Journal of the Korean Data & Information Science Society
2017, **28**(4), 797–810

http://dx.doi.org/10.7465/jkdi.2017.28.4.797
한국데이터정보과학회지

코플라함수를 이용한 극단치 강풍과 강수 분석

권태용[1] · 윤상후[2]

[1]대구대학교 일반대학원 통계학과 · [2]대구대학교 전산통계학과, 대구대학교 기초과학연구소

접수 2017년 5월 2일, 수정 2017년 6월 22일, 게재확정 2017년 6월 29일

요 약

한반도는 매년 태풍의 위험에 노출되어 있다. 태풍은 강풍과 강우가 동반되는 열대성 저기압으로 사회·경제적으로 막대한 피해를 유발한다. 현재의 자연재해 경고 시스템은 풍속과 강우를 구분하여 위험을 감지토록 설계되어 강풍과 폭우를 동반한 태풍의 위험을 경고하는데 한계점이 존재한다. 코플라모형은 확률변수들 사이의 복잡한 의존성 구조를 파악하기 위해 단변량분포의 집합을 다변량분포로 연결하는 모형으로 강우, 홍수, 가뭄 등의 분야에서 활발하게 연구되고 있다. 본 연구에서는 한반도에서 태풍에 가장 많이 노출된 도시인 부산과 제주도의 기상 관측소 (ASOS)에서 수집된 1904년 4월 9일부터 2015년 12월 31일까지 일강수량 (precipitation), 일최대풍속 (maximum wind speed) 자료를 이용하였다. 각 변수의 주변부확률을 추정하기 위해 두꺼운 꼬리 분포인 로그정규분포, 감마분포, 와이블분포를 고려하였다. 주변부 확률분포의 적합성검정은 Kolmogorov-Smirnov와 Cramer-von-Mises, Anderson-Darling 검정통계량을 이용하였다. 코플라모형을 위해 순위를 기반으로 한 유사자료 (pseudo observation)를 생성하여 두 변수 간 의존성을 추정하였다. 강풍과 폭우의 의존성을 설명하기 위한 코플라모형으로 타원형, 나선형, 극단치 코플라모형이 고려되었다. 코플라모형의 적합성은 Cramer-von-Mises로 검정하였고, 교차검증을 통해 최적모형을 선택하였다. 연구결과 일강우량과 풍속의 주변부 확률분포로 대부분 로그정규분포가 적합하였다. 부산의 일평균풍속에 따른 일강우량은 t 코플라, 일최대풍속에 따른 일강우량은 Clayton 코플라가 최적모형으로 선정되었다. 제주도의 일최대풍속에 따른 일강우량은 정규코플라, 일강우량에 따른 일평균풍속은 Frank 코플라, 일강우량에 따른 일최대풍속은 Husler-Reiss 코플라가 최적모형으로 선택되었다.

주요용어: 극단치 자료, 일최대강수량, 일최대풍속, 코플라모형, k겹 교차검증.

1. 서론

한반도는 매년 태풍의 위험에 노출되어 있다. 태풍은 강풍과 강우가 동반되는 열대성 저기압으로 사회·경제적 피해를 유발한다. 자연재해 경고 시스템은 풍속과 강우를 구분하여 위험을 감지토록 설계되어 강풍과 폭우를 동반한 태풍의 위험을 경고하는데 한계점이 존재한다. 본 연구에서는 부산과 제주도의 기상 관측소에서 수집된 풍속과 강우자료의 상관성이 고려된 코플라모형으로 자연재해 경고 시스템을 설계하여 강풍과 폭우를 동반하는 자연재해에 대한 위험을 경고하고자 한다.

코플라모형은 확률변수들 사이의 복잡한 의존성 구조를 파악하기 위해 단변량분포의 집합을 다변량분포로 연결시키는 모형으로 Sklar (1959)에 의해 제안되었으며 강우, 홍수, 가뭄 등의 분야에서 적용되고 있다. 기상 및 수문학 목적으로 국내외에서 코플라모형을 이용한 다변량분포에 관한 연구가 활발히 진

[1] (38453) 경상북도 경산시 진량읍 대구대로 201, 대구대학교 일반대학원 통계학과, 석사과정.
[2] 교신저자: (38453) 경상북도 경산시 진량읍 대구대로 201, 대구대학교 전산통계학과, 조교수.
E-mail: statstar@daegu.ac.kr

행되고 있다. 국내에서 Kwak (2016)은 이변량 경시적 자료의 조건부 결합 분포를 추정하기 위하여 회귀 모형과 코플라모형을 연구하였다. Choi 등 (2013)은 다양한 강수보험의 활성화에 필요한 강수횟수와 강수량을 확률분포에 적합하고 두 확률변수의 상관성을 코플라를 이용하여 분석하였다. 방재 및 가뭄에 대해 Park 등 (2015)은 강우와 바람사이의 상호의존성을 고려한 코플라모형 활용 가능성을 검토하였다. Joo 등 (2012)은 주변분포형과 코플라모형의 다양한 조합에 따른 강우량과 지속시간에 대한 이변량 빈도해석을 하였다. So 등 (2014)은 Clayton 코플라를 이용하여 이변량 결합가뭄지수를 산정하고 국내 활용성을 평가하였다. Kwak 등 (2012)은 코플라 이론을 이용하여 한강 상류유역과 남한강 상류유역을 대상으로 수문학적 가뭄의 결합확률분포를 추정하였다. Yoo 등 (2013)은 가뭄 지속기간과 심도의 상호관계를 고려한 코플라 함수를 적용하여 이변량 가뭄해석을 수행하였다. Kim 등 (2012)은 우리나라의 가뭄상황을 종합적으로 판단하고자 코플라 기반의 결합가뭄지수를 산정하여 활용하였다.

국외에서도 코플라 이론을 이용한 연구가 활발히 진행되고 있다. Kao 와 Govindaraju (2007)은 강우강도, 지속시간 변수 사이의 상관관계를 구현하기 위해 코플라모형을 분석하였다. Genest 와 Favre (2007)은 코플라모형이 주변분포의 특성을 유지하면서 이변량분포를 구축할 수 있는 이점이 있다고 밝혔다. Zhang 와 Singh (2006)은 코플라 방법을 이용하였을 경우에 이변량 빈도 해석의 한계를 완화할 수 있음을 보였다. Schozel 와 Frienderich (2008)은 일별강수량과 온도간의 코플라모형에 대한 적합성테스트의 문제점을 파악하고 보완하였다. Durante 와 Salvadori (2010)은 코플라 이론을 통해 적절한 수의 매개변수를 갖는 다변량 극한값에 대한 분석을 수행하고, 극한현상에 대한 재현기간을 평가하였다. Salvadori 와 Michele (2010)은 코플라 이론을 통해 새로운 다변량 극한분포를 쉽게 구축하는 방법을 제시하였다. Favre 등 (2004)은 코플라를 이용하여 캐나다 유역의 최대흐름에 대한 두 가지 어플리케이션을 구상하여 다변량분석을 수행하였다. Renard 와 Lang (2007)은 매우 간단한 코플라인 가우시안 코플라를 이용하여 중요성 결정, 지역 위험 분석, 방전주파수 모델로 수문곡선 설계 및 지역 주파수 분석 사용의 유용성을 입증하였다. Requena 등 (2013)은 이변량 코플라모형을 사용하여 홍수 최고점과 부피의 이항분포를 얻어 분석을 진행하였다.

