KIGAS Vol. 16, No. 1, February, 2012
(Journal of the Korean Institute of Gas)

국내 가스사고와 기상자료의 데이터마이닝을 이용한 가스사고 예측모델 연구

허영택* · **신동일**** · **†이수경**
*한국가스안전공사 · **명지대학교 화학공학과 · 서울과학기술대학교 안전공학과
(2012년 2월 3일 투고, 2012년 2월 27일 수정, 2월 27일 채택)

Data Mining of Gas Accident and Meteorological Data in Korea for a Prediction Model of Gas Accidents

Young-taeg Hur* · Dongil Shin · Su-Kyung Lee†**
Korea Gas Safety Corporation, Gyeonggi-do, 429-712, Korea
**Dept. of Chemical Engineering, Myongji University, Yongin, Gyeonggido 449-728, Korea*
Dept. of Safety Engineering, Seoul National University of Technology, Seoul 139-743, Korea
(Received February 3, 2012; Revised February 27, 2012; Accepted February 27, 2012)

요 약

본 연구에서는 국내 가스사고의 발생 환경을 분석하여 가스사고의 재발을 방지하고자 가스 사고를 유형별로 분석하였다. 가스사고는 지속적으로 발생하고 있고, 사고의 내용에서도 시기별, 날씨 등에 따라 가스사용 형태가 변하고 있어서 가스의 사용환경과 가스사고는 밀접한 관계가 있는 것으로 나타났다. 가스사고를 평균기온, 최고기온, 최저기온, 상대습도, 운량, 강수량 및 풍속의 7가지 기상요소별로 분석해 본 결과, 기온과 상대습도 등에 따라 영향을 받고 있은 것으로 나타났으며, 맑은 날, 풍속은 낮을 때 가스사고 발생빈도가 많았다. 가스사고 예측을 위하여 제시된 모델식을 활용하여 기상청의 일기예보 시스템과 연계하여 가스사고 발생 가능성을 실시간으로 제공하고, 회사의 업무시스템과 연계시켜 실시간으로 확인이 가능하도록 하여 가스사고 예방활동에 적극 활용할 수 있을 것으로 사료된다.

Abstract - Analysis on gas accidents by types occurred has been made to prevent the recurrence of accidents, through analysis of past history of gas accident occurring environment. The number of gas accidents has been decreasing, but still accidents are occurring steadily. Gas-using environment and gas accidents are estimated to be closely connected since gas-using types are changing by time period, weather, etc. in terms of accident contents. As a result of analysing gas accidents by 7 meteorological elements, such as the mean temperature, the highest temperature, the lowest temperature, relative humidity, the amount of clouds, precipitation and wind velocity, it has been found out that gas accidents are influenced by temperature or relative humidity, and accident occurs more frequently when the sky is clean and wind velocity is slow. Possibility of gas accidents can be provided in real time, using the proposed model made to predict gas accidents in connection with the weather forecast service. Possibility and number of gas accidents will be checked real time by connecting to the business system of Korea Gas Safety Corp., and it is considered that it would be positively used for preventing gas accidents.

Key words : data mining, accident warning and prediction, meteorological data, gas accident data, statistical analysis.

†교신저자:lsk@seoultech.ac.kr

Ⅰ. 서 론

산업의 발달과 더불어 국내의 가스산업도 양적인 팽창을 거듭해 오고 있으며, 사용량 증가와 함께 가스로 인한 사고도 매년 지속적으로 발생하고 있다. 국내의 경우에는 아직도 이렇다 할 기술적 지원 및 관련 자료의 지원 부족으로 한번 겪었던 가스사고의 경험을 반복해서 겪고 있는 실정이다. 따라서 효율적인 가스사고의 예방과 관련 정책의 수립하기 위해서는 가스사고 발생의 특성을 파악하여 예방하는 노력이 더욱 중요해 지고 있다. 현재 기상청(KMA, Korea Meteorologi- cal Administration)의 경우도 날씨변화에 따른 다양한 예보시스템을 갖추고, 생활과 밀접한 기상정보를 지수화 하여 일반 국민들에게 제공하고 있다. 산불위험지수, 식중독지수, 불쾌지수, 자외선지수, 대기오염기상지수, 황사영향지수 및 보건지수 등의 제공이 그 예이다. 따라서 과거의 가스사고 발생 특성의 분석과 활용은 가스사고 예방의 중요한 기초자료로 사용될 수 있다[1]. 또한 소방방재청(KNEMA, Korea National Emergency Management Agency)에서도 전국 광역자치단체별로 화재, 물놀이, 산불 등 10개 안전사고 유행에 대한 안전지수를 개발하여 주간단위로 예보하고 있다. 이 안전사고 위험지수는 과거에 주간 단위로 발생한 안전사고 중에서 가장 많이 발생한 주간단위의 안전사고를 100으로 기준해서 지수가 90이상이면 위험단계를 의미하는 것으로, 안전사고 지수를 국민의 안전사고 예방에 적극 활용하고 있다[2].

