location Information	location of the accident, diameter, address of the accident
Vessel Information (Ship Spec)	vessels tonnage, vessel type, shipment, unregistered vessel, load type. number of passengers, number of foreigners, etc.
Weather / Sea Characteristics Information	weather condition, cloudy/clear, wind direction, wind speed, digging, visibility distance, high / low tide (hour)
Accident (Risk) Information	1st accident, 2nd accident, 3rd accident (chain risk, compound accident)
Damage Result Information	rescuer (fair condition), injured person, dead person, missing person, type of ship damage, type of spilled oil, runoff
Incident Response Information / Input Resource Information	initial deployed force, initial arrival time, moving distance, additionally deployed force, input equipment / resources, response measures, system availability
Cause of the Accident	response vulnerability, alcohol intake, unlicensed, incident vulnerability (final cause of accident), oil spill cause
End of Incident Information	accident end date, accident end time, accident end location, final process results

각 속성에 대한 EDA를 위해서는 데이터를 수집해야한다. 데이터 수집은 텍스트 분석을 통해 도출된 속성에 대해 진행하며, 상황보고서에서 각 속성에 해당하는 값들을 추출해 엑셀(excel)로 DB화 한다. 이때 DB의 열(column)은 정의된 속성들이 되고, 행(row)은 개개의 사고 사례(case)가 된다.

4.2 데이터 파악

상황보고서로부터 수집한 사례(case) 수는 1,319개이며, 도출된 속성 중에서 결측치가 많아 수집에서 제외된 속성들을 제외하고 총 56개 속성 값을 추출해 분석에 활용했다. 수집한 속성별 데이터 수는 상황보고서에서 기록되지 않은 경우가 있어 개별 속성마다 결측치를 제외한 수집 case에는 차이가 있다.

수집된 데이터 중에는 분포나 결측치 등을 확인하여 변수조작(Feature Engineering) 과정을 거친 속성도 있다. 이는 범주형 구간을 재조정 하거나(예를 들어 1월~12월까지 月을 봄, 여름, 가을, 겨울의 계절적 특성 혹은 분기로 구분), 상위척도를 낮은 수준의 척도로 변환(비율척도인 부상자 수를 명목척도인 부상자 유무로 변경), 서로 다른 속성끼리 결합하여 새로운 변수를 만드는(신고 접수시간부터 현장 도착시간까지 시간 계산을 통해 골든타임 내 초기대응시간 여부 확인 등) 작업들이 포함된다.

EDA에 사용한 속성은 그 유사성을 고려하여 범주를 구분하였으며 속성 데이터별 타입(형태)과 척도, 수집된 사례 수(case)를 다음 Table 3과 같이 정리했다.

Table 3. Data definition for each attributes
(Type: Ca-Categorical, Nu-Numerical, Scale: N-Nominal, O-Ordinal, R-Ratio)