본 논문에서는 한반도에서 태풍에 가장 많이 노출된 도시인 부산과 제주도의 기상관측소 (ASOS)에서 수집된 1904년 4월 9일부터 2015년 12월 31일까지 총 112년간의 일강수량 (precipitation), 일최대풍속 (maximum wind speed), 일평균풍속 (average wind speed)자료를 이용하였다. 본 연구의 관심은 악기상으로 연간 최댓값을 추출하여 접근하였다. 일최대강수량이 발생한 날과 일최대풍속이 가장 높은 날, 일평균풍속이 가장 높은 날이 동일하지 않으므로 각각에 대해 매칭되는 값을 모델로 설정하였다. 또한, 코플라모형은 상관관계가 중요하므로 각 모델의 두 변수간 상관성이 가장 높은 지점을 문턱치로 지정하였다. 두 변수간의 상관관계가 입증되지 않은 모델은 제외하였다. 또, 각 변수들이 어떤 분포를 따르는지 확인하기 위해 두꺼운 꼬리 분포인 로그정규분포, 감마분포, 와이블분포에 적합 시켰으며, 코플라 이론을 이용하여 설정한 각 모델이 어떤 코플라 분포를 따르는지 확인하였다. 적합도 검정은 Kolmogorov-Smirnov (K-S)와 Cramer-von-Mises (CvM), Anderson-Darling (A-D)을 이용하였으며 코플라모형의 최적 적합은 leave-one-out cross validation을 이용하였다. 마지막으로 각 모델이 적합된 코플라와 경험적 코플라를 그래프를 통해 비교하였다. 이 연구를 통해 강풍과 강수 각각의 자연재해 위험성 예측이 아닌 코플라모형을 통한 두 변수의 상관성이 고려된 위험성 예측을 할 수 있다. 이를 이용해 보다 나은 자연재해 경고 시스템을 수립하여 각각이 위험하다고 판단되는 수치보다 낮은 수치라도 복합적인 피해를 고려해서 위험하다고 판단하여 피해를 줄일 수 있다고 본다.

2. 코플라 이론

2.1. 코플라 함수의 정의

코플라 함수는 여러 확률변수들 사이의 복잡한 구조를 파악하기 위한 방안으로 Sklar (1959)에 의해 제시되었다. 코플라 함수는 수학적으로 두 개 이상의 확률변수의 결합확률분포 (joint probability distribution)를 유도하기 어려운 경우에 두 확률변수의 주변확률분포 (marginal probability distribution)로부터 얻은 유사자료 (pseudo observation)로부터 함수를 추정한다. 즉, 코플라 함수는 단변량분포의 집합을 다변량분포로 연결시키는 함수를 말하며, 0에서 1까지의 n개의 균등분포 (uniform distribution)들의 결합분포함수로 정의된다.

$$C : [0,1]^n \to [0,1].$$

여기서 C는 코플라 함수를 의미한다. 모든 확률분포는 역계산법 (inverse transforming method)에 의해 균등분포로 변환 가능하므로, n차원의 균등분포의 결합확률분포로 다변량 결합확률분포를 유사하게 추정할 수 있다.

2.2. Sklar의 정리

Sklar 정리 (1959)는 다변량분포함수의 각각의 주변확률분포을 이용하여 다변량결합확률분포를 제시하였다. 코플라 함수는 단변량분포함수인 주변확률분포와 다변량분포함수인 결합확률분포를 연결하는 역할을 하며 변수간의 의존구조에 관한 정보를 담고 있다. Sklar 정리에 따르면 확률분포함수 F가 $F_1, \cdots, F_n$을 가지는 n차원의 함수라면, 실수 공간의 모든 랜덤 벡터 $\mathbf{x}$에 대해 n차 코플라 함수 C가 존재한다.

$$F(x_1, \cdots, x_n) = C(F_1(x_1), \cdots, F_n(x_n)).$$

만약 $F_1, \cdots, F_n$이 연속인 경우에는 코플라함수 C가 유일하게 존재한다. 또한 코플라함수 C는 역계산법에 의해 다음과 같이 주어진다.

$$C(u_1, \cdots, u_n) = C(F_1^{-1}(u_1), \cdots, F_n^{-1}(u_n)),$$

여기서 $u_i \in [0,1]$이고, F_1^{-1}는 F_i의 역함수이다.

2.3. 타원형 코플라 함수 (Elliptical copula function)

타원형 코플라 함수는 타원형분포 (elliptical distribution)와 관련된 함수로 대표적으로 정규코플라 함수와 t 코플라 함수가 있다. 정규코플라 함수는 평균벡터와 공분산행렬 $\sum$ 로 구성되고, t 코플라는 평균벡터와 상관행렬 R 그리고 자유도 v로 구성되어진다 (Table 2.1).

Table 2.1 Ellipticla copula functions

Copula	Function
Student-t	$\int_{-\infty}^{t_v^{-1}(u_1)} \cdots \int_{-\infty}^{t_v^{-1}(u_n)} \frac{\gamma(\frac{v+n}{2})\lvert R\rvert^{-1/2}}{\gamma(\frac{v+n}{2})\lvert v\prod\rvert^{n/2}}(1+\frac{1}{v}\underline{x}^T R^{-1}\underline{x})^{-\frac{v+n}{2}} dx_1 \cdots dx_n$
Normal	$\frac{1}{\lvert\sum\rvert^{1/2}} exp(-\frac{1}{2}\underline{x}^T(\sum^{-1}-I)\underline{x})$

2.4. 나선형 코플라 함수 (Archimedean copula function)

나선형 코플라 함수는 함수의 모수가 하나로 구성되어 수학적으로 다루기 쉽고 간단하다는 장점이 있다. 나선형 코플라 함수군에 속하는 코플라 함수는 Clayton 코플라, Frank 코플라, Joe 코플라 등이 있다. Clayton 코플라 함수는 왼쪽 꼬리의 강한 종속성을 잘 설명하고 Frank 코플라 함수는 자료간의 상관계수가 양 (+) 및 음 (-) 모두 추정이 가능한 코플라 함수이다 (Nelsen, 2006) (Table 2.2).

