적외선 신호 해석을 위한 해양 기상 표본 추출법

김윤식^{1,†} · Vaitekunas, D.A.²
선박해양플랜트연구소 미래선박연구부¹
W.R. Davis Engineering Ltd.²

A New Sampling Method of Marine Climatic Data for Infrared Signature Analysis

Yoonsik Kim^{1,†} · David A. Vaitekunas²
Advanced Ship Research Division, KRISO/KIOST, Daejeon¹
W.R. Davis Engineering Ltd., Ottawa, Canada²

This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(<http://creativecommons.org/licenses/by-nc/3.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper presents a new method of sampling the climatic data for infrared signature analysis. Historical hourly data from a stationary marine buoy of KMA(Korean Meteorological Administration) are used to select a small number of sample points ($N=100$) to adequately cover the range of statistics(PDF, CDF) displayed by the original data set ($S=56,670$). The method uses a coarse bin to subdivide the variable space ($3^5=243$ bins) to make sample points cover the original data range, and a single-point ranking system to select individual points so that uniform coverage ($1/N = 0.01$) is obtained for each variable. The principal component analysis is used to calculate a joint probability of the coupled climatic variables. The selected sample data show good agreement to the original data set in statistical distribution and they will be used for statistical analysis of infrared signature and susceptibility of naval ships.

Keywords : Marine climate buoy(해양 기상관측 부이), Sampling method(표본추출법), Infrared analysis(적외선 해석), Principal component analysis(주성분 분석법), Cumulative distribution function(누적분포함수), Correlation(상관관계)

1. 서론

함정의 적외선(IR, InfraRed) 스텔스 성능은 해양의 기상조건 변화에 매우 큰 영향을 받는다. 함정 적외선신호(IR signature), 대함 미사일에 의한 피탐거리(detection range) 등에 대한 해석을 위해서는 대상함이 운용될 수 있는 해상의 기상조건(기온, 수온 등)을 정확하게 적용하여 해석해야만 대상함의 신호 및 피탐특성을 정확하게 예측할 수 있다 (Vaitekunas, 2006). 설계중인 함정에 대해 대상함이 나타낼 것으로 예상되는 신호의 변화 범위를 예측하는 것은 건조될 함정에 대한 요구 신호성능(signature requirement)을 결정하는데 중요한 참고자료로 활용될 수 있다. 따라서 정확한 신호예측이 필요하며 신호예측결과에 가장 큰 영향을 미치는 기상조건의 설정은 함의 신호예측에서 가장 중요한 부분이라 할 수 있다.

해양기상 관측데이터는 그 방대한 관측데이터로 인해 직접 해석에 적용하는 것은 불가능하다. 이로 인해 국내외에서는 적외선 신호해석을 위해서 월별 평균기상조건을 설정하고 해석을 수행하는 방법을 활용해왔다 (Kim, 2012; Vaitekunas, 2010). 이 방법은 다음과 같이 요약할 수 있다. 먼저 기상변수(기온, 수온 등)별

로 월별 평균/표준편차를 구하고 12개월 각각에 대한 평균기상조건을 설정한다. 그리고 각 조건별로 함의 신호를 해석한 후 가장 큰 신호(또는 피탐거리)를 나타내는 달과 그 신호값을 구한다. 이후 각 기상요소들의 월별 표준편차를 이용한 영향 평가를 통해 함의 신호변화 특성을 고찰한 후 가장 큰 신호를 나타내는 기상조건('기준환경') 등을 설정한다.

이와 같은 기존의 기상조건 설정법은 다음과 같은 문제점들이 있다. 첫째, 모든 기상변수들이 독립적으로 변화할 수 있다고 가정함으로써 실제 기상특성과 차이를 발생시킬 수 있다. 즉, 기온, 수온, 상대습도, 풍속 등의 해양기상 변수들 중 임의의 변수는 다른 변수들의 변화특성과 무관하게 변화할 수 있다고 가정하고 있다. 그러나 실제 해양기상관측데이터를 통해 기상변수들간의 상관관계를 분석하면 서로 독립적으로 변화하지 않는다는 것을 알 수 있다. 둘째, 해석대상 함정의 운용조건과 해양기상조건의 상호작용을 고려하지 않음으로 인해 운용조건별 '기준환경'을 별도로 설정해야하고, 이로 인해 해석결과와 분석 과정이 매우 복잡한 결과를 초래할 수 있다. 예를 들어, 함 표면에 washdown을 적용하는 경우 'dry-ship'에 비해 기온, 수온 등의 외부 기상요소의 영향이 크게 변화하게 된다. 이로 인해 각 운용조건별로 서로 다른

‘기준환경’에서 해석이 수행됨으로써 신호저감 효과분석 등이 곤란해지는 문제점이 있다.