Attribute Category		Name of Attribute	Type	Scale	Case
Accident Induction Variable	Weather Situation	Wind Speed	Ca	O	728
		Wave & Speed (2~2.5m or 10~15m/s)	Ca	O	1319
		Visibility	Ca	O	705
		Weather Condition	Ca	N	1319
	Cause of Accident	Cause of the Accident	Ca	N	477
Accident Response Variable	Report Related	Means of Filing the Report	Ca	N	849
		The Reporter	Ca	N	1040
		Time of the Report	Ca	O	1299
		Stopover Number of Report Route	Nu	R	1319
	Response Activity	Initial Response(Status)	Ca	N	1319
		Moving Distance	Nu	R	386
		Response Vulnerability	Ca	N	1319
	Response Time	Receipt~Command Time	Ca	O	1182
		Command~Site Arrival Time	Ca	O	1139
		Receipt~Site Arrival Time	Ca	O	1142
		Total Response Period	Ca	O	1247
		Number of Reports	Nu	R	1319
	Committed Resource (Force)	Number of Initial Command Force	Nu	R	1319
		Number of Standby Force	Nu	R	1319
		Number of Actual Arrival Force	Nu	R	1319
		Initial Arrival Force	Ca	N	1212
		Number of Additionally Deployed Forces	Nu	R	1319
		Number of Total Coast Guard	Nu	R	1319
		Number of Civilian Deployed Forces	Nu	R	1319
		Number of Public Deployed Forces	Nu	R	1319
		Number of the Rest Deployed Forces	Nu	R	1319
		Number of Total Deployed Forces	Nu	R	1319
		Coast Guard Actual Arrival rate	Nu	R	1164
Accident Related Data Variable	Character-istic of Time	Year	Ca	N	1319
		Month	Ca	N	1319
		Time of the Accident	Ca	N	1303
	Character-istic of Space	Region	Ca	N	1319
		Address of the Accident	Ca	N	1298
		Location of the Accident	Ca	N	1313
	Accident Target	Target of Accident Report	Ca	N	1319
		Vessel Tonnage	Ca	N	1287
		Vessel Type	Ca	N	1304
		Number of Passengers	Nu	R	1287
Accident Result Variable	Accident Type	1st Accident	Ca	N	1296
		2nd Accident	Ca	N	1318
		3rd Accident	Ca	N	1319
		A Series Accident(Y/N)	Ca	N	1318
	Human Damage	Rescuer	Nu	R	1319
		Injured Person	Nu	R	1319
		Injured Person(Y/N)	Ca	N	1319
		Dead Person	Nu	R	1319
		Dead Person(Y/N)	Ca	N	1319
		Missing Person	Nu	R	1319
		Missing Person(Y/N)	Ca	N	1319
		Damage of Human Lives(Y/N)	Ca	N	1319
		Deaths, Missing(Y/N)	Ca	N	1319
	Material Damage	Material Damage Type	Ca	N	1133
		Material Damage(Y/N)	Ca	N	1133
		Towing(Salvage)(Y/N)	Ca	N	1319
	Pollution Damage	Oil Outflow(Y/N)	Ca	N	1319
		Natural Extinction(Y/N)	Ca	N	1319

4.3 속성 간 관계분석

앞서 정의한 속성 별 척도를 고려하여 상관계수를 도출해 상관성이 있는 조합들을 추려냈다. 상관계수는 상관성 정도를 나타내는 지표로 -1~1 사이의 값을 가지며, 통상 0.4 이상일 때 중간 정도의 상관관계, 0.7 이상일 때 강한 상관관계를 가지는 것으로 해석한다.

상관분석에서 속성 별 척도 조합이 명목-명목, 명목-연속인 경우 cramer's V, 서열-서열인 경우 spearman, 서열-비율, 비율-비율인 경우 pearson 상관계수를 사용했다.

분석 결과 0.4 이상의 상관성을 보이는 속성 조합을 다음 Table 4와 같이 정리하였으며, 관계에 대한 유의성 검증을 위해 유의확률(p-value < .05)을 함께 확인했다.

0.7 이상의 강한 상관관계를 가지는 속성 조합은 18개이며, 모두 p-value < 0.05로 유의한 것으로 나타났고, 그 외 중간 정도의 상관관계(0.4 이상~0.7 미만)를 가지는 속성 조합은 81개 중 70개가 유의한 것으로 나타났다. 따라서 99개의 속성 조합 중 11개를 제외한 88개의 속성이 서로 연관성이 있는 것으로 결론지을 수 있다.

Table 4. Correlation analysis result among attributes
(c.c: correlation coefficient, p-v: p-value)