따라서 가스사고 발생 요인을 종합적으로 분석하여 관련 기상조건에 따른 가스사고의 발생 형태를 도출해 내면 기상청이 일기와 관련된 각종 생활정보를 예보하듯 가스사고의 발생 가능성도 예측이 가능할 것으로 판단된다. 또한 확률적 또는 비확률적 방법을 통해 일정기간 동안에 발생 가능한 가스사고 건수의 예측은 불확실한 미래를 대비하는 것이므로 현재의 의사결정에 매우 중요한 영향을 미친다. 따라서 가스사고 건수의 예측은 가스사고 예방대책의 수립에 있어 필수적인 요소이다.

선행연구로서 가스사고 발생 환경분석을 통한 사고발생 모형고찰에서 가스사고 발생에 대한 환경 인자를 분석하였다[3].

본 연구에서는 기상요소에 따른 가스사고의 발생 가능성의 예측을 위한 적절한 모델을 제시하고 실시간 제공 및 가스사고 예방활동 등의 활용방안도 제시하고자 한다.

Ⅱ. 이론적 배경

가스사고의 발생가능성을 예측하기 위하여 사용된 통계적 기법의 이론적 내용을 기술하였다. 기상요소에 따른 가스사고의 예측식을 유도하기 위하여 데이터마이닝 기법의 하나이며, 전통적으로 활용되어온 통계기법이기도 한 인과형 예측기법 중 회귀분석과 상관분석 방법을 적용하였다.

2.1 단순선형회귀분석

회귀분석에서 가장 간단한 모형이 독립변수의 수가 하나이면서 함수형태가 선형인 단순선형회귀모형(simple linear regression model)이다. 종속변수를 y라 하고, 독립변수를 x라고 할 때 단순선형회귀모형은 식 (1)과 같다. 다만, $\epsilon_i \sim N(0, \sigma^2)$이고, 서로 독립이다.

$$y_i = \alpha + \beta x_i + \epsilon_i, i = 1,2,\cdots,n \quad (1)$$

여기서 회귀식은 독립변수 x의 일차식으로 나타내어지며, α와β는 절편과 기울기를 나타내는 회귀계수(regression coefficients)라 하고 앞의 회귀식을 모집단의 회귀모형이라 한다. ϵ_i는 i번째 측정치의 오차(error)이고, 오차항이라 한다. 오차항에 대해서는 정규성(normality), 등분산성(equal variance), 독립성(independence)을 일반적으로 가정하고 있다.

2.2 다중선형회귀분석

회귀분석의 실제 응용에서는 독립변수가 1개인 단순선형회귀분석보다는 독립변수가 2개 이상인 다중선형회귀분석(multiple linear regression analysis)이 보다 빈번하게 사용되고 있다. 그 이유는 독립변수 하나로만 종속변수를 설명하는데 있어 한계가 있기 때문이다. 즉, 종속변수는 보통의 경우 여러 가지 독립변수와 관계를 가지고 있다.

독립변수가 $k(\geq 2)$개 이고, 선형함수식을 갖는 회귀모형이 식 (2)와 같다고 하면,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i, \ i = 1,\cdots,n \quad (2)$$

중회귀모형의 일반적인 모형은 다음과 같은 행렬과 벡터로 표현된다. 관측값 $i(i = 1,2,\cdots,n)$에 대해 각각 식 (3), 식 (4), 및 식 (5)이므로, 식 (6)과 같이 행렬과 벡터로 표현할 수 있다.