이 연구에서는 실제 관측된 해양기상 데이터에서 함정의 신호 해석에 활용 가능하도록 적절한 표본집단을 추출하고 그 표본집단에 대하여 신호 해석을 수행하면 전체 관측데이터에 대한 해석 결과를 추정할 수 있을 것으로 가정하였다. 이러한 특성을 만족시키기 위해 표본집단 추출을 위한 요구조건을 설정하고 적절한 표본을 추출하기 위한 방법을 이 논문에서 제안한다. 추출된 기상 표본집단을 이용한 함정 적외선신호 해석 및 평가 과정에 대해서는 별도의 연구논문으로 발표할 예정이다.

2. 표본추출 절차

2.1 기상관측데이터 획득 및 전처리

기상청에서는 국내해상의 해양기상 관측을 위해 부이(buoy)와 등표를 설치 운영하고 있다 (KMA, 2013). 관측데이터들은 관련 연구 등을 위한 자료로 일반에 제공하고 있다. Fig. 1은 기상청에서 운영하고 있는 10개 부이들의 설치 위치를 나타낸 것이다. 부이를 이용한 관측자료들은 기온, 수온, 상대습도, 풍향/풍속, 파고, 파향 등으로써 다양한 기상 및 해양관측 자료를 제공하고 있다. 이 연구에서는 ‘동해’ 부이(Fig. 1의 No.5, 37.5°N/130.0°E)에서 2001년 5월부터 2013년 4월까지 1시간 간격으로 관측된 자료를 이용하였다. 함정의 적외선 신호해석을 위해 필요한 기상변수는 5가지(기온, 수온, 상대습도, 풍속, 풍향)이며, 나머지 변수들에 대한 관측결과는 활용하지 않았다.

Fig. 1 Locations of KMA buoys

기록된 기상관측 자료는 관측장비 오류 등으로 인해 오염된 자료가 포함될 수 있으므로 품질검사와정을 통해 오염된 자료들을 걸러낸다. 품질검사를 위한 방법은 기상청의 품질검사 기준 (KMA, 2006)에 따라 수행하였다.

일반적으로 해양기상데이터는 태양 일주운동(주간/야간)에 따른 기상변수의 변화는 크지 않으나 월별변화에 따른 기상변수의 변화는 매우 크게 나타나는 특징이 있다 (Kim, 2012). 이 때문에 특정 월에 데이터가 집중되지 않도록 하는 것이 필요하다. 월별 데이터 수를 비교하여 가장 적은 데이터를 갖는 달을 기준으로 나머지 달의 데이터는 난수(random number)를 이용하여 제거함으로써 각 월별 데이터 수를 유사하게 설정하였다. 본 연구에서 활용한 ‘동해’부이의 경우 최종 데이터는 56,670개를 획득하여 모집단으로 활용하였다.

2.2 표본추출을 위한 요구사항

표본집단을 구성하는 각 기상표본들은 평균, 표준편차 등을 이용하여 인위적으로 생성된 것이 아니고 실제 관측된 것들 중에 추출하는 방식을 적용하였다. 이것은 인위적으로 표본을 생성하는 경우 여러 기상변수들 간 상관관계(correlation)의 정확한 반영 및 실제 발생가능성을 담보하기 어려워질 수 있기 때문이다. 모집단에서 표본 추출을 위한 요구조건들은 다음과 같다.

첫째, 추출된 표본 데이터들의 5가지 기상변수들 각각은 모집단 기상변수들의 분포특성에 최대한 근접하도록 한다.

둘째, 추출된 표본집단이 모집단의 특정 구역에 집중되지 않고 고루 분포할 수 있도록 한다.

셋째, 가급적 다양한 기상조건을 포괄하도록 한다.

이때 추출할 표본집단의 크기(N)가 지나치게 클 경우 추후 수행할 적외선 신호해석의 경우의 수가 늘어나는 문제점이 있으므로 적절히 제한되어야하고, 지나치게 작을 경우 모집단의 통계적 특성을 반영하기 어렵게 된다. 이 논문에서는 적외선 신호 해석 소요 시간 등을 고려하여 표본집단의 크기 N=100로 설정하였다. 이 논문에서 표본추출 방법에 대한 설명을 위해서는 N=10의 조건을 이용하여 기술하도록 한다.

위 각 요구조건들을 적용하기 위한 구체적인 방법은 다음과 같다.