Correlation Attribute Combination		case	c.c	p-v
Wind speed	Wave & Speed (2~2.5m or 10~15m/s)	728	0.706	0.000
Weather Condition	Cause of the Accident	477	0.504	0.000
	Wind Speed	728	0.706	0.000
	Wave & Speed (2~2.5m or 10~15m/s)	1319	0.725	0.000
Cause of the Accident	Number of Passengers	457	0.416	0.000
	Vessel Tonnage	466	0.419	0.000
	Vessel Type	470	0.432	0.000
	Target of Accident Report	477	0.432	0.000
	Towing(Salvage)(Y/N)	477	0.452	0.000
	Material Damage Type	439	0.560	0.000
	1st Accident	469	0.682	0.000
	Material Damage(Y/N)	439	0.748	0.000
The Reporter	Oil Outflow(Y/N)	1040	0.412	0.000
Initial Response (Status)	Towing(Salvage)(Y/N)	1319	0.401	0.000
	1st Accident	1296	0.483	0.000
	Oil Outflow(Y/N)	1319	0.489	0.000
	Material Damage Type	1133	0.539	0.000
	Material Damage(Y/N)	1133	0.756	0.000
Moving Distance	Address of the Accident	383	0.458	0.029
	Initial Arrival Force	384	0.470	0.579
	Region	386	0.472	0.549
	Initial Response(Status)	386	0.481	0.317
	Vessel Type	383	0.485	0.428
	Month	386	0.487	0.183
	Total Response Period	382	0.490	0.456
	Time of the Accident	383	0.500	0.009
	Time of the Report	384	0.510	0.001
	Vessel Tonnage	383	0.511	0.002
	1st Accident	380	0.513	0.003
	Visibility	332	0.521	0.385
	Location of the Accident	386	0.526	0.000
	Year	386	0.535	0.223
	Wave & Speed (2~2.5m or 10~15m/s)	386	0.538	0.034
	Means of Filing the Report	333	0.541	0.123
	The Reporter	360	0.547	0.000
	Receipt~Command Time	386	0.559	0.021
	Material Damage Type	359	0.573	0.001
	Command~Site Arrival Time	378	0.586	0.000
	2nd Accident	386	0.625	0.118
	Receipt~Site Arrival Time	378	0.627	0.000
	Cause of the Accident	125	0.633	0.505
Command~ Site Arrival Time	Receipt~Site Arrival Time	1139	0.878	0.000
Number of Reports	Number of Actual Arrival Force	1319	0.405	0.000
	Receipt~Site Arrival Time	1142	0.449	0.000
	Number of Additionally Deployed Forces	1319	0.472	0.000
	Number of Total Deployed Force	1319	0.500	0.000
	Number of Total Coast Guard	1319	0.569	0.000
Number of Initial Command Force	Number of Actual Arrival Force	1319	0.798	0.000
Initial Arrival Force	Coast Guard Actual Arrival rate	1164	0.410	0.000
Number of Coast Guard	Initial Arrival Force	1212	0.499	0.000
	Number of Initial Command Force	1319	0.616	0.000
	Number of Additionally Deployed Forces	1319	0.756	0.000
	Number of Actual Arrival Force	1319	0.782	0.000
	Number of Total Deployed Force	1319	0.853	0.000
Number of Total Deployed Force	Number of Public Deployed Force	1319	0.450	0.000
	Number of Initial Command Force	1319	0.533	0.000

Correlation Attribute Combination		case	c.c	p-v
	Number of Additionally Deployed Forces	1319	0.640	0.000
	Number of Actual Arrival Force	1319	0.671	0.000
Address of the Accident	Missing Person(Y/N)	1298	0.710	0.006
	Missing Person	1298	0.710	0.006
	Region	1298	0.911	0.000
Target of Accident Report	Material Damage Type	1133	0.466	0.000
	Vessel Tonnage	1287	0.577	0.000
	Vessel Type	1304	0.577	0.000
	Material Damage(Y/N)	1133	0.646	0.000
	1st Accident	1296	0.673	0.000
	Oil Outflow(Y/N)	1319	0.686	0.000
Vessel Tonnage	Towing(Salvage)(Y/N)	1287	0.454	0.000
	Material Damage Type	1118	0.476	0.000
	Number of Rescuer	1287	0.484	0.000
	Number of Passengers	1264	0.507	0.000
	1st Accident	1264	0.511	0.000
	Oil Outflow(Y/N)	1287	0.556	0.000
	Material Damage(Y/N)	1118	0.671	0.000
	Vessel Type	1283	0.692	0.000
Vessel Type	Towing(Salvage)(Y/N)	1304	0.407	0.000
	Material Damage Type	1124	0.478	0.000
	Number of Rescuer	1304	0.489	0.000
	Number of Passengers	1276	0.514	0.000
	Oil Outflow(Y/N)	1304	0.527	0.000
	1st Accident	1281	0.609	0.000
	Material Damage(Y/N)	1124	0.670	0.000
Number of Passengers	Visibility	698	0.402	0.001
	Number of Rescuer	1287	0.998	0.000
1st Accident	Injured Person(Y/N)	1296	0.477	0.000
	Damage of Human Lives(Y/N)	1296	0.480	0.000
	Towing(Salvage)(Y/N)	1296	0.560	0.000
	Material Damage Type	1112	0.755	0.000
	Oil Outflow(Y/N)	1296	0.806	0.000
	Material Damage(Y/N)	1112	0.891	0.000
2nd Accident	Material Damage Type	1132	0.416	0.000
3rd Accident	Missing Person(Y/N)	1319	0.410	0.005
	2nd Accident	1318	0.431	0.000
	Missing Person	1319	0.516	0.000
Damage of Human Lives(Y/N)	Injured Person	1319	0.638	0.000
	Injured Person(Y/N)	1319	0.950	0.000
Material Damage Type	Towing(Salvage)(Y/N)	1133	0.424	0.000
	Oil Outflow(Y/N)	1133	0.491	0.000
Material Damage (Y/N)	Towing(Salvage)(Y/N)	1133	0.420	0.000

