이학석사 학위논문

선박사고에 영향을 주는 자연적 요인에 관한 통계 분석

Statistical Analysis on Natural Factors
Affecting Ship Accidents

지도교수 박 찬 근

2017 년 2 월

한국해양대학교 대학원

데이터정보학과

이 재 익

본 논문을 이재익의 이학석사 학위논문으로 인준함.

위원장 김 재 환 (인)

위 원 박 찬 근 (인)

위 원 장 길 웅 (인)

2017년 11월 28일

한국해양대학교 대학원

목 차

List of Tables

List of Figures

Statistical Analysis on Natural Factors Affecting Ship Accidents

Lee, Jae Ik

Department of Data Information
Graduate School of Korea Maritime University

Abstract

The ship accidents were caused by various factors. Various factors consist of human factors and natural factors. In this paper, we focus on natural factors. The purpose of this paper is to analyze ship accidents in Busan and Incheon and analyze the effects of natural factors on ship accidents using logistic regression analysis and decision trees. The Korean Peninsula is surrounded by three different seas, and each sea has different natural factors. The data of natural factors were obtained from the Korean Maritime Safety Tribunal and the Korea Meteorological Administration. Each data of tow factors was combined. Ship accident data, general weather data, and marine weather data were refined by propensity score matching(PSM) method. We observed that significant digging, wind speed, daily precipitation, and barometric pressure affected ship accidents in Busan. Also we found that wind speed, daily precipitation, and temperature affected ship accidents in Incheon. In future studies, the factors of ship accidents should be modeled by considering the regional difference.

KEY WORDS: Decision tree, Logistic regression, Natural factors, Propensity score matching, Ship accidents

제 1 장 서 론

2014년 4월 16일 세월호 참사는 안전관리의 현실을 보여주며 큰 충격을 몰고 온 사건이다. 절대 일어나지 않았어야 하는 사건으로 남아있다. 이 사고는 한 가지 원인으로 사고가 일어났다고 볼 수 없다. 다양한 원인이 결합되어 나타난 사고이며 기존 분석처럼 한 가지 원인으로만 사고가 일어났다고 할 수 없다. 기존 선박사고의 원인을 살펴보면 충돌 사고가 가장 많은 부분을 차지하고 좌초, 화재/폭발 등이 그 뒤를 잇고 있다. 하지만 이런 사고 발생의 근본적인 원인은 대부분 주의 태만, 상황인식 오류, 불충분한 훈련 등의 인적 오류라고 분석해오고 있다.

현대에 들어 해양운송수단의 비중이 크게 차지하고 있으므로 해양수송수단에 대한 위험도 및 안정성에 대한 평가가 중요한 문제가 되고 있다. 해양운송수단의 발달은 이러한 분야에 연구를 더욱 가속화 시키고 있다. 선박이 대형화되고 고속화 되었으며 선박의 운송품목이 크게 늘어났다. 이러한 발달로 인해 해양운송수단은 효율성과 사고 발생 시 큰 피해를 가져올 수 있는 위험성을 동시에 가지게 되었다. 따라서 이러한 선박의 발달과 운송에 대한 연구와 병행하여 안정성에 대한 연구가 필요하다 (곽수용, 2011).

전 세계적으로 해운산업이 계속적인 성장을 보이고 있다. 이와 더불어 우리나라 연안 해상물동량 역시 증가하는 추세에 있다. 그리고 해상교통량의 증가와 조선기술의 발달로 선박은 대형화, 고속화, 전용화 되고, 해양사고의 양상은 복잡화, 대형화 추세를 보이고 있다. 이로 인한 해상에서의 인명과 재산의 손실은 물론 환경오염도 큰 문제가 되고 있다.

국제해사기구(International Maritime Organization, IMO)에서는 해상에서의 안전을 확보하기 위해 아래의 Table 1에서처럼 각종 국제협약을 제정하여 선박안전관련 설비, 기준, 자격, 절차들을 규정함으로써 선박의 안전운항, 승무원과 여객의 인명보호, 오염으로부터 해양환경보호를 위해 많은 노력을 기울이고 있다.

그러나 이러한 노력에도 불구하고 해양사고는 지속적으로 발생하고 있다. 특

히 80년대 후반과 90년대 초반의 대형 해양사고로 인한 대형인명손실 및 환경오염의 발생은 선박안전에 대한 의식재고의 필요와 이를 극복하기 위한 방안을 찾고자하는 세계적인 움직임을 발생시켰다.

Table 1 선박 사고에 관한 국제 협약

선명	사고발생연도 / 장소	사고개요	관련 국제협약
Titanic	1912 / 북대서양	산과 충돌, 침몰하여 1,500여명	해상인명안전협약 (SOLAS)
Torry Canyon	1967 / 영국 근해	좌초로 인해 선적된 119,000톤의 원유를 Dover해협에 유출 편의치적선의 선원자질 저하가 문제화됨	해상오염방지협약 (MARPOL) 선원의 훈련, 자격증명 및 당직근무 기준에 관한 국제협약
Herald of Free Enterprise	1987 / 벨기에 지브로 근항 인근	Bow Door가 개방된 채 출항, 침수로 인한 전복으로 188명 사망	ISM Code 탄생배경이 된 해양사고 SOLAS 강화됨
Exxon Valdez	1989 / 미국, 알라스카 프린스 윌리암 수로	항해 중 좌초로 선적된 원유 45,000톤의 원유를 유출하고 30억불의 오염 피해 발생	OPA 90 (미국 오염방지법)
Scandinavian Star	1990 / 북해 해상	선실 화재로 159명의 여객 및 선원 사망	ISM Code 탄생을 가속화시킴

특히 사고원인의 심층 분석 결과에서 인적과실에 의한 것으로 밝혀짐에 따라 기술적인 측면만이 아닌 인적요인을 개선하여 선박을 안전을 향상시키고자 하는 방안의 마련에도 큰 관심이 모아졌다. 선박 사고에 대한 기존의 개선 방안들을 살펴보았을 때, 제도의 개선, 실제 종사자들에 대한 교육을 사고 예방에 대한 방법으로 제시되고 있는 것이 현실이다 (서용화, 2006).

선박사고에 대해서 그간 축적된 해양수산부의 공공데이터와 정부네트워크에 기반을 둔 대규모 데이터를 활용하여 새로운 국민서비스 제공이 가능하다. 빅데이터의 가공과 분석에 따라, 상황인식, 문제해결, 미래전망이 가능해지고 데이터가 경제적 자산과 경쟁력의 척도로 부각되고 있다. 광범위한 해상지역에서 생성되는 대용량 정보를 실시간으로 처리하여 선제적 대응, 분석을 통해 선박사고예방에 활용하려고 한다. 선박은 다른 교통과 다르게 바다에서 고립된 상태로 운행하는 특징을 가지고 있다. 그 때문에 한번 사고가 발생하게 된다면 인명구조 및 재산손해방지가 쉽지 않은 문제점이 있다. 과거 선박사고를 살펴보았을 때 선박사고의 구조 및 그 해결이 쉽지 않다는 것을 알 수 있다. 사고 후의 배상문제도 중요하지만 우선적으로 사고의 원인 규명을 철저히 하여 유사사고가 반복되지 않도록 예방 하는 것이 가장 중요하다.

기존 선박 사고에 대해서 구체적으로 여러 분석을 통해서 원인을 규명한 적이 없다. 간단한 빈도분석을 통하여, 사고 발생의 횟수를 표현. 이는 사고의 단편적인 모습만 확인 하는 것이며 기존 선박사고에 대한 통계적 분석을 바탕으로 합리적 예방 방안을 도출 할 수 있어야 한다. 선박사고의 원인을 한 가지 요인으로 나타내는 것이 아니라 다양한 요인의 관계를 분석하여 선박 운행 시 사고가 발생할 요인을 줄여야 한다.