Table 2.2 Archimedean copula functions

Copula	Function
Clayton	$[max(u^{-\theta}+v^{-\theta}-1;0)]^{-1/\theta}$
Frank	$-\frac{1}{\theta}log(1+\frac{(exp(-\theta u)-1)(exp(-\theta v)-1)}{exp(-\theta)-1})$
Joe	$1-[(1-u)^{\theta}+(1-v)^{\theta}-(1-u)^{\theta}(1-v)^{\theta}]^{1/\theta}$

2.5. 극단치 코플라 함수 (Extreme copula function)

극단치 코플라 함수는 서로 독립적이며 동일한 분포를 따르는 표본의 극단치의 한계를 보완하는 함수이다. 이변량 표본이 $(X_i, Y_i), i = 1, \cdots, n$로 나타나고 (X_i, Y_i)는 서로 독립적이라고 가정하면 표본에서 추출한 최댓값은

$$M_n = max(X_1, \cdots, X_n),$$
$$N_n = max(Y_1, \cdots, Y_n)$$

이다. 이변량결합확률분포함수 (H)의 주변확률분포가 F와 G를 가진다고 가정하면 M_n와 N_n의 확률분포함수는

$$Pr[M_n \le x] = F^n(x), Pr[N_n \le y] = G^n(y)$$

이다. 이변량결합확률분포함수는 다음식과 같이 표현된다.

$$Pr[M_n \le x, N_n \le y] = H^n(x, y).$$

Sklar 정리에 의해 분포함수 $H = H^1$는 코플라 C^1을 가진다. 즉, $H^1(x, y) = C_1(F(x), G(y))$가 성립된다. 따라서 다음 식도 성립이 된다.

$$H^n(x, y) = C_1^n(F^n(x)^{1/n}, G^n(x)^{1/n}).$$

H^n의 코플라 C_n은 $C_n(u, v) = C_1^n(u^{1/n}, v^{1/n})$이다. 극단치 코플라 함수로는 Gumbel 코플라 함수, Galambos 코플라 함수, Husler-Reiss 코플라 함수가 있다. 극단치 코플라 함수는 양의 의존성만 표시할 수 있다 (Table 2.3).

Table 2.3 Extreme copula functions

Copula	Function
Gumbel	$exp(-((-log(u))^{\theta}+(-log(v))^{\theta})^{1/\theta})$
Galambos	$uvexp([(-logu)^{-\delta}+(-logv)^{-\delta}]^{-1/\delta})$
Husler-Reiss	$exp((logu)\phi[\frac{1}{\delta}+\frac{1}{2}\delta log\frac{-logu}{-logv}]+(logv)\phi[\frac{1}{\delta}+\frac{1}{2}\delta log\frac{-logv}{-logu}])$

2.6. 모수추정법

코플라 함수의 모수추정법으로는 모수적 추정법과 비모수적 추정법이 있다. 모수적 추정법은 최대우도함수를 기반으로 한 최대우도법 (maximum likelihood method, MLE) 방법이 대표적이다. 최대우도법은 코플라 함수의 모수와 주변확률분포함수의 모수를 한꺼번에 추정한다. 그 외의 모수적 추정법으로 주변확률분포의 모수와 코플라 함수의 모수를 분리하여 두 단계에 걸쳐서 추정하는 inference function for margins (IFM) 방법이 있다 (Joe와 Xu, 1996). 비모수적 방법으로는 각각의 주변확률분포함수에 대해 모수적 분포함수를 가정하지 않고 경험적 분포함수로 추정한 후 코플라 함수의 모수를 추정하는 canonical maximum likelihood (CML) 방법이 있다 (Cherubini 등, 2004). 본 연구에서 고려된 모수추정법은 비교적 쉽고 널리 쓰이는 유사우도방법 (maximum pseudo likelihood)으로 CML 방법과 유사하지만 경험적 분포함수를 이용하지 않고 주변확률분포의 순위 (rank)로부터 재구성된 유사자료 (pseudo observation)를 이용하여 최대우도법으로 코플라 함수의 모수를 추정한다 (Shih와 Louis, 1995; Genest 등, 1995).

$$\hat{\theta} = argmaxL(\theta) = argmax \sum_{t=1}^{T} lnc(\hat{u}_1^t, \cdots, \hat{u}_n^t),$$

여기서 $u_i = Rank(x_i)/(T+1), T$는 관측치 수이다.

모수추정은 R프로그램의 "copula" 패키지의 fitCopula 함수를 이용하여 유사우도방법으로 추정하였다 (Hofert 등, 2015).

2.7. 모형의 적합도 검정

적합도 검정 (goodness-of-fit test)은 대상 자료로부터 얻어지는 경험적 분포와 가정한 확률분포가 얼마나 잘 일치하는가를 검정한다. 모집단의 확률분포형을 알지 못하면 기존의 확률분포형으로 모집단의 성질을 정확히 나타내기 어려우므로 다양한 기법을 통해서 많은 정보를 이용하여 분포형을 선정하는 것이 합리적이라 할 수 있다 (Shin 등, 2010). 본 연구에서는 K-S검정과 CvM검정, A-D검정을 이용하였다.

K-S검정은 표본자료의 누적분포함수를 비교하여 편차가 제일 큰 값으로 판정하는 방법으로 표본의 누적확률분포를 구하기 위하여 먼저 n개의 자료를 크기순으로 배열하여 다음 식과 같이 자료 값의 누적확률을 산정한다 (Smirnov, 1939).

$$S_n(x) = \frac{k}{n}, x_k \le x < x_{k+1},$$

여기서 $S_n(x)$는 자료를 크기순으로 배열했을 때, k번째 자료 값의 이론적 누적발생확률, n은 자료 수이다. 두 분포함수의 최대편차는 $D_n = max|F(x) - S_n(x)|$로 계산한다. 최대편차 D_n은 n의 크기에 따라 좌우되는 확률변수로 유의수준 α에 따라 적합성이 결정된다.

CvM검정은 표본자료의 누적분포함수 $F(x_i)$로 정의된 확률분포형을 모집단으로 갖는다는 가정을 검정하며, $\hat{\theta}$은 표본자료의 크기가 n인 자료에서 추정된 매개변수 집단이다. 검정통계량 W는 다음과 같이 계산된다 (Thompson, 1966).

$$W = \frac{1}{12n} + \sum_{t=1}^{n} \left[\frac{2i-1}{2n} - F(x_i) \right]^2,$$

여기서 $F(x_i)$는 i번째 순위통계량의 누적분포함수치이다. 또한 $W \le W_{i-\alpha}(n)$을 만족한다면, 적용한 분포형을 유의수준 α에서 기각할 수 없다.