5. 정책적 활용 방안

EDA를 통해 선박사고 관련 기존 통계 항목 대비 다양한 속성들이 존재함을 확인했다. 추가적으로 발굴된 속성들을 실제 정책(업무)에 활용하기 위해 전략별 키워드 분석과 EDA 도출 속성 간 매칭 작업이 필요하다.

5.1 수난대비기본계획 세부 전략 키워드 분석

먼저 조직에서 수립한 계획(문서)의 콘텐츠를 분석하여 전략별 과제의 측정 가능한 핵심어(keyword)를 추출해야 한다. 여기서 측정 가능한 핵심어란 EDA를 통해 발견된 속성과 매칭할 수 있는 내용을 의미한다.

Table 5는 해양경찰청에서 작성한 2차 수난대비기본계획에서 '목표 지향적 수색구조 활동 및 맞춤형 해양사고 대비' 전략 중 '목표 지향적 수색구조 활동' 추진과제 상세 내용에 대한 키워드 분석을 보여준다.

Table 5. Key word analysis samples for each detail task

Promotion Strategy : Goal oriented search and rescue activity and customized marine accident preparation.		
Promotion Assignment : Goal oriented search and rescue activity.		
Detailed assign -ment	Mobilization time goal system and arrival time management system	- Mobilization time goal system : set the least amount of time from mobilization command receipt time of to mobilization by rescue force - Arrival time management system : continually management shorten the arrival time by setting the arrival time from rescue force departure time to arrival time near the sea area **(Key word)** Rescue Force, Receipt Time, Mobilization Time, Least Amount of Time, Required Time, Departure Time, Arrival Time, Mobilization Distance
	Rescue rate of human lives and vessel etc. goal setting and management	- Setting sustainable search and rescue result goal (rescue rate of human lives, rescue rate of vessel, within 1 hour response rate) **(Key word)** Rescue of Human Lives Rate, Rescue of Vessel Rate, Response Rate (Within 1 Hour)

대부분 정부에서 작성하는 기본계획은 1개의 추진전략 아래 여러 추진과제로 구성되며, 추진과제 아래에도 여러 세부과제들이 포함되기 때문에 키워드 분석 시 상위 전략이나 과제가 아닌 하위의 세부 과제별로 내용을 분석해야 개별 특성을 반영한 키워드 도출이 가능하다.

5.2 전략별 키워드와 EDA 도출 속성 간 매칭

다음으로 EDA을 통해 도출된 속성과 전략별 핵심어 비교를 통해 서로 매칭 가능한 부분을 확인하는 것이다.

세부과제 중 '출동시간 목표제 및 도착시간 관리제'에 대한 세부 키워드를 살펴보면 '구조세력, 접수시간, 출동시간, 최단시간, 소요시간, 출발시간, 도착시간, 출동거리'로 요약된다. 이는 구조세력별로 출동지령 접수시간부터 출동까지 최단시간을 정하고 현장까지의 도착시간을 관리하여 신속하게 대응하기 위함으로 볼 수 있다.

세부 과제 키워드와 관련하여 '시간'과 '출동거리'를 중점으로 매칭할 수 있는 유의미한(p-value < .05) 속성조합을 정리해보면 Table 6과 같다.