해양안전심판원의 통계자료를 토대로 선박사고에 대한 원인을 빈도분석으로 살펴보면 대체적으로 인적과실에 의한 사고가 많이 발생한다. 인적과실의 유형은 다양하게 있는데 그 중에서도 '경계 소홀'이 가장 큰 원인이다. 항해기술이 발달하고, 법제도가 개선되고 있지만 선박사고는 지속적으로 늘어나고 있다. 자연적 요인은 통제가 가능한 부분이 아니기 때문이다. 따라서 본 논문에서는 부산, 인천의 연근해에 대해서 해양안전심판원의 선박 사고 자료와 기상청의

자료를 이용하였으며 이를 바탕으로 선박사고가 발생 하였을 때의 일반적 기상요인과 해양적 기상요인으로 분류된 자연적 요인을 분석하고 파악하여서 선박사고를 예방하려고 한다.

2. 이론적 배경

2.1 선박사고의 개념

선박사고는 현재 해양, 해난사고로 정의되고 있다. 대한 정의는 해양사고의 조사 및 심판에 관한 법률 제2조에 규정되어 있으며, 그 내용은 다음과 같다 (한국해양수산개발원, 2015).

해양사고란 해양 및 내수면에서 선박의 운용과 관련하여 발생한 아래의 경우를 포함한다.

가. 선박의 구조·설비 또는 운용과 관련하여 사람이 사망 또는 실종되거나 부상을 입은 사고를 의미한다. 여기서 선박의 구조·설비란 선박에 설치되어있는 모든 시설물을 말하며, 운용이란 선박의 항해, 정박, 하역 등 선박의 운항과 관련된 모든 것을 의미한다.

나. 선박의 운용과 관련하여 선박이나 육상시설·해상시설이 손상된 사고를 의미한다. 여기서 선박이란 수상 또는 수중을 항해하거나 항해할 수 있는 구조물을 말하며, 육상·해상시설이란 표지, 계선시설, 부두하역시설, 방파제, 해저전선, 어장 등과 같은 시설물을 말한다.

다. 선박이 멸실·유기되거나 행방불명된 사고를 의미한다.

라. 선박이 충돌·좌초·전복·침몰되거나 선박을 조종할 수 없게 된 사고를 의미한다.

마. 선박의 운용과 관련하여 해양오염 피해가 발생한 사고를 의미한다 (한국해양수산개발원, 2015).

위와 같이 5가지 관련하여 사고가 났을 경우 통계는 해양수산관서, 언론, 지자체에서 사고를 인지한 후 지방해양안전심판원에 보고하면, 지방해양안전심판

원에서 해양안전관리시스템에 입력한 자료를 집계한 자료이다 (한국해양수산개발원, 2015).

2.2. 선박사고의 종류

선박사고의 종류로는 Table 2와 같이 정리하였으며 충돌, 접촉, 좌초, 화재·폭발, 침몰, 기관손상, 조난, 인명사상 등의 종류가 있다.

또한 IMO의 기준에 따르면 선박의 운용과 관련하여 해양사고를 다음과 같이 분류하고 있다.

가. 사람이 사망하거나 중상을 입은 경우,

나. 사람이 행방불명된 경우,

다. 선박의 멸실, 추정멸실 또는 구조를 포기한 경우,

라. 선박에 손상이 발생한 경우,

마. 선박의 좌초, 운항불능 및 충돌,

바. 시설에 손상이 발생한 경우,

사. 환경 피해가 발생한 경우를 말한다.

다시 말하면 해양사고는 선박의 운용과 관련하여 발생한 선박, 인명, 시설피해 및 오염사고 등을 모두 포함하고 있으며, 해양사고 중 선박 간의 충돌사고는 교통사고와 유사한 외상에 의한 인명피해가 대부분이나 전복, 침몰 등의 사고는 익수로 인한 흡인성폐렴, 저체온증 등이 동반하여 중증 손상 가능성이 높다. 이 중 선박과 관련된 해양사고의 종류는 대부분 선박의 기관 손상 사고와 충돌 사고가 높은 비중을 차지하고 있다. 아래는 Table 2는 해양사고를 크게 8가지로 구분 하고 있다.

Table 2 선박사고의 종류

사고종류	내용
충돌	항해중이거나 정박 중임에도 불구하고 다른 선박에 부딪치거나 맞붙어 닿은 사고
접촉	다른 선박이나 해저를 제외하고 선박이 외부 물체나 외부의 시설물에 부딪치거나 맞붙어 닿아 손상이 생긴 사고
좌초	해저 또는 수면하의 난파선에 얹히거나 부딪혀 발생하는 사고
화재폭발	화재, 폭발사고는 최초의 사고로서 화재 또는 폭발이 발생하여 선박에 손상이 생긴 사고
침몰	기상악화 또는 외판 등의 균열, 파공, 절단 등에 의한 침수로 인해 선박이 가라앉은 사고
기관손상	주기관, 보조보일러 및 보조기기 등이 손상되어 발생된 사고
조난	기상악화, 부유물 접촉 또는 침수 등으로 선박에 손상이 발생 하였으나 기타 위의 어느 사건의 범주에 해당되지 않는 사고
사상	선박의 운용과 관련하여 사람이 사망, 행방불명되거나 부상당한 사고

2.3 상향점수매칭

상향점수매칭 (propensity score matching, PSM)은 Rosenbaum & Rubin (1983)이 제안한 것으로서, 외부요인에 대한 그룹간의 편향을 보정하기 위해 어렵지 않게 관찰할 수 있는 공변량 (covariate)을 이용하는 비모수적 방법론이다. 이 방법론의 주목적은 무작위과정을 통한 실험이 불가능한 상황에서 가상적 사실과 가장 유사한 비교집단을 만드는 것이다. 상향점수매칭 방법은 다음과 같은 2가지의 가정은 기초로 한다 (Rosenbaum, Robin, 1983; DEhijia & Wahba, 1998,1999; Zhao, 2000).

가정 1 : $(Y_0, Y_1) \perp Y | X$ [조건부 독립성의 가정(conditional independence assumption)].

가정 2 : $0 < \Pr(T+1|X) < 1$[공통영역의 가정(common support assumption)].

가정 1의 조건부 독립성의 가정은 공변량 X가 주어졌을 때, 처리 T의 할당 여부는 반응변수 Y_i들과 독립적이라는 가정이다. 이는 반응변수와 관련한 차이는 관찰된 변수에 의해 통제가 가능하고 따라서 관찰되지 않은 어떤 특성도 반응변수에 영향을 주지 않는다는 것을 의미한다. 가정 2의 공통영역의 가정은 분석 대상의 처리 T의 할당 확률은 공통 영역 내에 있다는 가정이다.

성향점수 (propensity score, PS) $P(X)$는 관찰된 공변량 X가 주어졌을 때, 관심집단에 할당 되도록 영향을 주는 독립변수에 대한 조건부 확률로서 정의된다. 이를 식으로 정리하면 다음과 같다.

$$\Pr opensity\, Score = P(X) = \Pr(T=1|P(X)) < 1. \qquad (2.3.1)$$

Rosenbaum & Rubin (1983)은 앞서 설명한 두 가지 가정이 만족 된다면 이를 X변수들의 함수로서 구해진 성향점수에도 적용할 수 있음을 증명하였고 다음과 같은 부명제로 정리했다.

부명제 1. $X \perp T|P(X)$.

부명제 2. $(Y_1, Y_2) \perp T|P(X), 0 < \Pr(T=1|P(X)) < 1$.

관심집단의 한 개체와 비교집단의 한 개체가 같은 성향점수의 값을 가진다면 이는 관심집단의 개체와 비교집단의 개체는 공변량 X에 대해 같은 분포를 가지는 것으로 생각할 수 있다. 다시 말해, 성향점수에 의한 짝짓기는 무작위과정을 통한 실험에서 얻은 결과와 같은 선택편향이 없는 결과를 추정 할 수 있음을 의미한다.