A-D검정은 경험적 분포함수와 가설 분포의 수직 차이를 바탕으로 EDF (empirical distribution function)통계량을 계산한다 (Anderson와 Darling, 1952).

$$n\int_{-\infty}^{\infty}(F_n(x)-F(x))^2\varpi(x)dF(x),$$

여기서 $\varpi(x)$는 가중치 함수이고 가중치 함수 $\varpi(x)=1$일 경우 위의 식은 CvM검정의 검정통계량과 같다. $\varpi(x)=[F(x)(1-F(x))]^{-1}$이면 위의 식은 $A^2=n\int_{-\infty}^{\infty}(F_n(x)-F(x))^2/F(x)(1-F(x))dF(x)$으로 표시된다. 여기서 $F(x)$는 추정된 매개변수를 포함하는 누적분포함수, $F_n(x)$는 경험적 분포함수를 나타내며 이때 A^2을 A-D검정통계량이라 한다.

적합도 검정은 R 프로그램의 "copula" 패키지의 gofCopula 함수를 이용하였다 (Hofert 등, 2015).

2.8. 최적 코플라 함수 선정방법

소표본 데이터일 때 분류기 성능측정의 통계적 신뢰도를 높이기 위해서 일반적으로 재샘플링 (resampling)기법을 사용하는데 대표적인 방법으로 k겹 교차검증 (k-fold cross validation)과 붓스트랩 (bootstrap)이 있다. k겹 교차검증은 수집된 샘플을 k개의 부표본 (subsample)로 나눈다. 이 k개의 부표본 중에서 하나를 모델의 테스트를 위한 검증데이터로 두고 남은 k-1개의 부표본들을 트레이닝 데이터로 사용한다. 모든 부표본들이 검증데이터로 정확히 한 번씩 사용될 때까지 k번 반복되고, 결과들을 평균하여 단일 평가 값을 구한다. 본 연구에서는 k겹교차검증을 기반으로 하여 이를 극단으로 가져가면 k를 데이터 관측치 수 n만큼 하는 leave-one-out cross validation을 이용하였다.

최적 코플라 함수 선정은 R 프로그램의 "copula" 패키지의 xvCopula 함수를 이용하였다 (Hofert 등, 2015).

3. 연구자료

본 연구에서는 위도 30도와 60도 사이의 중위도 지역에서 서쪽에서 동쪽으로 부는 탁월풍인 편서풍의 영향으로 매년 태풍의 위험에 노출된 부산과 제주도의 기상 관측소 (ASOS)에서 수집된 자료를 이용하여 분석 하였다. 1904년 4월 9일부터 2015년 12월 31일까지 총 112년간 수집된 일강수량 (precipitation), 일최대풍속 (maximum wind speed), 일평균풍속 (average wind speed)을 이용하였다. 본 연구의 관심은 악기상으로 연간 최댓값을 추출하여 접근하였다. 일최대강수량이 발생한 날과 일최대풍속이 가장 높은 날, 일평균풍속이 가장 높은 날이 동일하지 않으므로 각각에 대해 매칭되는 값을 정리하면 Table 3.1이다.

코플라 함수는 두 변수 간 상관성이 존재해야하므로 유사자료의 상관성이 존재하지 않은 모델은 분석에서 제외하였다. 극단치 코플라 함수는 상관성이 양수일 때 모수를 추정할 수 있다. 바람세기와 강수량 사이에는 음의 상관성이 존재하여 생존함수인 $S(t)=P[T>t]=1-F(t)$를 이용하였다. $S(t_1,\cdots,t_d)=\bar{C}(S_1(t_1),\cdots,S_d(t_d))$에 대해 $\bar{C}$는 코플라로 이변량의 주변확률분포 사이에 음의 상관성이 존재할 경우 사용된다.

모델1_1은 부산의 일평균풍속에서 추출한 최댓값과 이에 매칭되는 일강수량으로 구성하였으며 상관성이 나타났다. 모델1_2는 부산의 일최대풍속에서 추출한 최댓값과 이에 매칭되는 일강수량으로 구성하였으며 상관성이 보였다. 모델1_3의 경우 부산의 일강수량에서 추출한 최댓값과 이에 매칭되는 일평균풍속으로 구성하였으며 상관성이 보이지 않아 분석에서 제외하였다. 모델1_4 또한 부산의 일강수량에서 추출한 최댓값과 이에 매칭되는 일최대풍속으로 구성하였으나 상관성이 나타나지 않아 분석에서 제외하였다. 모델2_1은 제주도의 일평균풍속에서 추출한 최댓값과 이에 매칭되는 일강수량으로 구성하였

Table 3.1 Extreme variables and paired variables

Region	Model	Extreme variable (ref.)	Paired variable	Kendall's correlation
Busan	Model1_1	Average wind speed	Precipitation	O
	Model1_2	Maximum wind speed	Precipitation	O
	Model1_3	Precipitation	Average wind speed	X
	Model1_4	Precipitation	Maximum wind speed	X
Jeju	Model2_1	Average wind speed	Precipitation	X
	Model2_2	Maximum wind speed	Precipitation	O
	Model2_3	Precipitation	Average wind speed	O
	Model2_4	Precipitation	Maximum wind speed	O

으며 상관성이 보이지 않아 분석에서 제외하였다. 모델2_2의 경우 제주도의 일최대풍속에서 추출한 최댓값과 이에 매칭되는 일강수량으로 구성하였으며 상관성이 나타났다. 모델2_3은 제주도의 일강수량에서 추출한 최댓값과 이에 매칭되는 일평균풍속으로 구성하였으며 상관성이 보였다. 모델2_4는 제주도의 일강수량에서 추출한 최댓값과 이에 매칭되는 일최대풍속으로 구성하였으며 상관성이 나타났다. 문턱치 (threshold)를 변화해 가며 두 변수간의 상관성이 가장 높은 지점을 선택하였다. 문턱치는 2.5씩 증가시켰으며, 각 모델의 paired variable에 적용하였다.