Table 6. Keyword for strategy - EDA drawn attributes matching sample

Detailed assignment : Mobilization time goal system and arrival time management system		
(Key word) Rescue Force, Receipt Time, Mobilization Time, Least Amount of Time, Required Time, Departure Time, Arrival Time, Mobilization Distance		
Correlation Attribute Combination		**Correlation Coefficient**
Moving Distance	Address of the Accident	0.458
	Time of the Accident	0.500
	Time of the Report	0.510
	Vessel Tonnage	0.511
	1st Accident	0.513
	Location of the Accident	0.526
	Wave & Speed (2~2.5m or 10~15m/s)	0.538
	The Reporter	0.547
	Receipt~Command Time	0.559
	Material Damage Type	0.573
	Command~Site Arrival Time	0.586
	Receipt~Site Arrival Time	0.627
Command~Site Arrival Time	Receipt~Site Arrival Time	0.878

해당 전략에 활용할 수 있는 유의미한 속성 조합을 살펴보면 주로 시간, 공간, 기상(취약성), 사고결과(사고종류 및 피해유형) 등과 관련한 변수 조합으로 구성된 것을 알 수 있다. 즉, 해당 속성(조합)들이 (초기)대응시간을 관리하기 위해 고려해야할 요소들로 해석해 볼 수 있다(단, 상관관계가 있다고 해서 인과관계가 있다고 해석할 수는 없다).

6. 연구결론 및 향후 연구 과제

통계는 각종 의사결정을 뒷받침할 수 있는 강력한 증거 중 하나로 정책과정이나 각종 연구에 활용할 목적으로 생산된다. 특히 분석 및 설득을 위한 기초자료가 된다는 점에서 통계의 역할은 매우 크지만 실제로 이러한 통계의 중요성에 비해 활용도는 제한적인 수준이다. 현재 개방된 통계는 단순 결과 요약 자료 수준이며 공급자 위주로 생산되어 수요자 관점에서 가치 창출을 위한 수단으로는 부족한 측면이 있다.

따라서 이러한 문제를 보완하기 위해 현재 제공되는 통계 항목 외에 정책이나 연구에 다양하게 활용할 수 있는 추가 속성을 탐색했다. 본 연구에서는 해양경찰청에서 발간하는 해상조난사고 통계 항목을 보완하기 위한 추가 속성을 도출하였으며, 해양경찰서에서 작성하는 선박사고 상황보고서 텍스트 분석을 통해 추가할 수 있는 속성들을 수집했다.

EDA 결과, 유의확률(p-value < .05)을 만족하는 상관계수 0.7 이상의 강한 상관관계가 있는 속성 조합 18개와 중간 정도의 상관관계(0.4 이상 0.7 미만)를 가지는 속성조합 70개, 총 88개의 조합을 발굴했다.

이 속성들을 실제 정책에 적용할 수 있는지를 검토하기 위해 수난대비기본계획 세부 전략 키워드 분석을 실시하고, EDA를 통해 도출한 속성들과의 매칭을 통해 전략에 활용할 수 있는 속성들을 찾아보았다.

행정 자료로부터 다양한 속성을 탐색하고 의미 있는 정보를 발견하는 과정을 통해 조직은 크게 4가지 효과를 기대할 수 있다. 첫 번째, 기존 승인통계 항목을 구체화하거나 보완할 수 있고 두 번째, 전략 과제들을 평가하기 위한 적합한 속성이 있다면 지표로 활용할 수 있다. 세 번째로 승인통계로 활용하지 않더라도 내부 업무 수행에 있어 더욱 구체적이고 깊은 분석 자료로서 가치가 있으며, 마지막으로 발견한 속성이 현재 수집되지 않고 있는 속성이더라도 정보로서 활용 가치가 있다면 해당 속성을 지속적으로 수집하기 위한 시초가 될 수 있다.

본 연구는 해양경찰청 상황보고서에서 발견할 수 있는 새로운 속성들로부터 의미 있는 정보를 발굴하기 위한 탐색적 성격을 지니고 있다. 이 연구는 기존에 개방된 해상조난사

고 통계 정보 외에 새로운 속성들과 유의미한 조합들을 제시하여 정책 및 연구 분야에서 활용성을 높이기 위한 기반을 제시했다는데 가장 큰 의미가 있다.