2.2.1 모집단과 표본집단의 유사 확률분포 유지

표본집단의 확률분포특성을 모집단과 유사하게 갖도록 하기 위하여 먼저 다음과 같은 확률밀도함수(PDF, Probability Density Function)와 누적분포함수(CDF, Cumulative Distribution Function)를 모집단의 각 기상변수별로 계산한다.

$$PDF = f(x), \text{ such that, } P(a < X \leq b) = \int_a^b f(x) dx \quad (1)$$

$$CDF = F(x) = \int_{-\infty}^x f(x) dx, \quad f(x) = \frac{d}{dx} F(x) \quad (2)$$

$$F_X(x) = P(X \leq x)$$

Fig. 2(a)는 모집단의 기온(T_a), 수온(T_s), 상대습도(RH)의 CDF를 나타내고 있으며, Fig. 2(b)는 풍속(W_s), 풍향(W_d)의 CDF를 각각 나타내고 있다. 그리고 Fig. 3은 기온의 PDF를 나타낸 것이다.

Fig. 2(a) CDF distributions of T_a , T_s , RH

Fig. 2(b) CDF distributions of W_s , W_d

Fig. 3 PDF distributions of air temperature(T_a)

각 기상변수들의 PDF, CDF 분포를 보면 정규분포(normal distribution)와 매우 상이한 형태를 띠는 것을 확인할 수 있다. Fig. 2(a)에 나타낸 기온의 CDF와 Fig. 3에 나타낸 기온의 PDF를 비교하면 그 특징을 쉽게 파악할 수 있다. CDF의 기울기가 큰 구간에서 PDF가 큰 값을 가지는 특징을 그래프와 정의식에 의해 확인할 수 있다. 이러한 특징을 이용하여 표본 추출을 위한 구간을 설정한다.

표본을 추출하기 위해 모집단의 각 기상변수별 CDF를 추출하고자 하는 표본의 수(N)만큼 등간격으로 분할한다(Fig. 4, 5의 세로축 참조). 이때 각 CDF값에 해당하는 기온은 불균일한 분포를 나타내고, 각 CDF 구간에 대응되는 기상변수 구간들을 가지게 된다(Fig. 4, 5의 가로축 참조). Fig. 4와 Fig. 5는 $N=10$ 인 경우에 대해, 기온(T_a)과 수온(T_s)의 CDF를 균등분할한 결과를 나타내고 있다. 즉, $1/N = 0.1$ 간격으로 CDF를 균등분할하면 그때의 기온, 수온의 분할 간격은 기온, 수온의 PDF(Fig. 3 참조) 값에 반비례하여 결정됨으로써 PDF가 큰 값을 갖는 구간에서는 좁은 구간으로 분할된다. 이렇게 분할된 각 기상변수(x)들의 구간들을 '슬롯'('Slot')라고 칭한다.

Fig. 4 Division of CDF into N slots(T_a)

Fig. 5 Division of CDF into N slots(T_s)

$$Slot(x_j) = x_j - x_{j-1}, j = 1, 2, \dots, N \quad (3)$$

$$x_j = F^{-1}(j/N)$$

따라서 각 기상변수별 슬롯의 수는 추출할 표본의 수와 같아지며, 각 변수별(기온, 수온 등)로 각 슬롯에서 중복되거나 빠짐없이 하나씩의 표본을 추출하면 추출된 표본들이 갖는 확률분포는 모집단이 갖는 확률분포에 근접하게 된다. 그러나 예외적인 처리를 해야 하는 경우도 발생할 수 있다. 이 경우는 관측된 데이터의 범위와 정밀도에 비해 지나치게 많은 구간을 분할할 때 발생한다. 예를 들어, 기록된 상대습도 데이터의 정밀도가 1%단위로 표기되고, 관측된 데이터가 20~100%까지인 경우 최대 80개 이상의 슬롯을 만들지 못하는 문제가 발생한다. 이러한 예외처리 방안은 이후의 절에서 별도로 기술하도록 한다.

2.2.2 표본추출구역 다각화 방안

Fig. 6과 Fig. 7은 기상관측데이터(모집단)에서 기온과 수온, 기온과 상대습도를 각각 표시한 것이다.

Fig. 6 Distribution of observed data(Ta vs. Ts)

Fig. 7 Distribution of observed data(Ta vs. RH)

Fig. 6에서 확인할 수 있는 바와 같이 기온과 수온은 매우 강한 상관관계를 갖는다. Table 1은 5개의 기상변수들 간의 상관관계계수(correlation coefficients)를 정리한 것이다. 기온과 수온이 가장 강한 상관관계를 나타내고 있으며, 기온과 상대습도, 그리고 수온과 상대습도의 순으로 상관관계를 가짐을 확인할 수 있다.