Table 3.2 Kendall correlation coefficient by threshold

Threshold	Model1_1		Model1_2		Model2_2		Model2_3		Model2_4	
	Kendall	p-value	Kendall	p-value	Kendall	p-value	Kendall	p-value	Kendall	p-value
0.0	0.075	(0.323)	0.011	(0.884)	0.160	(0.025)	0.213	(0.003)	0.209	(0.003)
2.5	0.118	(0.210)	0.036	(0.684)	0.170	(0.039)	**0.249**	**(0.001)**	0.209	(0.003)
5.0	0.092	(0.378)	0.030	(0.740)	0.195	(0.037)	0.163	(0.093)	**0.221**	**(0.002)**
7.5	0.219	(0.043)	0.000	(1.000)	**0.248**	**(0.018)**	0.050	(0.696)	0.199	(0.010)
10.0	0.265	(0.024)	0.000	(1.000)	0.219	(0.043)	0.267	(0.137)	0.259	(0.006)
12.5	0.231	(0.056)	0.018	(0.846)	0.179	(0.132)	-0.405	(0.106)	0.315	(0.007)
15.0	0.275	(0.031)	0.049	(0.616)	0.164	(0.205)	0.000	(1.000)	0.182	(0.153)
17.5	0.308	(0.020)	0.055	(0.588)	0.164	(0.205)	-0.333	(1.000)	0.283	(0.060)
20.0	0.379	(0.005)	-0.006	(0.958)	0.212	(0.107)			-0.066	(0.742)
22.5	**0.434**	**(0.002)**	0.047	(0.669)	0.196	(0.144)			0.000	(1.000)
25.0	0.434	(0.002)	0.080	(0.470)	0.196	(0.144)	.	.	0.333	(0.750)
27.5	0.357	(0.013)	0.096	(0.413)	0.196	(0.144)	.	.	0.333	(0.750)
30.0	0.322	(0.029)	0.127	(0.287)	0.196	(0.144)	.	.	0.333	(0.750)
32.5	0.428	(0.005)	0.142	(0.245)	0.196	(0.144)	.	.	1.000	(0.333)
35.0	0.400	(0.012)	0.166	(0.183)	0.196	(0.144)	.	.	.	
37.5	0.400	(0.012)	0.166	(0.183)	0.151	(0.269)	.	.	.	
40.0	0.239	(0.171)	0.228	(0.077)	0.151	(0.269)	.	.	.	
42.5	0.277	(0.126)	0.228	(0.077)	0.151	(0.269)	.	.	.	
45.0	0.213	(0.258)	0.262	(0.050)	0.151	(0.269)	.	.	.	
47.5	0.213	(0.275)	0.197	(0.158)	0.254	(0.076)	.	.	.	
50.0	0.248	(0.244)	**0.356**	**(0.017)**	0.254	(0.076)	.	.	.	

상관성이 확인되지 않은 모델1_3, 모델1_4, 모델2_1을 제외한 나머지 다섯 개의 모델에 대하여 분석을 진행하였다.

Tabel 3.2를 보면 모델1_1에 대하여 문턱치가 22.5와 25일 때 가장 높은 상관성이 측정되는 것을 확인할 수 있다. 두 경우의 값이 같으므로 모델1_1의 최적의 문턱치로 22.5를 선택하였다. 모델1_2에 대하여 문턱치가 50일 때 가장 높은 상관성이 측정되는 것을 확인할 수 있어 최적의 문턱치로 50을 선택하였다. 모델2_2에 대하여 문턱치가 7.5일 때 가장 높은 상관성이 측정되는 것을 확인할 수 있으므로 최적의 문턱치는 7.5가 선택되었다. 또한 모델2_3에 대하여 문턱치가 2.5일 때 가장 높은 상관성이 측정되는 것을 확인할 수 있어 최적의 문턱치로 2.5가 선택되었다. 마지막으로 모델2_4에 대하여 문턱치가 5일 때 가장 높은 상관성이 측정되는 것을 확인할 수 있어 최적의 문턱치로 5가 선택되었다.

다음으로 최적의 문턱치를 적용하여 선택된 데이터에 대한 각 변수가 어떤 분포를 따르는지 확인하였으며 몇몇의 두꺼운 꼬리 분포가 고려되었다.

Table 3.3 The result of goodness-of-fit for marginal distribution

Model	Variable	Marginal distribution	Kolmogorov-Smirnov	Cramer-von-Mises	Anderson-Darling
Model1_1	Average wind speed*	Log-Normal	**0.109**	**0.040**	**0.302**
		Gamma	0.108	0.041	0.324
		Weibull	0.155	0.108	0.834
	Precipitation	Log-Normal	**0.124**	**0.070**	**0.466**
		Gamma	0.150	0.118	0.735
		Weibull	0.183	0.192	1.118
Model1_2	Maximum wind speed*	Log-Normal	**0.082**	**0.030**	**0.253**
		Gamma	0.091	0.030	0.252
		Weibull	0.131	0.060	0.441
	Precipitation	Log-Normal	**0.199**	**0.170**	**1.048**
		Gamma	0.203	0.179	1.093
		Weibull	0.226	0.242	1.439
Model2_2	Maximum wind speed*	Log-Normal	**0.087**	**0.073**	**0.477**
		Gamma	0.106	0.093	0.592
		Weibull	0.152	0.209	1.250
	Precipitation	Log-Normal	0.181	0.372	2.293
		Gamma	0.213	0.348	2.488
		Weibull	**0.180**	**0.263**	**1.734**
Model2_3	Precipitation*	Log-Normal	**0.066**	**0.064**	**0.495**
		Gamma	0.049	0.029	0.299
		Weibull	0.075	0.053	0.549
	Average wind speed	Log-Normal	**0.074**	**0.089**	**0.561**
		Gamma	0.084	0.115	0.843
		Weibull	0.103	0.218	1.423
Model2_4	Precipitation*	Log-Normal	**0.061**	**0.083**	**0.610**
		Gamma	0.051	0.036	0.346
		Weibull	0.074	0.048	0.550
	Maximum wind speed	Log-Normal	**0.088**	**0.109**	**0.614**
		Gamma	0.107	0.156	1.068
		Weibull	0.136	0.295	1.908

*:Extreme variable

Tabel 3.3을 보면 두꺼운 꼬리 분포인 로그정규분포, 감마분포, 와이블분포에 적합시켰을때 적합성검정인 K-S, CvM, A-D에 대한 통계량을 확인하여 3가지 분포 중 어떤 분포에 가장 적합한지 알 수 있다. 모델1_1에 대하여 일평균풍속의 분포를 확인하면 로그정규분포에 가장 적합한 것을 볼 수 있고, 일강수량의 분포도 로그정규분포에 접합한 것을 볼 수 있다. 모델1_2의 일최대풍속 분포를 확인하면 로그정규분포를 따르는 것을 볼 수 있고, 일강수량의 분포도 로그정규분포를 따르는 것을 볼 수 있다. 모델2_2의 경우 일최대풍속의 분포를 확인하면 로그정규분포에 가장 적합한 것을 볼 수 있고, 일강수량의 분포는 와이블분포에 적합한 것을 볼 수 있다. 모델2_3을 보면 일강수량의 분포를 확인했을 때 로그정규분포에 가장 적합한 것을 볼 수 있고, 일평균풍속의 분포도 로그정규분포를 따르는 것을 볼 수 있다. 마지막으로 모델2_4에 대하여 일강수량의 분포를 확인하면 로그정규분포에 가장 적합한 것을 볼 수 있고, 일최대풍속의 분포도 로그정규분포에 적합한 것을 볼 수 있다.