하지만 비구조적인 상황보고서로부터 데이터를 수집하고 EDA를 수행함에 있어 다음과 같은 몇 가지 한계점을 가진다. 가장먼저 분석에 사용한 사례(case) 수가 제한적이다. 분석을 위해서는 1개의 case에 모든 속성 값이 누락되지 않고 채워져 있어야 하지만 상황보고서는 비구조적 서식이기 때문에 작성자마다 작성 내용이 달라 수집 불가능한 속성 값이 생긴다. 이러한 문제로 전체 속성 값이 모두 수집된 사례 수가 적다는 문제가 있고, 실제 속성 간 분석 시 그 속성 값이 모두 갖춰진 사고 case만을 대상으로 분석을 진행해야 했다.

때문에 더 많은 case로 EDA를 수행할 경우 또 다른 상관성을 발견할 수도 있다. 물론 분석결과 해석에 있어 발견된 상관관계가 실제 상황의 이해에 도움이 되지 않을 수 있다. 자료 규모가 커지게 될 경우 통계적으로 유의한 상관관계를 찾을 가능성이 높아지기 때문이다. 따라서 자료의 속성과 분석 방법 선택, 분석 틀에 대한 많은 질문이 필요하다.

또한 비구조적인 상황보고서로부터의 데이터 활용은 데이터 수집(결측치) 뿐만 아니라 정체(전처리) 및 분석기법 결정에도 영향을 미친다. 보고서가 서술형으로 작성되기 때문에 작성자마다 동일한 의미에 대해 다양한 단어, 기호, 약어, 단위를 사용하고 있다. 따라서 수집한 데이터를 활용하기 위해서는 분석에 적합한 형태로 데이터를 정제(전처리)하는 과정이 반드시 필요하다.

빅데이터 시대, 데이터의 중요성이 커짐에 따라 데이터 기반 접근 방식은 수많은 분야에서 의사결정 과정을 지원해 왔다. IT의 발전으로 이제는 방대한 규모의 데이터 처리가 가능해지고, 다양한 유형의 데이터를 분석할 수 있게 되면서 기존보다 다양한 속성들을 정책 결정에 활용할 수 있게 되었다.

문제는 여전히 많은 조직이 데이터에 대한 가치를 인지하지 못하고 있다는 것이며, 이로 인해 데이터를 활용하기 위한 체계도 마련되지 않은 것이 현 실정이다. 조직에서 생산된 데이터는 그 조직 활동 결과에 대한 사실(fact)을 담고 있기 때문에 생산 시점부터 수집, 가공, 분석만 잘하면 의미 있는 정보와 지식을 창출해 각종 의사결정에 활용할 수 있다.

이제는 어떤 조직이든 생산된 데이터를 무의미하게 버리지 않고 활용할 수 있도록 데이터를 탐색하고자 하는 노력이 필요한 시점이다. 또한 향후에는 데이터로부터 발견된 새로운 정보를 바탕으로 가설을 세우고 테스트하기 위한 확증적 데이터 분석을 통해, 도출 속성에 대한 신뢰성을 증대시킬 수 있어야 할 것이다.

References

[1] Anderson, C.(2008), The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, Vol. 16, No. 7.

[2] Ahn, T. H.(2015), Data compilation methods through the use of administrative data: Specifically analysed in the field of the Mining and Manufacturing Industry Survey, Korea University Graduate School of Public Administration.

[3] Behrens, J. T.(1997), Principles and procedures of exploratory data analysis. Psychological Methods, Vol. 2, No. 2, pp. 131-160.

[4] Chae, C. J., Y. S. Park, S. H. Jo, S. Y. Kang, H. Lee, and H. B. Kim(2019), A Study on the Emergency Response Empowerment for Captain Based on the Analysis of Maritime Accidents, Journal of the Korean Society of Marine Environment and Safety, Vol. 25, No. 4, pp. 413-422.