짝짓기를 수행하기에 앞서 성향점수를 추정해야 한다. 성향점수는 판별분석 또는 로지스틱 회귀분석을 사용하여 추정할 수 있다. 이 두 방법을 이용하면 관측된 공변량들이 주어진 조건 하에서 처리할당에 대한 확률의 추정치를 구할 수 있다.

성향점수의 추정을 마친 후에는 성향점수의 값을 비교하여 같은 값을 가진 짝을 만들어야 한다. 그러나 성향점수, 즉 $P(X)$는 연속적인 변수이므로 관심집단과 비교집단이 완전히 동일한 성향점수 값이 아니더라도 짝을 지을 수 있는 적절한 기준 및 방법이 필요하다.

이러한 문제의 해결을 위해서 Rosenbaum & Rubin (1983)은 최근접 거리 짝짓기 방법 (nearest available matching), 마할라노비스 거리 짝짓기 방법 (Mahalanobis metric matching), 반경 내 최근접 마할라노비스 거리 짝짓기 방법 등이 있다.

2.4 로지스틱 회귀분석

이번 논문은 종속변수는 사고 유무이며, 기상, 조류, 시간대, 시설, 선박 상황, 선박 크기 등 사고가 발생할 수 있는 독립변수들을 파악하여 기존 사고원인들에 대하여 분석하여 선박사고가 발생하는 유무에 대해서 예측 하는 것이다. 로지스틱 회귀분석은 다음 식과 같이 나타낼 수 있으며 Y가 사고유무이므로 1일 경우는 "예", 0일 경우는 "아니요" 일 조건으로 분석을 실시하여 그에 따른 확률 값을 추정하는 것이다.

$$P(Y=1|Y=0)=\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n. \tag{2.4.1}$$

위에 주어진 선형함수는 상한과 하한이 없으므로 확률을 모형화 하는 데에는 사용될 수가 없다. 통상적 최소제곱 (ordinary leasts squares, OLS)법이 사용될 수 없는 또 다른 이유가 있다. 반응변수 Y 는 이항 (binomial)확률변수로서, 따라서 그의 분산이 □의 함수가 될 것이며 ⁇에 의존한다. 즉 등분산 (equal variance; homoscedasticity)의 가정이 성립되지 않는다 (Chatterjee, Hadi, 2006).

로지스틱 회귀모형 (logistic regression analysis)에서의 모수의 추정은 일반적으로 최대우도법 (Maximum likelihood method)을 이용한다. n개의 자료가 있는 경우, n 쌍의 반응 변수와 설명변수를 (y_1, x_1), ⋯ , (y_n, x_n)로 표기하자. 모수에 대한 우도함수 (likelihood function)는 다음과 같이 주어진다.

$$L(\alpha,\beta_1,\cdots,\beta_p)=\prod_{i=1}^{n}\Pr(Y=y_i|X=x_i)=\prod_{i=1}^{n}\theta(x_i)^{y_i}(1-\theta(x_i))^{1-y_i}. \tag{2.4.2}$$

위의 우도함수에 대한 로그우도함수는 아래와 같다.

$$\begin{aligned}\log(L)&=\sum_{i=1}^{n}\log(\Pr(Y=y_i|X=x_i)) \\ &=\sum_{i=1}^{n}\left[y_i(\alpha+x_i^T\beta)+\log\left\{1+\exp(\alpha+x_i^T\beta)\right\}\right].\end{aligned} \tag{2.4.3}$$

식 (2.4.3)을 최대화하는 모수$(\alpha, \beta_1, \cdots, \beta_p)$는 해석적으로 구할 수 없으며, Newton-Raphson 방법 등을 이용한 반복 알고리듬을 통해 수치적 (numerical)으

로 구한다. 이는 SAS, R 등 통계분석 프로그램에 구현되어 있다 (SAS의 GENMOD procedure 및 R의 glm함수). 로지스틱회귀모형으로부터 구한 모수 추정량의 특성 및 모수의 해석은 표준적인 통계교과서에 잘 기술되어 있다 (Montgomery et al., 2006).

전통적 통계기법들 가운데 통계학적, 방법론적 문제점들을 덜 내포하고 있다고 평가되는 것은 로지스틱 회귀분석이라고 할 수 있다 (김종덕, 1999).

로지스틱 회귀분석은 판별분석과 그 분석 목적이 유사하나, 판별분석은 독립변수의 형태가 등간척도 이상의 연속형 변수이어야 다. 그러나 로지스틱회귀분석은 명목척도 또는 서열척도와 연속형 변수도 포함되고, 변수들이 다변량 정규분포를 한다는 가정이 힘들 때 사용하는 기법이다 (김충련, 1993).

로지스틱 회귀분석은 종속변수와 독립변수들의 관계를 비선형적으로 표현하기 때문에 비선형적인 관계를 보이고 있는 현상들의 예측이나 분석에 적합한 분석방법이다.

2.5 의사결정나무

의사결정나무는 의사결정규칙 (decision rule)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류 (classification)하거나 예측 (prediction)을 수행하는 분석방법이다. 분석과정이 나무구조에 의해서 표현되기 때문에 판별분석 (discriminant analysis), 회귀분석 (regression analysis), 신경망 분석 (neural networks analysis) 등과 같은 방법들에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다. 의사결정나무는 분류 또는 예측을 목적으로 하는 어떤 경우에도 사용 될 수 있으나 분석의 정확도보다는 분석과정의 설명이 필요한 경우에 더 유용하게 사용된다 (최종후 등, 1998).

Table 3 의사결정 나무 응용분야

세분화	예측개체를 비슷한 특성을 갖는 몇 개의 그룹으로 분할하여 각 그룹별 특성을 발견하고자 하는 경우
분류	여러 예측변수(predicated variable)에 근거하여 목표변수(target variable)의 범주를 몇 개의 등으로 분류하고자 하는 경우
예측	자료로부터 규칙을 찾아내고 이를 이용하여 미래의 사건을 예측하고자 하는 경우
차원축소 변수선택	매우 많은 수의 예측변수에서 목표변수에 큰 향을 미치는 변수들을 골라내고자 하는 경우
교호작용효과의 파악	여러 개의 예측변수들이 결합하여 목표변수에 작용하는 교호작용을 악하고자 하는 경우
범주의 병합 및 연속형 변수의 이산화	범주형 목표변수의 범주를 소수의 몇 개로 병합하거나, 연속형 목표변수를 몇 개의 등으로 범주화 하고자 하는 경우

다시 말해서 유용한 입력변수를 찾아내고 입력변수간의 다양한 교호작용 즉, 두개 이상의 변수가 결합하여 목표변수에 어떻게 영향을 주는지를 찾아내는 알고리즘이다. 또한 선형성 (linearity)이나 정규성 (normality) 또는 등분산성 (equal variance)등의 가정을 필요로 하지 않는 비모수적 방법이 다. 다음의 Table 4는 의사결정나무의 구성요소를 설명하고 있다 (최종후 외, 1999).

Table 4 의사결정나무의 구성요소

마디	내용
뿌리마디(Root node)	나무구조가 시작되는 마디로서 전체자료로 이루어져 있음
자식마디(Child node)	하나의 마디로부터 분리되어 나간 2개 이상의 마디
부모마디(Parent node)	자식마디의 상위마디
끝마디(Terminal node)	각 나무줄기의 끝에 위치한 마디로 잎이라고도 하며 결국 끝마디의 개수만큼 분류규칙이 생성되는 것
중간마디 (Internal node)	나무구조의 중간에 있는 끝마디가 아닌 마디
가지(Branch)	하나의 마디로부터 끝마디까지 연결된 일련의 마디

일반으로 의사결정나무 분석은 Table 5와 같은 단계를 거친다 (Berry and Linoff:1997; 강철, 서두성, 최종후:1998)

Table 5 의사결정나무 분석 단계

순서	내용
의사결정나무의 형성	분석의 목적과 자료구조에 따라서 한 분리 기준과 정지규칙을 지정하여 의사결정나무를 얻는다.
가지치기	분류오류를 크게 할 위험이 높거나 부적절한 규칙을 가지고 있는 가지를 제거한다.
타당성 평가	이익도표나 위험도표는 검정용 자료에 의한 교차타당성 등을 이용 하여 의사결정나무를 평가한다.
해석 및 예측	의사결정나무를 해석하고 분류 및 예측모형을 설정한다.