4. 연구결과

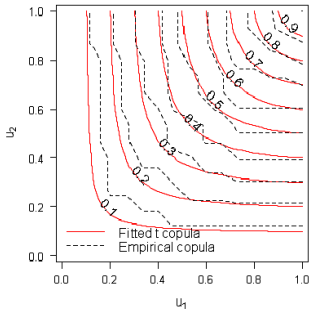
코플라모형은 총 8가지를 고려하였으며 모델을 코플라모형에 적합 시켰을 때 적합도 검정 결과와 leave-one-out cross validation의 값을 통하여 최적의 코플라모형 적합을 확인할 수 있다. Table 4.1을 보면 모델1_1은 적합도 검정 결과 Clayton와 Joe 코플라를 제외한 나머지 6가지 코플라에 대하여 적합한 것을 확인할 수 있으며, 이 중 leave-one-out cross validation의 값이 가장 큰 t 코플라가 최적의 코플라모형 적합인 것으로 나타났다. 모델1_2는 적합도 검정 결과 Clayton 코플라에 대해서만 적합한 것을 확인할 수 있어 Clayton 코플라가 최적의 코플라모형 적합인 것으로 나타났다. 모델2_2는 적합도 검

Table 4.1 The selection of best copula model (CvM)

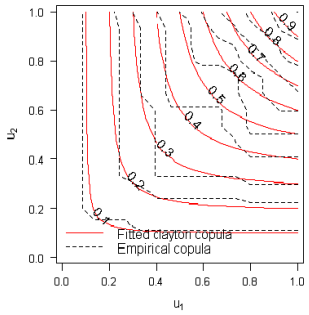
Model	Copula	Goodness-of-fit		Leave-one-out cross validation
		Statistic	p-value	
Model1_1	t	**0.028**	**(0.175)**	**4.121**
	Normal	0.031	(0.176)	3.106
	Gumbel	0.026	(0.223)	4.078
	Frank	0.036	(0.063)	3.789
	Clayton	0.064	(0.012)	2.334
	Joe	0.039	(0.048)	3.169
	Galambos	0.026	(0.244)	3.811
	Husler-Reiss	0.028	(0.192)	2.995
Model1_2	t	0.041	(0.047)	1.282
	Normal	0.040	(0.038)	0.741
	Gumbel	0.073	(0.002)	-0.316
	Frank	0.050	(0.022)	1.478
	Clayton	**0.022**	**(0.317)**	**4.285**
	Joe	0.132	(0.001)	-1.172
	Galambos	0.069	(0.003)	-0.257
	Husler-Reiss	0.067	(0.001)	-0.745
Model2_2	t	0.037	(0.092)	1.959
	Normal	**0.030**	**(0.194)**	**2.619**
	Gumbel	0.041	(0.052)	1.469
	Frank	0.041	(0.037)	1.924
	Clayton	0.055	(0.013)	1.252
	Joe	0.056	(0.012)	0.664
	Galambos	0.036	(0.110)	2.025
	Husler-Reiss	0.031	(0.153)	2.302
Model2_3	t	0.035	(0.080)	2.962
	Normal	0.029	(0.146)	4.546
	Gumbel	0.049	(0.029)	1.952
	Frank	**0.029**	**(0.113)**	**4.924**
	Clayton	0.057	(0.005)	3.403
	Joe	0.081	(0.002)	0.290
	Galambos	0.046	(0.032)	2.304
	Husler-Reiss	0.043	(0.041)	2.386
Model2_4	t	0.037	(0.043)	3.608
	Normal	0.034	(0.080)	4.477
	Gumbel	0.031	(0.140)	4.200
	Frank	0.039	(0.030)	3.933
	Clayton	0.083	(0.000)	0.908
	Joe	0.037	(0.110)	3.861
	Galambos	0.028	(0.183)	4.958
	Husler-Reiss	**0.026**	**(0.239)**	**5.340**

정 결과 Frank와 Clayton, Joe 코플라를 제외한 나머지 5가지 코플라에 대하여 적합한 것을 확인할 수 있으며, 이 중 leave-one-out cross validation의 값이 가장 큰 정규코플라가 최적의 코플라모형 적합인 것으로 나타났다. 모델2_3은 적합도 검정 결과 t, Frank, 정규코플라에 대해서 적합한 것을 확인할 수 있으며, 이 중 leave-one-out cross validation의 값이 가장 큰 Frank 코플라가 최적의 코플라모형 적합인 것으로 나타났다. 모델2_4는 적합도 검정 결과 t와 Frank, Clayton 코플라를 제외한 나머지 5가지 코플라에 대하여 적합한 것을 확인할 수 있으며, 이 중 leave-one-out cross validation의 값이 가장 큰 Husler-Reiss 코플라가 최적의 코플라모형 적합인 것으로 나타났다.

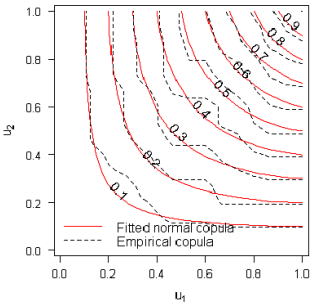
Figure 4.1을 통해 적합된 코플라와 경험적 코플라가 유사함을 확인할 수 있다. Figure 4.2를 통해 두 변수의 상관성이 고려된 실제 관측값의 위험성을 예측할 수 있다. 모델1_1의 경우 기존 위험수준 예측