[5] Cho, H. K., B. S. Park, D. H. Kang, and S. S. Kim(2017), The Main factor and Counterplan for Marine accidents in Korea, Journal of fishries and marine sciences education, Vol. 29, No. 3, pp. 746-756.

[6] Choi, J. Y.(2016), Toparchy occupation statistics writing study through administRn data matching, Korean University Graduate School paper of masters degree.

[7] Good, I. J.(1983), The philosophy of exploratory data analysis. Philosophy of science, Vol. 50, No. 2, pp. 283-295.

[8] Hong, J. U.(2015), A Study On data Fusion Using Statistical Matching, Sungkyunkwan University.

[9] Howlett, M.(2009), Policy analytical capacity and evidence based policy making: Lessons from Canada, Canadian public administRn, Vol. 52, No. 2, pp. 153-175.

[10] Jang, W. J. and J. S. Keum(2004), An Analysis on the Models of Occurrence Probability of Marine Casualties, Journal of The Korean Society of Marine Environment & Safety, Vol. 10, No. 2, pp. 29-34.

[11] Kim, D. S.(2018), A Study on the Prevention of Ship Collision in Low Visibility: Focusing on the Role of Korea Coast Guard, Korean Association of Maritime Police Science, Vol. 8, No. 3, pp. 71-85.

[12] Kim, J. Y.(2016), Hello, DATA SCIENCE, Hanbit Media.

[13] Kwon, D. C.(2017), Statistics is not just numerical value, Policy, Health and welfare forum, korea health and social affairs researcher, Vol. 250, No. 1, pp. 2-4.

[14] Lee, E. G.(2017), Agricultural statistics writing technique advancement way utilizing administRn data - mainly for

fishing industry total investigation and fishery business, Korea University Graduate School of Public AdministRn paper of masters degree.

[15] Lee, K. H.(2016), A Study on the Actual Condition and the Countermeasure of Marine Accidents, Korean Association of Police Science, Vol. 18, No. 6, pp. 27-54.

[16] Lee, K. J., M. K. Kim, J. Y. Ahn, and K. H. Choi(2012), A case study on the selection of representative statistics for systematic management of administrative statistics, Journal of the korean data&information science society, Vol. 23, No. 1, pp. 63-70.

[17] Lee, Y. J., S. K. Kang, and J. Y. Gu(2019), A Study on Marine Accident Ontology Development and Data Management: Based on a Situation Report Analysis of Southwest Coast Marine Accidents in Korea, Journal of the Korean Society of Marine Environment and Safety, Vol. 25, No. 4, pp. 423-432.

[18] Lee, Y. J., S. K. Kang, and J. Y. Gu(2020), The Initial Reaction Analysis by Ocean Safety Information Classification System : Focused on Boating Accidents of the Central Part Seas, Korean Association of Maritime Police Science, Vol. 10, No. 1, pp. 67-86.

[19] National Statistical Office(2020), 2020 statistics based Policy Evaluation, Daejeon: National Statistical Office.

[20] Noh, C. K.(2002), A Study on the Developments of the Salvage & Oil Spills Response, Journal of Navigation and Port Research, Vol. 26, No. 6, pp. 549-554.

[21] Oh, S. Y., K. Yoon, and K. Oh(2017), present situation research about Government Statistics Establishment and utilization for Evidence-based policy, Korea Institute of Public Administration.

[22] Park, B. S.(2018), AdministRn data and research data matching by statistical technique, Hannam University Graduate School paper of masters degree.

[23] Park, T. G., S. J. Kim, Y. S. Chu, T. S. Park, K. J. Ryu, and Y. W. Lee(2018), Reduction plan of marine casualty for small fishing vessels, Journal of the Korean Society of Fisheries and Ocean Technology, Vol. 54, No. 2, pp. 173-180.

[24] Seltman, H. J.(2018), Experimental design and analysis, pp. 61-100.

[25] Seo, M. S. and S. J. Bae(2002), The Study on the Analysis of Marine Accidents and Preventive Measures, Journal of fishries and marine sciences education, Vol. 14, No. 2, pp. 149-160.

Received : 2020. 12. 02.
Revised : 2020. 12. 16. (1st)
: 2020. 12. 23. (2nd)
Accepted : 2020. 12. 28.