Table 1 Correlation coefficients of climatic variable

	Ta(°C)	Ts(°C)	RH(%)	Ws(m/s)	Wd(°)
Ta(°C)	1				
Ts(°C)	0.869	1			
RH(%)	0.554	0.372	1		
Ws(m/s)	-0.322	-0.215	-0.129	1	
Wd(°)	-0.217	-0.172	-0.264	0.132	1

이러한 상관관계뿐만 아니라 모집단은 매우 광범위하게 분포되어 있는 특성이 있다. 추출된 표본을 이용한 해석결과가 광범위한 기상변화조건에 대한 해석결과를 대표하기 위해서는 다양한 조합에 해당되는 표본을 추출하는 것이 필요하다. 따라서 각 기상변수들에 대해 구역을 분할하고 각 구역별로 표본을 추출하는 방법을 적용하였다. 분할하는 구역의 수(B)는 추출하고자 하는 표본의 수(N)보다 커야하며, 전체 구역의 수는 각 변수별 분할 구간의 곱으로 결정된다. 즉, 식 (4)와 같다.

$$B_{tot} = B_{Ta} \times B_{Ts} \times B_{RH} \times B_{Ws} \times B_{Wd} \quad (4)$$

각 변수별 분할 구간의 수를 동일하게 하는 경우 $B_{tot} = B_X^5$ 가 된다. 따라서 $N=100$ 인 경우 $N < B_{tot} = B_X^5$ 이므로 $B_X \geq 3$ 이 된다. 즉, 각 변수별로 세 구간 이상으로 분할해야 한다. 이때 분할 기준은 각 변수별 CDF값을 기준으로 한다. $B_X = 3$ 인 경우 $CDF(X) = 0.33, 0.67$ 을 기준으로 분할한다. 이와 같이 분할하면 각 분할 구간 내에 위치하는 관측데이터의 수가 동일해진다. 예를 들어 기온(Ta)의 $CDF = 0.33, 0.67$ 에 해당되는 $Ta = 9.9(^{\circ}\text{C}), 18.8(^{\circ}\text{C})$ 이고, 수온(Ts)은 $13.7(^{\circ}\text{C}), 20.1(^{\circ}\text{C})$ 이다 (Fig. 2(a) 참조). 이와 같은 방식으로 다른 기상변수(상대습도, 풍속, 풍향)에 대해서도 구간을 나눌 수 있고 모든 변수들에 대해 구간을 적용하면 5차원적 구역(block)으로 설정되며 전체 구역의 수는 $3^5 = 243$ 개가 된다. 이해를 돕기 위해 만일 기온과 수온만을 고려하는 것으로(2차원 문제)로 단순화시키면 구역의 수는 9개가 되며 Fig. 8에 나타낸 것과 같이 각 구역을 표시할 수 있다. 이때 각 구역이 포함하는 관측데이터의 수가 동일한 것은 아니다. 수평으로 분할된 구역들(수온 기준 분할) 3개에는 모두 같은 수의 데이터가 존재하고, 수직으로 분할된 구역들(기온 기준 분할) 3개도 같은 수의 데이터가 존재하지만, 수평과 수직으로 모두 분할된 구역들은 동일한 데이터가 존재하지 않고 각 구역별로 다른 수의 데이터가 포함된다. 이러한 현상은 차원의 수를 2차원(기온 - 수온)에서 5차원(기온 - 수온 - 상대습도 - 풍속 -

풍향)까지 올리는 경우도 동일하게 나타난다. 즉, 243개의 구역들에는 서로 다른 개수(n)의 관측데이터들이 포함되어 있으며, 관측데이터를 전혀 포함하지 않는 ($n=0$) 경우도 다수 생길 수 있다 (예, Fig. 8에서 구역-9).

Fig. 8 Domain decomposition of climatic data(for B=3)

추출된 표본이 다양한 기상조건을 대표하도록 하기위해서 각 구역별로 1개씩의 표본만을 추출하도록 하였다. 이때 각 구역에 포함되어 있는 관측데이터의 수(n)는 구역마다 다르며, 많은 데이터를 포함하는 구역($n/S > 0.5$)과 적은 데이터를 포함하는 구역($n/S < 0.5$)을 나누고, 번갈아가며 구역을 선택하였다. 구역을 선택하는 과정에서는 난수를 이용하였다.