이상과 같은 과정에서 정지기, 분리기, 평가기 등을 어떻게 지정 하느냐에 따라서 서로 다른 의사결정나무가 형성된다 (최중후, 1999).

3. 연구방법

3.1 자료 수집 절차

이번 논문의 연구가설은 아래의 Fig. 1과 같다.

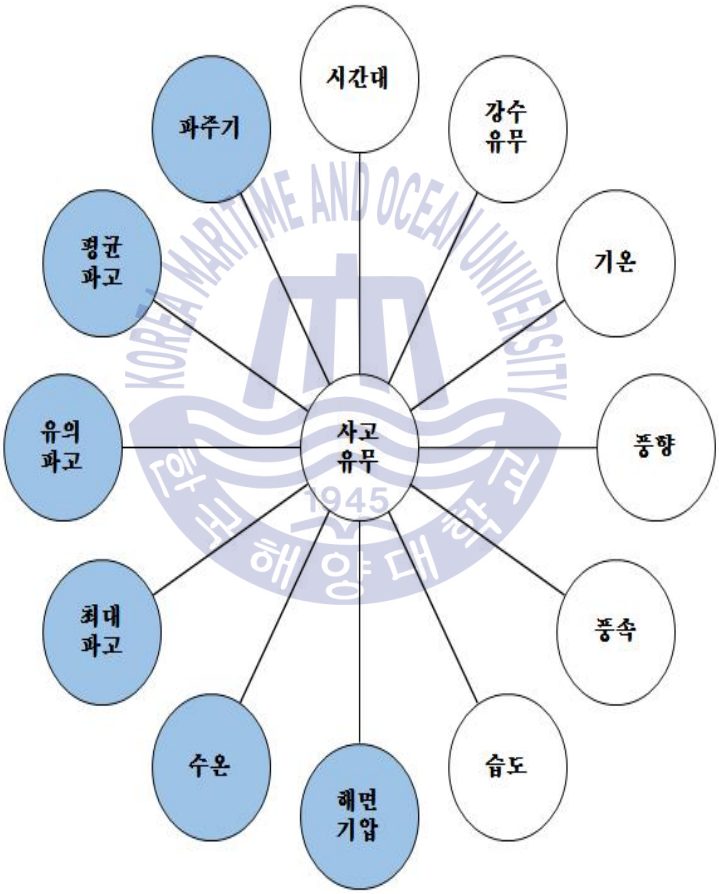

Fig. 1 연구 가설

‘자연적요인 (해양적 요인, 일반 기상적 요인)이 선박사고에 영향이 있다’ 라는 연구 가설을 세우고 분석을 진행하였다.

분석에 사용한 데이터는 기상청과 해양안전심판원의 공공데이터를 이용했고, 종류는 총 3가지로 선박사고, 기상, 해양 데이터를 수집 하였다 (해양사고현황, 2015; 한국해양수산개발원, 2015).

자료 수집 절차는 아래 Fig. 2와 같다.

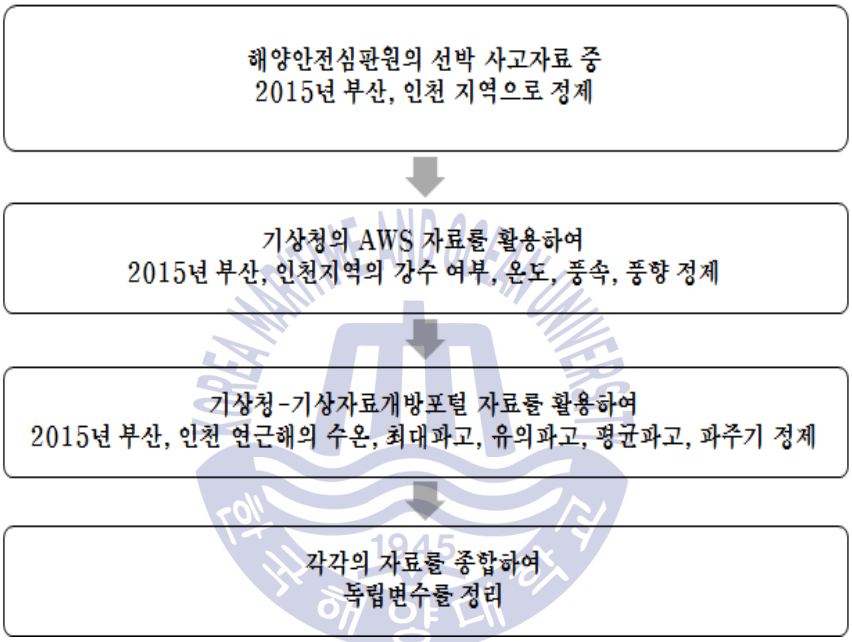

Fig. 2 자료 수집 절차

첫 번째로 선박사고 데이터는 선박의 사고정보에 대한 데이터로, 사고 위치를 기반으로 선박종류, 사고내용 등을 알 수 있는 데이터이다. 해양 안전심판원[1]의 선박사고 중 2015년 부산, 인천 연근해의 선박사고 데이터를 정제 하였다.

두 번째로 기상 데이터는 기상청[2]의 방재기상 (AWS: Automatic Weather

1) https://www.kmst.go.kr/inform/sagoInformationGMTView.jsp
2) http://www.kma.go.kr/weather/observation/aws_table_popup.jsp

System) 자료를 활용하여 기온, 강수, 풍향 등 항해 중 영향을 미칠 수 있는 기상에 관한 데이터이다. 2015년 부산, 인천 연근해의 강수여부, 온도, 풍속, 풍향의 데이터를 정제 하였다.

세 번째로 해양 데이터는 기상청에서 기상자료개방포털[3] 데이터를 활용했으며, 해양 기상부이를 이용해 해양의 파고, 수온 등을 관측한 데이터이다. 2015년 부산, 인천 연근해의 수온, 최대파고, 유의파고, 평균파고, 파주기 데이터를 정제 하였다.

네 번째로 정리 데이터는 위의 해양사고, 기상, 해양 데이터를 종합한 데이터로서, 각 데이터에서 독립변수를 선정하여 새롭게 정제한 데이터를 만들었다.

마지막으로 상향점수매칭 데이터는 상향점수매칭 방법을 이용하여 8,760개의 데이터에 대해서 파악이 가능한 각각에 대해서 부산 연근해 208개, 인천 연근해 80개로 정제하여 분석 하였다.

3) https://data.kma.go.kr/data/sea/selectBuoyList.do?pgmNo=50

3.2 자료 소개

해양사고 데이터는 해양안전심판원에서 제공하는 우리나라 해상에서 일어난 해양사고정보에 대한 데이터로, 사고 위치를 기반으로 선박종류, 사고내용 등을 나타낸 데이터이다. 해양사고의 기간은 1996년부터 2015년이며, 앞의 기간 동안 데이터를 선정하여 총 19,049개의 데이터 중 기상 데이터와 해양 데이터를 구할 수 있는 2015년도 부산, 인천 연근해에서 발생한 선박사고 데이터 129, 108개를 각각 활용했다 (한국해양수산개발원, 2015; 해양안전심판원, 2015).

기상 데이터는 기상청의 AWS자료에서 전국 해안에서 관측되는 데이터를 이용 했다. 강수, 기온, 풍향, 풍속, 습도, 해면기압 등을 측정한 데이터이며, 조사 기간은 2015년 1월 1일 00:00부터 2015년 12월 31일 23:00까지 1시간 마다 측정한 부산, 인천 지역 데이터를 이용했다.