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_k x_{k1} + \epsilon_1 \quad (3)$$

$$y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_k x_{k2} + \epsilon_2 \quad (4)$$

$$\vdots \qquad\qquad \vdots$$

$$y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_k x_{kn} + \epsilon_n \quad (5)$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \; Y = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix} \quad (6)$$

$$, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}, \; \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon n \end{pmatrix}$$

2.3 상관분석

상관분석은 두 변수 간에 얼마나 밀접한 선형관계를 가지고 있는 가를 분석하는 통계기법으로, 두 변수간의 관계의 강도를 상관관계(correlation analysis)라 한다. 상관관계는 분석방법에 따라 단순히 두 개의 변수가 어느 정도 강한 관계에 있는 가를 측정하는 단순상관분석(simple correlation analysis)과 3개 이상의 변수들 간의 관계에 대한 강도를 측정하면 다중상관분석이라 하는데, 다중상관분석에서 다른 변수들과의 관계는 고정되고, 두 변수만의 관계에 대한 강도를 나타내는 것을 편상관분석(partial correlation analysis)이라고 한다. 이때, 상관관계가 $0<\rho\leq+1$ 이면 양의 상관, $-1\leq\rho<0$ 이면 음의 상관, $\rho=0$이면 무상관이라고 한다. 그러나 0인 경우 상관이 없다는 것이 아니라 선형의 상관관계가 아니라는 것이다. 본 연구에서는 이와 같은 이론을 근거로 사용되고 있는 피어슨 상관계수(Pearson correlation coefficient)를 활용하여 상관관계분석을 실시하였다.

Table 1. Model and equation used in regression

Model	Equation
Linear	$E(Y_t) = A_1 + A_2 t$
Quadratic	$E(Y_t) = A_0 + A_1 t + A_2 t^2$
Cubic	$E(Y_t) = A_0 + A_1 t + A_2 t^2 + A_3 t^3$
Exponential	$E(Y_t) = A_0 e^{A_1 t}$
Logistic	$E(Y_t) = (\frac{1}{u} + A_0 A_1^t)^{-1}$

Ⅲ. 통계적 기법을 이용한 가스사고 예측

본 연구에서는 상관분석을 통하여 기상요소와의 가스사고와의 관계분석을 위해 SPSS프로그램을 활용하여 요소별 상관분석과 회귀분석을 실시하여 가스사고의 발생가능 예측식을 제시하였다[4]. 회귀분석에서 사용된 예측 모델식은 linear, quadratic, cubic, exponential, logistic 모형의 수식으로 이를 나타내면 Table 1과 같다.

3.1. 기상 요소별 가스사고 영향도

가스사고 예측모형을 구성하기 위하여 다음 평균기온, 최고기온, 최저기온, 상대습도, 운량, 강수량, 평균풍속과 가스사고의 발생에 대하여 분석하였다.

가스사고 예측 연구 모형을 구성하기 위하여 개별 요소의 가스사고와의 관련성을 회귀분석을 통해 분석하여 독립변인들이 가스사고에 미는 영향정도(R^2)를 산출하였고, 각 회귀식의 구성 변인의 영향력의 크기를 표준화하여 표준화된 결정계수(adjusted R^2)를 산출하였다. 또한 표준화된 결정계수(adjusted R^2)들을 상대적 영향력의 크기 백분비로 환산하였다.

그 결과, Table 2에서와 같이 평균풍속이 가장 높은 설명력을 나타내었으며, 그 다음은 상대습도, 평균기온 순으로 영향력이 높은 것으로 나타났다.