2.2.3 다양한 기상조건 포괄 방안

편의상 Fig. 8에서 구역-8을 기준으로 설명하면, 구역 내에 포함되어 있는 데이터들에 대해 차별적인 우선순위를 갖도록 설정하였다. 동일한 구역내에 포함되어 있더라도 그 데이터의 발생확률을 계산하여 낮은 발생확률을 갖는 데이터에 높은 우선순위를 부여하였다. 예를 들면 Fig. 8에서 점 A와 점 B를 비교하면 점 A는 분포영역 경계에 근접한 특성이 있으며, 점의 발생확률도 점 B에 비해 낮은 특징을 갖는다. 즉, 가능하면 분포형상의 경계에 위치한 점을 우선적으로 추출하기 위해 점 A에 높은 우선순위를 부여하는 방법을 사용하였다.

이와 같은 방법을 적용하기 위해서는 각 데이터들이 발생하는 발생확률을 계산해야 한다. 관측된 기상데이터들의 발생확률을 계산하기 위해서는 기상변수들간의 상관관계를 고려한 발생확률을 계산해야 한다. 즉, 결합확률(joint probability)을 계산해야 한다. 각 기상변수들은 서로 상관관계를 가지므로(즉, 서로 독립이 아니므로) 각 변수별 발생확률을 구한 후 각각을 곱하여 구할 수 없다. 이와 같이 서로 결합된(coupled) 변수들의 발생확률을 구하기 위해서 주성분분석법(PCA, Principal Component Analysis)을 이용하였다 (Jolliffe, 2010; Wikipedia, 2013). 주성분분석법을 이용하여 각 기상변수(X_k)를 서로 독립인 변수 Y_k 로 변환한

후, Y_k 에 대한 발생확률을 계산함으로써 각 데이터들의 발생확률을 계산할 수 있다. 이때 Y_k 는 물리적인 특성을 갖는 변수가 아니며 단지 수학적으로만 정의되는 변수이다.

주성분분석법을 위해 기상변수 X_k 에서 독립변수 Y_k 로 변환하는 절차는 다음과 같다.

- (1) X_k 의 각 변수들간의 공분산행렬(covariance matrix, C)를 구한다.
- (2) C 의 고유치(eigenvalues) λ_k 와 고유벡터(eigenvectors) ϕ_k 를 구한다.
- (3) λ_k 의 크기순으로 ϕ_k 를 정렬하고 고유벡터로 구성된 행렬(Φ)을 구성한다.
- (4) $Y = X \Phi$ 로 변환한다.

위와 같이 변환하고 Y_k 각 변수들 간의 상관관계계수를 구하면 아래 Table 2와 같이 나타난다. 즉, 변수 Y_k 는 서로 독립임을 확인할 수 있다.

Table 2 Correlation coefficients between variable Y_k

	Y1	Y2	Y3	Y4	Y5
Y1	1				
Y2	0	1			
Y3	0	0	1		
Y4	0	0	0	1	
Y5	0	0	0	0	1

Y_k 에 대한 PDF는 Y_k 에 대한 CDF($F^k = F(Y_k)$)를 이용하여 아래 식 (5)과 같이 구할 수 있다.

$$P_i(Y_k) = (F_{i+1}^k - F_{i-1}^k) / (Y_{k,i+1} - Y_{k,i-1}), \quad (5)$$

for $k = 1, \dots, 5, i = 1, \dots, S$

5개의 변수 Y_k 에 대한 PDF를 아래 식 (6)과 같이 곱하면 각 점의 결합확률(joint probability)를 구할 수 있다.

$$P_{J_i} = P_i(Y_1) \times P_i(Y_2) \times \dots \times P_i(Y_5) \quad (6)$$

식 (6)과 같이 구한 결합확률 값은 매우 작은 값을 가진다. 따라서 수치적 안정성 확보를 위해 아래 식 (7)과 같이 전체 합을 이용하여 표준화(normalize) 한 후 사용하였다.

$$P_{J_i} = P_{J_i} / \sum_{m=1}^S P_{J_m} \quad (7)$$

위와 같이 구한 각 관측데이터의 발생확률들을 서로 비교함으로써 낮은 발생확률을 갖는 점이 높은 우선순위를 갖도록 설정하였다.