데이터의 개수는 총 365일 × 24시간 = 8,760개 이며, 강수, 기온, 풍향, 풍속, 습도, 해면 기압 등 13개의 변수를 가진 데이터를 2015년 부산, 인천의 강수여부, 온도, 풍속, 풍향으로 데이터를 정제하여 사용했다.

해양 데이터는 기상청에서 기상자료개방포털 자료를 활용했으며, 해양 기상부이를 이용해 해양의 파고, 수온 등을 관측한 데이터이다. 수온, 최대파고, 유의파고, 평균파고, 파주기를 측정한 데이터 이며, 데이터의 기간은 2015년 1월 1일 00:00부터 2015년 12월 31일 23:00까지 1시간 마다 측정한 부산, 인천 지역 데이터를 이용했다.

데이터의 개수는 총 365일 × 24시간 = 8,760개 이며, 수온, 최대파고, 유의파고, 평균파고, 파주기 5개의 변수를 가진 데이터를 이용해 2015년 부산, 인천 연근해의 해양 데이터로 정제해서 사용했다.

해양사고데이터, 기상데이터, 해양데이터를 하나의 데이터로 종합을 하여 하나의 정리 데이터로 만들어 분석에 이용했다. 위의 해양사고, 기상, 해양 데이터를 하나의 데이터로 종합하여 정제한 이유는 만약 해양사고 데이터의 독립변수로만 봤을 때, A 데이터가 매우 유의하다고 할 수 있지만, 기상 데이터의 독

립변수로만 봤을 때, A 데이터가 유의하지 않다고 할 수 있으므로 모든 상황을 한 번에 확인 할 수 있도록 해양사고, 기상, 해양데이터를 종합하여 정제를 했다.

이 자료에 대해서 2015년 1월 1일 00:00 ~ 2015년 12월 31일 23:00 까지 시간을 나누고 사고 유무에 따라서 자료를 분류하였다. 사고 자료는 종속변수의 수가 적기 때문에 분석을 위해서 상향점수매칭을 활용하여 한 번 더 정제하였다. 상향점수매칭을 통한 데이터는 아래 Table 6과 같이 독립변수를 사용하였으며, 부산 208개, 인천 80개로 최종적으로 자료를 정리하여 분석에 활용 하였다.

Table 6 변수 설명

변수	설명
시간	2015년 1월 1일 00:00부터 ~ 2015년 12월 31일 23:00
시간대	1(0,1,2시) 2(3,4,5시) 3(6,7,8시) 4(9,10,11시) 5(12,13,14시) 6(15,16,17시) 7(18,19,20시) 8(21,22,23시)
강수유무	No = 0 , Yes = 1
강수	강수15 : 자료시간에서 과거 15분간 내린 강수의 양 강수60 : 자료시간에서 과거 60분간 내린 강수의 양 강수6H : 자료시간에서 과거 6시간 내린 강수의 양 강수12H : 자료시간에서 과거 12시간 내린 강수의 양 일강수 : 오늘 00시 00분부터 자료시간까지 내린 강수의 양
기온	연속형(continuous) 변수
풍향	풍향1 : 1분 풍향(degree, 16방위) 풍향10 : 10분 평균 풍향(degree, 16방위) E, ENE, ESE, N, NE, NNE, NNW, NW, S, SE, SSE, SSW, SW
풍속	풍속1 : 1분 평균 풍속(m/s) 풍속10 : 10분 평균 풍속(m/s)
습도	연속형(continuous) 변수
해면기압	연속형(continuous) 변수
수온	연속형(continuous) 변수
최대파고	파도의 최고 높이 / 연속형(continuous) 변수
유의파고	파도 높이 상위 30%에 대한 평균 / 연속형(continuous) 변수
평균파고	파도의 평균 높이 / 연속형(continuous) 변수
파주기	파도의 주기 / 연속형(continuous) 변수

부산	208	인천	80

4. 통계분석 결과

4.1 로지스틱 회귀분석 결과

상향점수매칭을 통하여 정제한 후에 독립변수를 사용하여, 사고유무에 대해서 분석 하였다.

Table 7 부산 연근해의 로지스틱 회귀분석 결과

	B	S.E,	Wald	자유도	유의확률	Exp(B)
일 강수	-0.232	0.104	4.962	1	0.026	0.793
풍속1	0.270	0.107	6.344	1	0.012	1.309
습도	0.030	0.017	3.086	1	0.079	1.030
해면기압	-0.031	0.022	2.021	1	0.155	0.969
유의파고	1.491	0.637	5.476	1	0.019	4.442
상수항	27.831	22.665	1.508	1	0.219	1.221E12

위의 표 에서 B의 부호가 (+) 는 일 강수, 풍속1, 습도, 유의파고로 나타내며 해양선박사고가 난 집단으로 분류되고 해면기압은 (-) 부호를 나타내며 해양선박사고가 나지 않는 집단으로 분류 된다. 분류집단 예측력이 있기 위해서는 변수의 유의성을 확인해야 한다. 확인해 본 결과 습도, 해면기압을 제외한 일 강수, 풍속, 유의파고 유의수준 0.05보다 작으므로 대립가설이 채택됨을 알 수 있다. 즉, 선박사고가 습도, 해면기압을 제외한 일 강수, 풍속, 유의파고 영향을 받는다고 할 수 있다.

이를 통하여 소속 집단을 예측하면 해양선박사고가 날 확률을 P로 두고 로지스틱 회귀모형을 만들면 다음과 같다.

$$\frac{P}{1-P} = \exp(27.831 - 0.232 \times \text{일강수} + 0.270 \times \text{풍속1} + 0.030 \times \text{습도} - 0.031 \times \text{해면기압} + 1.491 \times \text{유의파고}). \quad (4.1.1)$$

일 강수가 1 증가하면 선박사고는 exp(-0.232)배가 증가 하고, 풍속1이 1 증가하면 선박사고는 exp(0.270)배 증가, 유의파고가 1 증가하면 exp(1.491)배 증가함을 알 수 있다.

Table 8 부산 연근해의 로지스틱 회귀분석 분류정확도

		예측		
		사고 X	사고 O	분류정확 %
실제	사고 X	40	64	38.5
	사고 O	8	96	92.3
전체 %				65.3

사고가 나지 않은 집단 104 case 중 40 case, 그리고 사고가 발생한 집단에 속한 104 case 중 96 case가 제대로 분류 되었으며, 분류정확도는 65.3%로 나타났다 (이학식, 2013; 김채희, 2015).

Table 9 인천지역의 로지스틱 회귀분석 결과

	B	S.E,	Wald	자유도	유의확률	Exp(B)
일 강수	-.876	.295	8.803	1	.003	.416
풍속1	.316	.343	.846	1	.358	1.372
습도	-.616	1.918	.103	1	.748	.540
해면기압	.195	.094	4.315	1	.038	1.215
유의파고	-5.638	3.196	3.111	1	.078	.004
기온	.454	.173	6.904	1	.009	1.575
상수항	-139.497	255.481	0.298	1	0.585	0.000

위의 표 에서 B의 부호가 (+) 는 풍속1, 해면기압, 기온으로 나타내며 선박사고가 난 집단으로 분류되고 해면기압은 (-) 부호를 나타내며 선박사고가 나지 않는 집단으로 분류 된다. 분류집단 예측력이 있기 위해서는 변수의 유의성을 확인해야 한다. 확인해 본 결과 일 강수, 해면기압, 기온이 유의수준 0.05보다 작으므로 대립가설이 채택됨을 알 수 있다. 즉, 선박사고는 일 강수, 해면기압, 기온에 영향을 받는다고 할 수 있다.

이를 통하여 소속 집단을 예측하면 해양선박사고가 날 확률을 P로 두고 로지스틱 회귀모형을 만들면 다음과 같다.

$$\frac{P}{1-P} = \exp(27.831 - 0.876 \times \text{일강수} + 0.316 \times \text{풍속1} - 0.616 \times \text{습도} + 0.195 \times \text{해면기압} - 5.638 \times \text{유의파고} + 0.454 \times \text{기온}). \quad (4.1.2)$$

일 강수가 1 증가하면 선박사고는 exp(-0.876)배가 증가 하고, 풍속1이 1 증가하면 exp(0.316)배 증가, 해면기압이 1 증가하면 선박사고는 exp(0.195)배 증가 하는 것을 알 수 있다.