3.2. 가스사고 발생 예측

본 연구에서는 통상 최고기온과 최저기온의 평균값을 평균기온으로 설정하여도 큰 차이가 없는 것으로 나타나서 사고예측을 위해 평균기온은 예보되는 최고기온과 최저기온의 평균값을 적용하고, 평균풍

Table 2. Relative effectiveness of meteorological elements

구 분	R^2	Adjusted R^2	상대적 영향력(%)	영향 순위
평균기온	.337	.332	17.548	3
최고기온	.321	.315	16.649	4
최저기온	.239	.234	12.368	5
상대습도	.351	.347	18.340	2
운량	.066	.035	1.850	7
강수량	.115	.11	5.814	6
평균풍속	.526	.519	27.431	1

속의 경우에도 예보되는 풍속으로 하여 가스사고의 발생가능성을 예측하였다.

7가지 기상요소를 이용한 가스사고 예측 및 기상요소 중 다소 설명력이 떨어지는 운량과 강수량을 제외하고, 이를 다시 분석하였다.

3.2.1 7가지 기상요소를 이용한 가스사고 예측

평균기온, 최고기온, 최저기온, 상대습도, 운량, 강수량 및 풍속의 7가지 추정 회귀식을 사용하여 가스사고를 예측하였다. 그 결과, 기상요소별 추정 회귀식은 정리하여 나타내면 Table 3과 같다.

기상요소에 의한 가스사고의 발생을 예측하기 위하여 각 기상요소에 따른 회귀식에 Table 2의 영향도를 곱하여 합한 값을 가스사고 발생 예측식으로 하고, 이를 수식으로 나타내면 식 (7)과 같다.

$$Q_1 = (Y_0 \times 0.17548) + (Y_1 \times 0.16649) + (Y_2 \times 0.12368) + (Y_3 \times 0.18340) + (Y_4 \times 0.01850) + (Y_5 \times 0.05814) + (Y_6 \times 0.27431) \quad (7)$$

여기서 Q_1은 7가지 기상요소에 의한 예측식, Y_0은 평균기온, Y_1은 최고기온, Y_2는 최저기온, Y_3는 상대습도, Y_4는 운량, Y_5는 강수량, Y_6는 풍속이다.

수식의 특성상 사고발생 빈도가 (-)의 값이 나오는 경우는 0으로 산정한다. 각 수식을 적용해 본 결과, 국내에서 나타날 수 있는 기상요소 값 중 평균기온, 최고기온, 최저기온, 강수량, 풍속에서 (-)의 값이 나오는 것으로 나타났다. Table 4와 같이 기상요소별 한계수치 이상인 경우는 그 한계수치의 값을 적용한다.

회귀식의 합에 따라 가스사고의 발생가능성을 매우 높음, 높음, 보통, 낮음, 매우 낮음의 5등급으로 구분하여 위험지수로 나타내었다. 위험지수의 목적은 가스 사용자의 경각심의 고취와 가스 안전관리 상황의 적절한 대응에 있다 하겠다.

각 함수에 3.1절에서 분석된 기상조건별 가스 사고 발생이 높은 순으로 각각의 수치를 회귀식에 대입하여 사고빈도를 추출하였다. 그 결과는 Table 5와 같다.

또한 평균기온, 최고기온, 최저기온, 상대습도, 운량, 강수량, 풍속에 대해 계산된 회기식인 Table 3에 Table 5의 각 기상요소별 사고빈도 값을 대입하여 값을 추출하였다. 이 추출된 값을 예측식 (7)에 대입하여 예측값을 계산하고, 지수화하여 그 결과를 다음과 같은 Table 6과 Fig. 1로 나타내었다.

3.2.2 5가지 기상요소를 이용한 가스사고 예측

기상요소 중 설명력이 다소 떨어지는 운량과 강수량을 제외하고, 이를 다시 분석하였다. 이 경우 기상

Table 3. Regression equation for each meteorological element

구 분	회귀식
평균기온	$Y_0 = 0.311 \times T_{mean} - 0.002 \times T^2_{mean} + 0.0001 \times T^3_{mean} + 4.594$
최고기온	$Y_1 = 0.233 \times T_{max} + 0.012 \times T^2_{max} - 0.001 \times T^3_{max} + 2.732$
최저기온	$Y_2 = 0.256 \times T_{min} - 0.007 \times T^2_{min} + 0.0001 \times T^3_{min} + 5.788$
상대습도	$Y_3 = -0.435 \times H + 0.013 \times H^2 - 9.14\times10^{-5} \times H3 + 5.352$
운량	$Y_4 = -6.173 \times C + 0.554 \times C^2 + 35.96$
강수량	$Y_5 = -0.008 e^P + 2.606$
풍속	$Y_6 = -0.429 e^W + 65.333$