2.3 표본추출 절차

2.3.1 표본추출 우선순위 설정

추출 후보가 되는 모든 관측데이터들은 각 기상변수별 해당 슬롯의 합(m)과 발생확률(P_j)값을 가진다. 이때 m 은 5 이상의 값을 가질 수 있으며, m 값을 결정하는 과정에서 2.2.1절에서 언급한 바와 같이 CDF를 기준으로 N 개로 구간을 분할할 때 발생할 수 있는 중복조건에 대해 고려해야 한다. 관측데이터의 구분 가능한 최대 데이터 개수는 기상변수별로 다음과 같이 나타낼 수 있다.

$$\text{구분가능 데이터 개수}(D) = (\text{최대값} - \text{최소값}) / \text{유효정밀도} \quad (8)$$

위 식 (8)에서 D 가 N 보다 작은 경우나 특정구간에 데이터들이 집중되어 있는 경우 중복조건이 발생한다. 즉, 슬롯의 폭이 0이 되는 조건으로써 아래 식 (9)와 같이 된다.

$$x_j = x_{j-1} = \dots = x_{j-p} \quad (9)$$

$$\text{여기서, } x_j = F^{-1}(j/N)$$

예를 들어, 상대습도 관측데이터가 1%단위로 표기된 경우 전체 관측데이터의 3%에 해당하는 데이터가 상대습도(RH) = 60%의 값을 나타내는 경우를 가정하면, $F^{-1}(0.80) = F^{-1}(0.81) = F^{-1}(0.82) = 60\%$ 와 같은 결과를 나타내고, 해당 슬롯의 간격이 0이 된다. 이와 같은 경우 간격이 0이 되는 슬롯을 삭제하고 다음 슬롯에 중복 추출될 수 있도록 설정하였다. 이 과정을 거치면 중복 구간에 해당되는 데이터는 해당 슬롯의 합(m)이 5 이상의 값이 될 수 있다. 각 기상변수별로 해당되는 슬롯의 수를 s 라 할 때,

$$s_{Ta} = s_{Ts} = s_{Ws} = s_{Wd} = 1$$

$$s_{RH} = 3, \text{ 이므로}$$

$$m = \sum_{k=1}^5 s_{X_k} = 7 \text{ 이 된다.}$$

이와 같이 중복 가능한 조건을 고려하여 각 후보점들의 해당 슬롯의 합(m)과 발생확률(P_j)을 이용하여 각 후보점들의 우선순위(Rank)를 다음과 같이 설정한다.

$$R_i = \frac{m_i}{P_{ji}}, \quad i = 1, \dots, S \quad (10)$$

2.3.2 표본추출 우선순위 조정

한 점이 추출되면 그 후 그 점과 동일 슬롯에 포함되는 점들의 m 값은 줄어들게 된다. 예를 들어, 첫 번째 추출한 점의 기온이 15(°C), 수온이 12(°C)일 때(Fig. 8의 점 A) 기온이 14.2 - 17.1(°C)에 해당되는(Fig. 4 참조) 모든 점들(예, Fig. 8의 점 C)은 s_{Ta} 값이 1만큼(중복이 아닌 경우 1에서 0으로) 줄어들게 되고 수온이 10.5 - 12.1(°C)에 해당되는(Fig. 5 참조) 모든 점들(예, Fig. 8의 점 D)은 s_{Ts} 값이 1만큼 줄어들게 된다. 즉 먼저 추출된 표본에 의해 나머지 후보점들의 해당 슬롯 값이 변화하게 되며, CDF 구간별로 한 개씩 표본추출을 위해 다른 후보점들은 추출될 가능성이 크게 줄어들 수 있다. 이러한 과정을 기상변수가 기온과 수온만 갖는 2차원 경우를 가정하여 그림으로 도시하면 Fig. 9에 나타난 것과 같다. 점 A를 추출한 경우 점 A의 기온과 수온이 위치하는 구간과 중복되는 모든 점들(Fig. 9에서 회색 구역 내의 모든 점들)은 이후 표본 선택과정에서 해당슬롯의 합(m)이 줄어들게 된다.

Fig. 9 Schematic illustration of point selection process

2.3.2 표본추출 절차

앞서 기술한 내용을 바탕으로 전체 S 개의 관측데이터(모집단)에서 N 개의 표본을 추출하는 과정을 요약하면 다음과 같다.

- (1) 모집단의 각 기상변수별로 CDF(F)를 계산하고, N 개의 구간으로 분할하여 슬롯을 설정한다. 이때, 중복 추출가능 구간을 확인하고 중복 추출가능 표본의 수도 설정한다.
- (2) 모든 후보점들의 각 기상변수별로 해당 슬롯의 합(m)과 결합확률(P_j)을 계산하여 우선순위(R)를 계산한다.
- (3) N 개보다 많은 수의 구역을 생성하기 위하여 CDF를 이용하여 각 기상변수별로 구역을 나눈다. 그리고 각 구역별로 포함

된 데이터 수(n)를 확인하고 각 구역별 데이터 비(n/S)를 구한다.