Table 10 인천 연근해의 로지스틱 회귀분석 분류정확도

		예측		
		사고 X	사고 O	분류정확 %
실제	사고 X	16	24	40
	사고 O	0	40	100
전체 %				70

사고가 나지 않은 집단 40 case 중 16 ase, 그리고 사고가 발생한 집단에 속한 40 case 중 40 case가 제대로 분류 되었으며, 분류정확도는 70%로 나타났다 (이학식, 2013; 김채희, 2015).

4.2 의사결정나무분석 결과

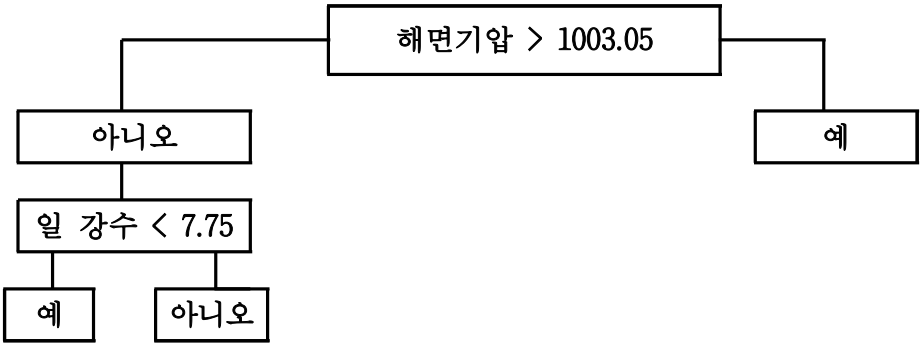

Fig. 3 부산 연근해의 의사결정나무 분석결과

부산 연근해에서 선박사고가 발생하는 경우를 살펴보면, 해면 기압이 큰 경우에 발생하며, 해면기압이 1003.05 보다 작은 경우에 일 강수량이 7.75 보다 작은 경우에 사고가 발생한다는 것을 알 수 있다.

Table 11 부산 연근해의 의사결정나무 분류정확도

		예측		
		사고 X	사고 O	분류정확 %
실제	사고 X	104	0	100
	사고 O	8	96	92.3
전체 %				96.2

사고가 나지 않은 집단 104 case 중 104 case, 그리고 사고가 발생한 집단에 속한 104 case 중 96 case가 제대로 분류 되었으며, 분류정확도는 96.2%로 나타났다 (이학식, 2013; 김채희, 2015).

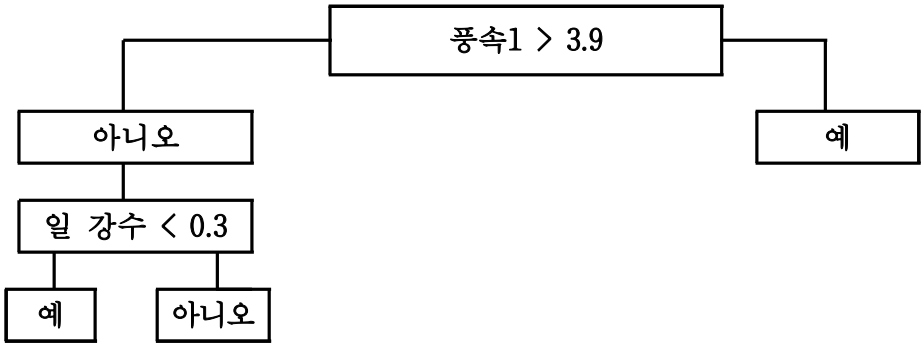

Fig. 4 인천 연근해의 의사결정나무 분석결과

인천 연근해에서 풍속이 3.9 이상인 경우 선박사고가 발생하며, 그렇지 않은 경우에는 일 강수가 0.3보다 작을 때 발생하는 것을 알 수 있다.

Table 12 인천 연근해의 의사결정나무 분류정확도

		예측		
		사고 X	사고 O	분류정확 %
실제	사고 X	40	0	100
	사고 O	0	40	100
전체 %				100

사고가 나지 않은 집단 40 case중 40 case, 그리고 사고가 발생한 집단에 속한 40 case중 40 case가 제대로 분류 되었으며, 분류정확도는 100%로 나타났다 (이학식, 2013; 김채희, 2015).

5. 고찰 및 결론

부산, 인천 연근해의 선박사고가 자연적 요인에 영향을 받는지에 관해 분석을 했다. 데이터는 기상청과 해양안전심판원의 선박사고, 기상, 해양 데이터를 이용하였으며, 총 8760개의 데이터로 정제를 했다. 그 후, 상향점수매칭 방법을 이용하여 부산 연근해 208개, 인천 연근해 80개의 데이터로 줄여 각각의 분석을 진행하였다.

R프로그램을 이용하여 로지스틱 회귀분석을 실시 전에 일, 월, 시, 시간대, 사고유무, 강수량, 일 강수, 기온, 풍향, 풍속, 습도, 해면기압, 수온, 최대파고, 유의파고, 평균파고, 파주기 등 독립변수 중 다중공선성 (multicollinearity)이 존재하는 변수를 분산팽창인수 (variance inflation factor, VIF)을 확인 후 제거 한 후의 지역별로 독립변수를 이용하여 로지스틱 회귀분석을 실시했다.

상향점수매칭 데이터로 정제하기 전 정리 데이터를 이용하여 분석을 실시했을 때, 각 독립변수에 대해서 해양선박사고유무에 영향을 끼치는 유의한 독립변수를 없었다. 반면, 상향점수매칭데이터로 정제 후 로지스틱 회귀분석을 실시했을 때, 부산 연근해에서는 습도, 해면기압을 제외한 일 강수, 풍속1, 유의파고가 해양선박사고유무에 영향을 끼치는 것을 알 수 있고 인천 연근해에서는 풍속1, 풍향10, 습도를 제외한 일 강수, 기온, 파주기가 해양선박사고유무에 영향을 끼치는 것을 확인 할 수 있었다.

결론적으로 로지스틱 회귀모형으로는 다음과 같은 결과를 알 수 있다.

부산 연근해에서는 유의파고가 높아지면 해양선박사고는 exp(1.491)=4.442배 더 발생한다. 풍속1이 강해지면 해양선박사고는 exp(0.270)=1.309배 더 발생한다. 일 강수량이 줄어들면 해양선박사고는 exp(-0.232)=0.793배 줄어든다.

인천 연근해에서는 일 강수량이 줄어들면 선박사고는 exp(-0.876)=0.416배 줄어든다. 풍속1이 강해지면 선박사고는 exp(0.316)=1.372배 더 발생한다. 기온이 높아지면 선박사고는 exp(0.195)=1.215배 더 발생한다.

부산, 인천 연근해의 선박사고에 대해서 로지스틱 회귀분석을 통하여 알아본 결과 지역에 따라서 다른 자연적 요인이 영향을 준다는 것을 알 수 있었고 로지스틱 회귀모형은 사고가 발생하지 않았을 경우보다 사고가 발생할 경우를 더 잘 예측 하였다.

의사결정 나무에서는 부산 연근해는 해면기압과, 일 강수, 인천 연근해에서는 풍속1과 일 강수가 영향을 주는 것을 알 수가 있었다. 그 결과 선박사고에 원인이 지역별로 상이한 점이 있다는 것을 확인 할 수 있었다.

각 모형의 분류정확도는 아래 Table 13과 같이 나타났다.