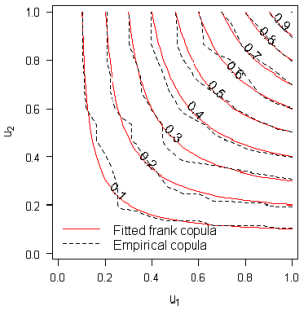
(a) model1_1's t copula



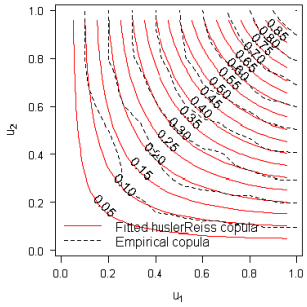
(b) model1_2's clayton copula



(c) model2_2's normal copula



(d) model2_3's frank copula



(e) model2_4's husler-reiss copula

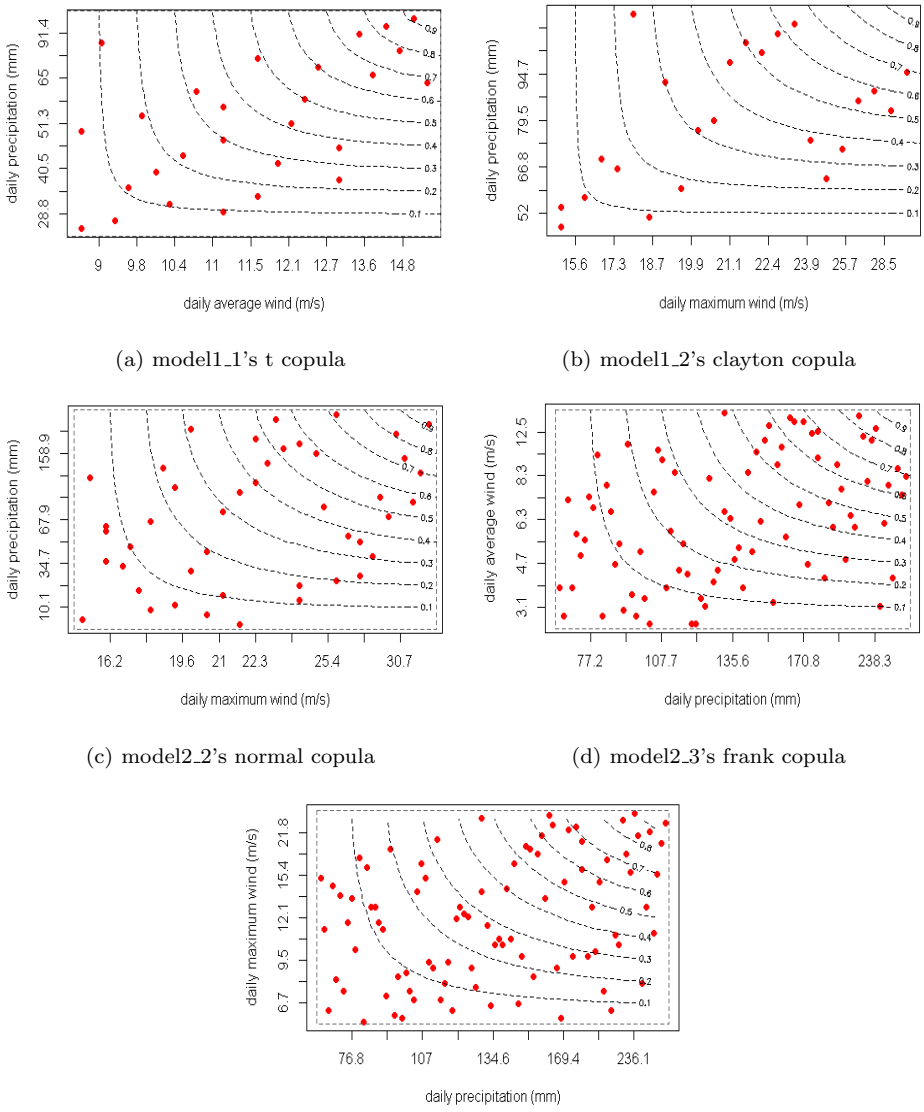
Figure 4.1 Comparison of empirical copula and fitted copula

을 만약 일평균풍속이 14.8을 넘으면 위험수준이라고 판단하거나, 일강수량이 91.4 이상이면 위험하다고 판단한다면 이보다 작은 수치라도 두 변수의 상관성을 고려한 포물선을 넘는다면 위험수준이라고 예측할 수 있다. 이처럼 각 변수의 값이 기존 위험수준에 미치지 못하더라도 코플라 이론을 통한 두 변수의 상관성이 고려된 포물선을 넘는다면 위험하다고 예측할 수 있다.

모델1_1의 경우 일평균풍속과 일강수량 모두 로그정규분포를 따르는 것으로 나타났으며 두 변수를 코플라모형에 적합했을 경우 t 코플라에 적합한 것을 볼 수 있었다. 모델1_2의 경우 일최대풍속과 일강수량 모두 로그정규분포를 따르는 것으로 나타났으며 두 변수를 코플라모형에 적합했을 경우 Clayton 코플라에 적합한 것을 볼 수 있었다. 모델2_2의 경우 일최대풍속은 로그정류분포에 일강수량은 와이블분포를 따르는 것으로 나타났으며 두 변수를 코플라모형에 적합했을 경우 정규코플라에 적합한 것을 볼 수 있었다. 모델2_3의 경우 일강수량과 일평균풍속 모두 로그정규분포를 따르는 것으로 나타났으며 두 변수를 코플라모형에 적합했을 경우 Frank 코플라에 적합한 것을 볼 수 있었다. 모델2_4의 경우 일강수량과 일최대풍속 모두 로그정규분포를 따르는 것으로 나타났으며 두 변수를 코플라모형에 적합했을 경우 HuslerReiss 코플라에 적합한 것을 볼 수 있었다.

5. 결론 및 향후연구

부산지역에서 연도별 평균풍속의 최댓값일 때 일강수량 데이터를 이용한 강풍과 강수의 코플라모형은



(a) model1_1's t copula

(b) model1_2's clayton copula

(c) model2_2's normal copula

(d) model2_3's frank copula

(e) model2_4's husler-reiss copula

Figure 4.2 Real data with fitted copula

t 코플라에 적합하였고, 연도별 최대풍속의 최댓값일 때 일강수량 데이터를 이용한 강풍과 강수의 코플라모형은 Clayton 코플라에 적합하였다. 제주도지역에서 연도별 최대풍속의 최댓값일 때 일강수량 데이터를 이용한 강풍과 강수의 코플라모형은 정규코플라에 적합하였고, 연도별 일강수량의 최댓값일 때 평균풍속 데이터를 이용한 강풍과 강수의 코플라모형은 Frank 코플라에 적합하였으며, 연도별 일강수량의 최댓값일 때 최대풍속 데이터를 이용한 강풍과 강수의 코플라모형은 Husler-Reiss 코플라에 적합

Table 4.2 The result of copula model

Region	Model	Variable	Marginal distribution	Copula	Estimate	s.e
Busan	Model1_1	Average wind speed*	Log-Normal	t	0.670	(0.100)
		Precipitation	Log-Normal			
	Model1_2	Maximum wind speed*	Log-Normal	Clayton	1.503	(0.882)
		Precipitation	Log-Normal			
Jeju	Model2_2	Maximum wind speed*	Log-Normal	Normal	0.430	(0.179)
		Precipitation	Weibull			
	Model2_3	Precipitation*	Log-Normal	Frank	2.278	(0.833)
		Average wind speed	Log-Normal			
	Model2_4	Precipitation*	Log-Normal	HuslerReiss	0.955	(0.173)
		Maximum wind speed	Log-Normal			