* T_{mean} : 평균기온, T_{max} : 최고기온, T_{min} : 최저기온, H : 상대습도, C : 운량, P : 강수량, W : 풍속

Table 4. Limited and applied value for each meteorological element

기상요소	한계값	적용값
평균기온(℃)	-12미만	-12
최고기온(℃)	25.4초과	25.4
최저기온(℃)	-15미만	-15
강수량(mm)	5.7초과	5.7
풍속(m/s)	5초과	5

Table 5. Frequency of accident by each meteorological element on degree of danger

구분	Extreme (매우 높음)	High (높음)	Moderate (보통)	Low (낮음)	Very low (매우 낮음)
평균기온	9.1<	9.0~6.1	6.0~4.6	4.5~3.6	3.5>
최고기온	5.0<	4.9~3.5	3.4~3.0	2.9~2.5	2.4>
최저기온	8.4<	8.3~6.5	6.4~4.6	4.5~2.4	2.3>
상대습도	6.5<	6.4~5.5	5.4~4.5	4.4~3.5	3.4>
운량	33<	32~25	24~22	21~19	18>
강수량	2.5<	2.4~2.2	2.1~1.8	1.7~1.4	1.3>
풍속	60<	59~50	49~40	39~25	24>

Table 6. Index of gas accident by frequency of accident

예측값(Q_1)	위험지수	위험수준
22이상~25미만	81~100	Extreme(매우 높음)
18이상~22미만	61~80	High(높음)
14이상~18미만	41~60	Moderate(보통)
9이상~14미만	21~40	Low(낮음)
1이상~9미만	0~20	Very low(매우 낮음)

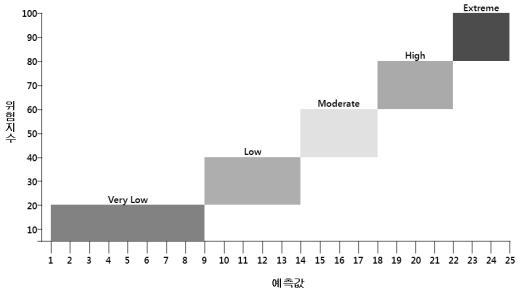

Fig. 1. Index distribution of gas accident.

Table 7. Relative influence for each meteorological element (exception of the amount of clouds and precipitation)

구 분	R^2	Adjusted R^2	상대적 영향력(%)	영향 순위
평균기온	.337	.332	19.000	3
최고기온	.321	.315	18.000	4
최저기온	.239	.234	13.400	5
상대습도	.351	.347	19.900	2
풍속	.526	.519	29.700	1

요소의 축소로 인해 각 기상요소별 상대적 영향력이 변경되어야 하고, 이로 인해 위험지수도 변경되어야 한다. 운량과 강수량을 제외한 상대적 중요도는 Table 7에 나타내었다.

따라서 상대적 중요도를 고려한 5가지 기상요소에 대한 가스사고발생 예측식을 나타내면 식 (8)과 같다.

Table 8. Index of gas accident by frequency of accident

예측값(Q_2)	위험지수	위험수준
23이상~25미만	81~100	Extreme(매우 높음)
19이상~23미만	61~80	High(높음)
15이상~19미만	41~60	Moderate(보통)
10이상~15미만	21~40	Low(낮음)
1이상~10미만	0~20	Very low(매우 낮음)

$$Q_2 = (Y_0 \times 0.19) + (Y_1 \times 0.18) + (Y_2 \times 0.134) + (Y_3 \times 0.199) + (Y_6 \times 0.297) \quad (8)$$

여기서 Q_2는 5가지 기상요소에 의한 예측식, Y_0은 평균기온, Y_1은 최고기온, Y_2는 최저기온, Y_3는 상대습도, Y_6는 풍속이다.