Table 13 분석방법별 분류 정확도

로지스틱 회귀분석 분류정확도		의사결정 나무 분류정확도	
부산	65.3	부산	96.2
인천	70	인천	100

2015년도 부산, 인천 연근해에 선박사고가 발생할 확률에 대해서는 의사결정 나무가 분석 정확도가 상대적으로 더 높은 분류정확도가 나온 것을 확인 할 수 있었다. 이번 논문을 통하여 부산, 인천 연근해에 발생하는 선박사고의 자연적 요인에 차이가 있다는 것을 알 수 있다.

국·내외로 해양선박사고는 많은 인적 요인과 자연적 요인으로 일어나고 있다. 지역별로 차이가 발생하는 자연적 요인들에 대하여 개별적으로 분석을 진행하여야 한다. 지역별 특색에 맞는 사고예방 모델을 개발하여 해양선박사고를 줄이는 것이 앞으로 남은 과제라고 생각한다.

앞으로의 연구에서는 Fig.5는 개발 완료 후의 프로그램에 대해서 나타낸 것이다.

부두 인근 해상 지역을 그리드로 나누고 각 그리드 지역에 사고 확률을 각각

의 분석을 통하여 계산해서 사고 확률에 따른 위험 등급을 지도상에 나타내는 것이다. 지리정보시스템 (geographic information system, GIS)의 공간분석기능과 공간통계를 이용하여 과거 데이터를 통해서 선박 사고 장소에 대해서 분석을 하려고 한다. 기존 사고 지역에 대해서 버퍼분석 (buffer analysis)을 통하여 타일화 시키고 GIS를 활용하여 공간 보간을 위해 샘플점에 대해 사고지역에 크리깅 보간법 (Kriging Interpolation)을 적용하고, 크리깅은 선형, 원형, 구형, 지수형 그리고 가우스형을 선정하여 활용 할 것이다. 사고지역에 관측점간의 거리에 반비례하여 가중치를 할당하는 방법이며, 스플라인은 완화곡선에 적용하여 구간별로 별도의 다항식을 채택하여 적용하는 방법이다. 또한 크리깅 보간법은 기하학적 거리에 대한 상관관계가 아니라 통계학적인 상관성을 이용하는 것으로서 반 베리오그램(Semivariogram)의 형태에 따라 다양한 형태를 가지게 된다.

선박사고에 대한 장기간 축적된 데이터로부터 과거와 현재의 규칙성과 상관관계를 밝히고, 이를 토대로 선박사고에 대한 발생 여부를 예측하고 정성적 정보의 양과 종류가 과거보다 대폭 증가하여 이상 징후 감지, 고위험 지역 발생 경고 등 포괄적인 리스크 관리가 가능할 것이다.

Fig. 5 개발 완료 후의 프로그램의 모습

참고문헌

강철, 서두성, 최종후, 1998. *Enterprise Minier의 의사결정나무분석 알고리즘*, SAS 사용자 컨퍼런스 발표자료집. 서울:SAS-Korea.

곽수용, 2011. *국내 해양선박사고의 원인요인 선정 및 정량적 위험도 분석*. 석사학위논문. 부산대학교.

김재희, 2015. *R을 이용한 회귀분석*. 파주:자유아카데미.

김종덕, 1999. *회귀분석*. 서울:세종출판사.

서용화, 2006. *선박 안전을 위한 해양 사고 사례 분석*. 석사학위논문. 부산대학교.

이학식, 2013. *SPSS 20.0 매뉴얼*. 서울:집현재.

최종후, 한상태, 강철, 김은석, 1998. *AnswerTree를 이용한 데이터마이닝 의사결정나무분석*. 서울 : SPSS아카데미

최종후 등, 1999. *데이터마이닝 의사결정나무의 응용*. 대전:통계청

한국해양수산개발원, 2015. *해운통계요람*. 부산:한국해양수산개발원.

해양안전심판원, 2015. *해양사고현황*. 세종:해양안전심판원.

Berry, M. J. A., Linoff, G. S., 1997. *Data Mining Techniques*. New York:John Wiley & Sons, Inc.

Chatterjee, Hadi, 2006. *Regression Analysis by Example*. Fourth Edition. Wiley.

Dehejia, Rajeev H, and Sadek Wahba, (1999). *Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs*. Journal of the American Statistical Association. 94(448): 1053-1062.

Montgomery, D. C., Peck, E. A., Vining, G. G., 2006. *Introduction to linear regression analysis*. the 4th Edition. Wiley.

Rosenbaum, P. R., Rubin, D. B., 1983. *The central role of the propensity score in*

observational studies for causal effects. Biometrika, 70(1): 41-55.

Zhao, Zhong, 2000. *Using Matching to Estimate Treatment Effects: Data Requirements, Sensitivity, and An Application. Unpublished Working Paper*. Baltimore: The Johns Hopkins University