- (4) 데이터 비(n/S)를 큰 순으로 정렬한 후, 데이터 비가 상위 50%에 속하는 구역(큰 구역)과 하위 50%에 속하는 구역(작은 구역)으로 분류하고, 큰 구역과 작은 구역을 번갈아가며 난수발생기를 이용하여 표본을 추출할 구역을 선택한다. 한번 선택된 구역은 다시 선택하지 않고 구역별로 한 개의 표본 추출을 원칙으로 한다.
- (5) 선택된 구역에 포함된 후보점들 중 우선순위(R)값이 가장 큰 점을 추출한다.
- (6) 표본이 선택 될 때마다 모든 후보점들의 해당 슬롯의 합(m)과 우선순위(R)를 갱신한다.
- (7) 원하는 개수의 표본이 추출될 때까지 (4)–(6)의 과정을 반복한다.

위의 표본추출 절차에 따라 $N=100$ 개의 표본 추출을 시도하였다. 그러나 표본에 대한 요구사항이 매우 복잡하여 $N=100$ 로 시작하면 100개의 표본을 모두 추출하지 못하는 문제가 발생하였다. 그 원인은 5개의 기상변수 모두에 대해 이전에 추출된 표본과 CDF 구간이 중복되지 않도록 요구하는 조건을 만족할 수 있는 표본을 찾지 못하는 경우가 발생하기 때문이다. $N=100$ 로 시작하는 경우 약 83개의 표본만을 추출할 수 있었다. 그래서 초기에 지정한 표본의 수를 증가시켜 $N_{ini}=123$ 개로 수정함으로써 최종적으로 $N_{final}=100$ 개의 표본을 추출할 수 있었다.

3. 표본추출 결과 검증

추출된 100개의 표본집단에 대하여 PDF와 CDF를 각각 구하여 모집단과 비교하였다. 모집단은 전체 데이터 수가 $S=56,670$ 개라서 PDF와 CDF계산 결과가 부드러운 변화를 나타낸다. 그러나 표본집단의 PDF와 CDF는 표본집단 내에서만 계산된 것임으로 인해 $1/100=0.01$ 간격으로 나타난다.

Figs. 10 ~ 14는 5개의 기상변수들에 대한 관측자료와 추출된 표본집단($N=100$)의 PDF와 CDF를 각각 비교한 것이다.

각각의 분포특성을 확인하면 대부분의 기상변수들은 해석적인 함수(analytic function) 형태로 표현하기 곤란한 분포양상을 나타내고 있으며, 풍속의 경우 Weibull 분포 (Ang & Tang, 2007; Vaitekunas, 2010)에 매우 근접한 결과를 확인할 수 있다.

Table 3은 모집단과 표본집단에 대하여 각 기상변수별 최대값, 최소값, 평균값(mean), 표준편차(standard deviation)를 비교한 것이다. 최대값과 최소값은 확률분포특성상 매우 발생확률이 낮음으로 인해 모집단과 다소 차이가 남을 확인할 수 있다. 그러나 평균과 표준편차는 모집단과 표본집단 간에 매우 근접한 결과를 보여주고 있음을 확인할 수 있다. 이러한 결과는 표본추출과정에서 평균, 표준편차에 대한 별도의 요구조건이 적용되지 않았지만 각 변수별 분포특성을 잘 따르도록 설정한 표본추출방법에 의한 결과로 생각된다.

Fig. 10(a) Comparison of PDF(air temperature)

Fig. 10(b) Comparison of CDF(air temperature)

Fig. 11(a) Comparison of PDF(sea temperature)

Fig. 11(b) Comparison of CDF(sea temperature)

Fig. 13(a) Comparison of PDF(Wind speed)

Fig. 12(a) Comparison of PDF(relative humidity)

Fig. 13(b) Comparison of CDF(Wind speed)

Fig. 12(b) Comparison of CDF(relative humidity)

Fig. 14(a) Comparison of PDF(Wind direction)

Fig. 14(b) Comparison of CDF(Wind direction)

Table 3 Comparison between buoy data and samples

		Min.	Max.	Mean.	St.Dev.
Ta (°C)	Meas.	-9.3	31.0	14.0	7.3
	Sample	-2.0	27.9	13.8	7.0
Ts (°C)	Meas.	1.5	29.2	17.1	5.0
	Sample	7.0	28.0	17.0	4.9
RH (%)	Meas.	13.0	100.0	71.2	16.5
	Sample	35.0	98.0	70.7	16.5
Ws (m/s)	Meas.	0.0	20.9	5.7	3.1
	Sample	0.2	17.2	5.9	3.2
Wd (°)	Meas.	0.0	360.0	192.3	106.0
	Sample	1.0	354.0	186.4	106.8

Fig. 15는 추출된 표본들의 분포특성을 나타내고 있다. Fig. 6에 나타난 관측데이터의 분포특성과 유사한 것을 확인할 수 있다.