또한 평균기온, 최고기온, 최저기온, 상대습도 및 풍속의 기상요소별 사고빈도를 앞서 구한 예측식 (8)에 대입하여 예측값을 계산하였다. 추출한 예측값은 지수화하여 Table 8에 나타내었다.

Ⅳ. 결 론

본 연구에서는 가스사고의 지역별 발생현황과 발생시기의 계절적 요인 및 기상청의 기상자료를 종합적으로 분석·적용하여 가스사고의 발생가능성과 발생건수를 예측할 수 있는 모델을 도출하고, 그 모델의 활용가능성 여부를 제시하였다. 본 연구를 통한 결론은 다음과 같다.

(1) 각 가스사고별 기상자료를 통계프로그램인 SPSS v15.0을 활용하여 상관분석과 회귀분석을 실시하였으며, 그 결과 풍속에 대한 회귀식의 설명력이 52.5%로 가장 높은 것으로 나타났으며, 다음으로 상대습도가 35.1%, 평균기온이 33.7%, 최고기온이 32.1%, 최저기온이 23.9%, 강수량이 11.5%, 운량이 6.6%로 나타났다.

(2) 7가지 요소 중 강수량 및 운량의 설명력이 타 요소 보다 낮으나 전체 예측식에서 보면 설명력이 낮지 않기 때문에 운량과 강수량을 포함한 경우와 큰 차이가 없어 7가지 기상요소를 모두 감안한 회귀식을 가스사고 발생 예측식으로 활용하여도 무방할 것으로 판단하였다.

(3) 가스 사고빈도의 정도에 따라 가스사고 발생 위험지수를 매우 높음, 높음, 보통, 낮음, 매우 낮음

등 5단계로 구분하였으며, 예측값이 22이상 25미만이면 매우 높음, 18이상 22미만이면 높음, 14이상 18미만이면 보통, 9이상 14미만이면 낮음, 1이상 9미만이면 매우 낮음으로 위험지수를 정하여 가스사고 예방활동에 활용할 수 있도록 하였다.

이상과 같이 본 연구에서는 가스사고의 예측도 향상과 적절한 예방대책 수립을 위하여 지속적인 연구의 필요성을 제안하였다. 가스안전전문기관인 한국가스안전공사에서도 가스사고의 발생가능성과 건수를 회사의 업무시스템과 연계시켜 실시간으로 확인이 가능하도록 하여 가스사고 예방활동에 적극 활용할 수 있을 것으로 사료되며, 제시된 모델식을 활용하여 기상청의 일기예보 시스템과 연계, 가스사고 발생가능성을 시각화시켜 실시하여 일반 국민들에게도 제공할 수 있을 것으로 본다.

참고문헌

[1] 기상청(http://www.kma.go.kr/)

[2] 소방방재청(http://www.nema.go.kr/)

[3] 허영택, 이수경, "가스사고 발생 환경분석을 통한 사고발생 모형 고찰, *KIGAS*, **14**(2), 27-33, (2010)

[4] 손충기, 백영균 등 3명, *내가하는 통계분석 SPSS (제4판)*, 학지사, (2007)

[5] 한국가스안전공사, *고압가스통계*, (2009)

[6] Young-Taeg Hur, Ha-Kyung Lim, and Su-Kyung Lee, "The Reformation of Gas Technical Standards System", *KIGAS*, **12**(3), 20-23, (2008)

[7] 한국가스안전공사, *가스사고연감*, 한국가스안전공사 사고점검처, (1998~2009)

[8] 성혁제, 안동균 등 4명, *데이터 분석을 위한 통계기법 및 SPSS 활용*, 동일출판사, (2003)

[9] 위영민, 송경빈 등 3명, "결정계수 기반의 데이터 마이닝을 이용한 특수일 최대 전력 수요 예측", *대한전기학회 하계학술대회 논문집*, 552-553, (2008)

[10] 이인범, "미국의 가스사고 조사 및 분석기법", 가스안전지, 통권 제182호, 14-20, (2006)

[11] Dong-Il Seol, "Climatological Characteristics of Monthly Wind Distribution in a Greater Coasting Area of Korea", *J. of the Korean Society of Marine Environment & Safety*, **12**(3), 185-192, (2006)