부록

```
/* ************************************** */
/* ************************************** */
/* Matching Macro */
/* ************************************** */
/* ************************************** */
%MACRO OneToManyMTCH (
Lib, /* Library Name */
Dataset, /* Data set of all patients */
depend, /* Dependent variable that indicates Case or Control */
/* Code 1 for Cases, 0 for Controls */
SiteN, /* Site/Hospital ID */
PatientN, /* Patient ID */
matches, /* Output data set of matched pairs */
NoContrls); /* Number of controls to match to each case */
/* ********************** */
/* Macro to Create the Case and Control Data sets */
/* ********************** */
%MACRO INITCC(CaseAndCtrls,digits);
data tcases (drop=cprob)
tctrl (drop=aprob) ;
set &CaseAndCtrls. ;
/* Create the data set of Controls */
if &depend. = 0 and prob ne . then
do;
cprob = Round(prob,&digits.);
Cmatch = 0;
Length RandNum 8;
```

```
RandNum=ranuni(1234567);
Label RandNum='Uniform Randomization Score';
output tctrl;
end;
/* Create the data set of Cases */
else if &depend. = 1 and prob ne . then
do;
Cmatch = 0;
aprob =Round(prob,&digits.);
output tcases;
end;
run;
%SORTCC;
%MEND INITCC;
/* ********************* */
/* Macro to sort the Cases and Controls data set */
/* ********************* */
%MACRO SORTCC;
proc sort data=tcases out=&LIB..Scase;
by prob;
run;
proc sort data=tctrl out=&LIB..Scontrol;
by prob randnum;
run;
%MEND SORTCC;


/* ********************* */
/* Macro to Perform the Match */
/* ********************* */
%MACRO MATCH (MATCHED,DIGITS);
data &lib..&matched. (drop=Cmatch randnum aprob cprob start oldi curctrl
```

```
matched);
/* select the cases data set */
set &lib..SCase ;
curob + 1;

matchto = curob;
if curob = 1 then do;
start = 1;
oldi = 1;
end;
/* select the controls data set */
DO i = start to n;
set &lib..Scontrol point = i nobs = n;
if i gt n then goto startovr;
if _Error_ = 1 then abort;
curctrl = i;
/* output control if match found */
if aprob = cprob then
do;
Cmatch = 1;
output &lib..&matched.;
matched = curctrl;
goto found;
end;
/* exit do loop if out of potential matches */
else if cprob gt aprob then
goto nextcase;
startovr: if i gt n then
goto nextcase;
END; /* end of DO LOOP */
/* If no match was found, put pointer back*/
nextcase:
```

```
if Cmatch=0 then start = oldi;


/* If a match was found, output case and increment pointer */
found:
if Cmatch = 1 then do;
oldi = matched + 1;
start = matched + 1;
set &lib..SCase point = curob;
output &lib..&matched.;
end;
retain oldi start;
if _Error_=1 then _Error_=0;
run;
/* get files of unmatched cases and controls */
proc sort data=&lib..scase out=sumcase;
by &SiteN. &PatientN.;
run;
proc sort data=&lib..scontrol out=sumcontrol;
by &SiteN. &PatientN.;
run;
proc sort data=&lib..&matched. out=smatched (keep=&SiteN. &PatientN.
matchto);
by &SiteN. &PatientN.;
run;
data tcases (drop=matchto);
merge sumcase(in=a) smatched;
by &SiteN. &PatientN.;
if a and matchto = . ;
cmatch = 0;
aprob =Round(prob,&digits.);
run;
```

```
data tctrl (drop=matchto);
merge sumcontrol(in=a) smatched;
by &SiteN. &PatientN.;
if a and matchto = . ;
cmatch = 0;
cprob = Round(prob,&digits.);
run;
%SORTCC
%MEND MATCH;

/* ********************* */
/* Macro to call Macro MATCH for each of the 8-digit to 1-digit matchs */
/* ********************* */
%MACRO CallMATCH;
/* Do a 8-digit match */
%MATCH(Match8,.0000001);
/* Do a 7-digit match on remaining unmatched*/
%MATCH(Match7,.000001);
/* Do a 6-digit match on remaining unmatched*/
%MATCH(Match6,.00001);
/* Do a 5-digit match on remaining unmatched*/
%MATCH(Match5,.0001);
/* Do a 4-digit match on remaining unmatched */
%MATCH(Match4,.001);
/* Do a 3-digit match on remaining unmatched */
%MATCH(Match3,.01);
/* Do a 2-digit match on remaining unmatched */
%MATCH(Match2,.1);
/* Do a 1-digit match on remaining unmatched */
%MATCH(Match1,.1);
%MEND CallMATCH;
/* ********************* */
```

```
/* Macro to Merge all the matches files into one file */
/* ********************* */
%MACRO MergeFiles(MatchNo);
data &matches.&MatchNo. (drop = matchto);
set &lib..match8(in=a) &lib..match7(in=b) &lib..match6(in=c) &lib..match5(in=d)
&lib..match4(in=e)
&lib..match3(in=f) &lib..match2(in=g) &lib..match1(in=h);
if a then match_&MatchNo. = matchto;
if b then match_&MatchNo. = matchto + 10000;
if c then match_&MatchNo. = matchto + 100000;
if d then match_&MatchNo. = matchto + 1000000;
if e then match_&MatchNo. = matchto + 10000000;
if f then match_&MatchNo. = matchto + 100000000;
if g then match_&MatchNo. = matchto + 1000000000;
if h then match_&MatchNo. = matchto + 10000000000;
run;
%MEND MergeFiles;

/* ******************************** */
/* ******************************** */
/* Perform the initial 1:1 Match */
/* ******************************** */
/* ******************************** */
/* Create file of cases and controls */
%INITCC(&LIB..&dataset.,.00000001);
/* Perform the 8-digit to 1-digit matches */
%CallMATCH;
/* Merge all the matches files into one file */
%MergeFiles(1)


/* ******************************** */
```

```
/* ******************************** */
/* Perform the remaining 1:N Matches */
/* ******************************** */
/* ******************************** */
%IF &NoContrls. gt 1 %Then %DO;
%DO i = 2 %TO &NoContrls.;
%let Lasti=%eval(&i. - 1);
/* ********** */
/* Start with Cases from the last Matched Cases file and the remaining Un-Matched */
/* Controls. NOTE: The Unmatched Controls file (Scontrol) is created at end of the */
/* previous match */
/* Select the Matched Cases from the last Matched File */
data &LIB..Scase;
set &matches.&Lasti.;
where &Depend. = 1;
run;
/* ********** */
/* Perform the 8-1 digit matches between Matched Cases and the Unmatched Controls */
%CallMATCH;
/* ********** */
/* Merge the 8-digit to 1-digit matches files into one file */
%MergeFiles(&i.)
%DO m = 1 %TO &Lasti.;
data &matches.&i.;
set &matches.&i.;
if &Depend.=0 then Match_&m. = .;
run;
%END;
```

```
/* ********** */
/* Determine which OLD Controls correspond to the kept Cases */
%DO c = 1 %TO &Lasti.;
/* Select the KEPT Cases */
proc sort data=&matches.&i. out=skeepcases (keep = Match_&c.);
by Match_&c.;
where &Depend. = 1;
run;
/* Get the OLD Controls */
proc sort data = &matches.&Lasti. out = soldcontrols&c.;
by Match_&c.;
where &Depend. = 0 and Match_&c. ne . ;
run;
/* Get the OLD Controls that correspond to the kept Cases */
data keepcontrols&c.;
merge skeepcases (in = a) soldcontrols&c. (in = b);
by Match_&c.;
if a;
run;
%END;
/* ********** */
/* Combine all the OLD Controls into one file */
data keepcontrols;
set keepcontrols1 (obs=0);
run;
%DO k = 1 %TO &Lasti.;
data keepcontrols;
set keepcontrols keepcontrols&k.;


run;
%END;
```

```
/* ********** */
/* Append the OLD matched Controls to the new file of matched cases and
controls */
data &matches.&i.;
set &matches.&i. keepcontrols;
run;
/* ********** */
/* If there are more matches to be made, add the previously matched, but
not kept, */
/* controls back into the pool of unmatched controls */
%if &i. lt &NoContrls. %then %do;
%DO z = 1 %TO &Lasti.;
/* Select all the KEPT Cases */
proc sort data=&matches.&i. out=skeepcases (keep = Match_&z.);
by Match_&z.;
where &Depend. = 1;
run;
/* Select all the OLD Controls */
proc sort data = &matches.&Lasti. out = soldcontrols&z.;
by Match_&z.;
where &Depend. = 0 and Match_&z. ne .;
run;
/* Keep the OLD Controls that correspond to the NOT KEPT Cases */
/* Drop the previuos Match_X variable */
data AddBackControls&z. (drop = Match_&z.);
merge skeepcases (in = a) soldcontrols&z. (in = b);
by Match_&z.;
if b and not a;
run;
%END; /* End DO */

/* Drop the previuos Match_X variable */
```

```
data &LIB..Scontrol (drop = Match_&lasti. );
set &LIB..Scontrol;
run;
/* Append */
%DO y = 1 %TO &Lasti.;
data &LIB..Scontrol;
set &LIB..Scontrol AddBackControls&y.;
run;
%END; /* End DO */
%end; /* End IF */
%END; /* End Main DO */
%END; /* End Main IF */
/* ************************************ */
/* ************************************ */
/* Save the final matched pairs data set */
/* ************************************ */
/* ************************************ */
/* Sort file by Treatment Variable */
proc sort data=&matches.&NoContrls. out = &lib..&matches.;
by &depend.;
run;
%MEND OneToManyMTCH;
proc logistic  data=accident  desc;
model accident=day_week time_zone;
output out=accident_propen pred=prob;
run;
data accident_propensity;
set accident_propen;
hid=1;
run;

%OneToManyMTCH(work,accident_propensity,accident,hid,no,Matches_pre,1);
```

감사의 글

졸업논문을 마무리하며 지난 2년이란 시간을 돌이켜보게 되었습니다. 후회가 되고 아쉬운 일도 많았지만, 스스로 성장할 수 있는 시간이었습니다. 많은 분들의 격려와 도움으로 대학원 생활을 무사히 마칠 수 있었습니다.

먼저 연구자로서의 자세와 어떤 연구를 하여야 하는지에 대해서 지도해주신 지도교수인 박찬근 교수님께 존경과 감사의 말씀을 드립니다. 또한 심사위원으로서 부족한 논문에 대해서 아낌없는 충고와 조언을 해주신 김재환 교수님과 장길웅 교수님께 감사드립니다.

학부, 대학원 생활을 거치면서 인생의 선배이자 교수로서 때로는 따뜻하게, 때로는 따끔하게 옳은 길로 갈 수 있게 격려해주신 배재국 교수님, 홍정희 교수님, 김익성 교수님, 손미정 교수님께도 감사드립니다. 또한, 대학원 생활을 순조롭게 이어갈 수 있도록 배려해주신 정지영 조교님께도 감사를 드립니다. 같이 연구실 생활을 하면서 서로 기댈 수 있는 든든한 동기 선배들에게도 고마움을 전합니다.

마지막으로 제가 선택한 길을 믿고 응원해준 가족들에게 정말 감사드립니다.