Fig. 15 Distribution of sample data (Ta vs. Ts)

4. 결 론

본 논문에서는 기상청에서 운영하고 있는 부이에서 관측된 해양기상 관측데이터를 함정의 적외선신호 해석용으로 활용하기 위하여 제한된 수의 표본데이터를 추출하는 방법을 제안하였다. 표본추출을 위한 여러 요구조건들을 적용함으로써 추출된 표본들이 방대한 양의 해양기상관측데이터가 갖는 통계적 특성을 잘 따르도록 하고, 다양한 기상조건을 포함할 수 있도록 하였다. 표본추출을 위한 주요 요구조건은 다음과 같이 요약된다.

- (1) 모집단의 각 기상변수별로 누적분포함수를 구하고 CDF를 추출하고자 하는 표본의 수(N)만큼 균일하게 분할한다(슬롯 생성). 이후 각 슬롯별로 하나의 표본만을 추출함으로써 표본집단과 모집단의 CDF가 유사한 분포특성을 갖도록 한다.
- (2) 모집단의 여러 기상변수들의 분포범위를 충분히 포괄하기 위하여 구역을 분할한 후, 각 구역별로 하나의 표본만을 추출하도록 한다.
- (3) 각 구역별 표본추출과정에서 해당 구역에 포함된 데이터들 중 우선순위를 계산한다. 이때 각 데이터 점들의 발생확률 계산을 위해 주성분분석법을 적용하였다.

추출된 기상표본들과 관측데이터의 확률밀도함수(PDF), 누적분포함수(CDF)의 비교를 통해 매우 근접한 통계적 특성을 갖는 표본집단을 구성할 수 있음을 확인하였다.

이 논문에서 제안된 방법으로 추출된 기상표본데이터들은 다양한 해양기상조건에 따른 함정의 적외선신호 해석 및 평가 방법 개선을 위한 연구에 활용되며, 관련 연구결과는 별도의 논문으로 발표할 예정이다.

이 논문에서 제시한 표본추출 방법은 실제 관측된 기상데이터 또는 실험데이터에서 적절한 표본을 추출하여 해석 및 평가를 수행할 필요가 있는 분야에도 다양하게 활용될 수 있을 것으로 기대된다.

후 기

본 논문은 산업통상자원부에서 지원한 “에너지 절감을 위한 선박 저항감소 및 추진성능 향상 핵심 기술개발”과제(PNS2000)와 한국해양과학기술원 주요사업(PES156G)의 지원으로 수행되었음을 밝히며, 지원에 감사드립니다.

References

- Ang, A.H-S. & Tang, W.H., 2007. *Probability Concepts in Engineering – Emphasis on Applications to Civil and Environmental Engineering*. 2nd Ed. John Wiley & Sons, Inc: San Francisco.

- Jolliffe, I.T., 2010. *Principal Component Analysis*, 2nd Ed. Springer-Verlag: New York.
- Kim, Y., 2012. A Study on the Infrared Signature of a Naval Ship under the Marine Climate. *Journal of the Society of Naval Architects of Korea*, 49(3), pp.264-272.
- Korea Meteorological Administration (KMA), 2013. *KMA Business-Marine Observation*. [Online] Available at: <http://web.kma.go.kr/aboutkma/biz/observation07.jsp> [Accessed 18 July 2013].
- Korea Meteorological Administration (KMA), 2006, *User Manual of the Real Time Quality Control System of the Climate Observation Data(I)*. Korea Meteorological Administration: Seoul.
- Vaitekunas, D.A., 2006, Validation of shipIR(v3.2): methodology and results. *Proceeding of the SPIE 6239, Targets and Backgrounds XII: Characterization and Representation*, 62390K (04 May 2006), doi:10.1117/12.666790.
- Vaitekunas, D.A., 2010, IR susceptibility of naval ships using ShipIR/NTCS. *Proceeding of the SPIE Defense, Security and Sensing*, Orlando, Florida, USA, 5-9 April 2010, (SPIE Paper No. 7662-31).
- Wikipedia, 2013. *Wikipedia - Principal component analysis*. [Online] Available at: http://en.wikipedia.org/wiki/Principal_component_analysis [Accessed 19 July 2013].

김 윤 식

Vaitekunas, D.A.