*:Extreme variable

하였다. 이 분석결과를 통해 강풍과 강수 각각의 모형에 대한 자연재해 위험성 예측이 아닌 두 변수의 상관성이 고려된 코플라모형을 통한 자연재해에 대한 위험성 예측으로 보다 나은 자연재해 경고 시스템을 수립할수 있다고 본다. 풍속과 강수량 각각의 위험?수준 판단되는 수치보다 낮은 수치라도 복합적인 피해를 고려해서 위험하다고 판단하여 사회·경제적 피해를 더 줄일 수 있다고 본다. 본 연구에서는 제주도와 부산지역을 연구하였지만, 앞으로 전국 70개 기상관측소를 대상으로 확대하여 분석하고, 이변량 t분포와 비교할 예정이다.

References

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "Goodness of Fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, **23**, 193-212.

Cherubini, U., Luciano, E. and Vecchiato, W. (2004). *Copula methods in finance*, John Wiley & Sons.

Choi, C. H., Lee, H. S. and Ju, H. C. (2013). Analyzing rainfall patterns and pricing rainfall insurance using copula *Journal of the Korean Data & Information Science Society*, **24**, 603-623.

Durante, F. and Salvadori, G. (2010). On the construction of multivariate extreme value models via copulas. *Environmetrics*, **21**, 143-161.

Favre, A. C., Adlouni, S. E., Perreault, L., Thiemonge, N. and Bobee, B. (2004). Multivariate hydrological frequency analysis using copulas. *Water Resources Research*, **40**, W01101, doi:10.1029/2003WR002456.

Genest, C. and Favre, A. C. (2007). Everything you always wanted to know about copula but afraid to ask. *Journal of Hydrology*, **12**, 347-368.

Genest, C., Ghoudi, K. and Rivest, L. P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Printed in Great Britain*, **82**, 543-552.

Hofert, M., Kojadinovic, I., Maechler, M. and Yan, J. (2015). *Copula: multivariate dependence with copulas*, R package version 0.999-14, http:CRAN.R-project.org/package=copula.

Joe, H. and Xu, J. J. (1996). *The estimation method of inference functions for margins for multivariate models*, Department of Statistics, University of British Columbia.

Joo, K. W., Shin, J. Y. and Heo, J. H. (2012). Bivariate frequency analysis of rainfall using copula model. *Journal of Korea Water Resources Association*, **45**, 827-837.

Kao, S. C. and Govindaraju, R. S. (2007). A bivariate frequency analysis of extreme rainfall with implications for design. *Journal of Geophysical Research*, **112**, D13119, doi:10.1029/2007JD008522.

Kim, S.D., Ryu, J. S., Oh, K. R. and Jeong, S. M. (2012). An application of copulas-based joint drought index for determining comprehensive drought conditions. *Journal of Korean Society of Hazard Mitigation*, **12**, 223-230.

Kwak, J. W., Kim, D. G., Lee, J. S. and Kim, H. S. (2012). Hydrological drought analysis using copula theory. *Journal of the Korea Society of Civil Engineers*, **32**, 161-168.

Kwak, M. J. (2016). Estimation of the joint conditional distribution for repeatedly measured bivariate cholesterol data using nonparametric copula. *Journal of the Korean Data & Information Science Society*, **27**, 689-700.

Nelsen, R. B. (2006). *An introduction to copulas*, Springer, New York.

Park, J. B., Kal, B. S. and Heo, J. R. (2015). The study to estimate the fitness of bivariate rainfall frequency analysis considering the interdependence between rainfall and wind speed. *Journal of Korean Society of Hazard Mitigation*, **15**, 103-110.

Renard, B. and Lang, M. (2007). Use of gaussian copula for multivariate extreme value analysis: some case studies in hydrology. *Advances in Water Resources*, **30**, 897-912.

Requena, A. L., Mediero, L. and Garrote, L. (2013). A bivariate return period based on copula for hydrologic dam design: accounting for reservoir routing in risk estimation. *Hydrologic Earth System Sciences*, **17**, 3023-3038.

Salvadori, G. and Friederichs, P. (2010). Multivariate multiparameter extreme value models and return peiods: a copula approach. *Water Resources Research*, **46**, W10501, doi:10.1029/2009WR009040.

Schoelzel, C. and Friederichs, P. (2008). Multivariate non-normally distributed random variables in climateresearch introduction to the copula approach. *Nonlinear processes in Geophysics*, **15**, 761-772.

Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, **51**, 1384-1399.

Shin, H. J., Sung, L. M. and Heo, J. H. (2010). Derivation of modified anderson-darling test statistics and power test for the gumbel distribution. *Journal of Korea Water Resources Association*, **43**, 813-822.

Sklar, A. (1959). Fonctions de reparation rma n dimensions et leurs marges. *Publication de l'institute de statistique de l'universite de Paris*, **8**, 229-231.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin of Mathematical University of Moscow*, **2**, 3-16.

So, J. M., Sohn, K. H. and Bae, D. H. (2014). Estimation and assessment of bivariate joint drought index based on copula functions. *Journal of Korea Water Resources Association*, **47**, 171-182.

Thompson, R. (1966). Bias of the one-sample cramer-von mises test. *Journal of the American Statistical Association*, **61**, 246-247.

Yoo, J. Y., Shin, J. Y., Kim, D. K. and Kim, T. W. (2013). Drought risk analysis using stochastic rainfall generation model and copula functions. *Journal of Korea Water Resources Association*, **46**, 425-437.

Zhang, L. and Singh, W. P. (2006). Bivariate flood frequency analysis using the copula method. *Journal of Hydrologic Engineering*, **11**, 150-164.

Journal of the Korean Data & Information Science Society
2017, **28**(4), 797–810

http://dx.doi.org/10.7465/jkdi.2017.28.4.797
한국데이터정보과학회지

Analysis of extreme wind speed and precipitation using copula

Taeyong Kwon[1] · Sanghoo Yoon[2]

[1]Department of Statistics, Daegu University
[2]Department of Statistics and Computer Science, Daegu University & Institute of Basic Science, Deagu University

Received 2 May 2017, revised 22 June 2017, accepted 29 June 2017

Abstract

The Korean peninsula is exposed to typhoons every year. Typhoons cause huge socioeconomic damage because tropical cyclones tend to occur with strong winds and heavy precipitation. In order to understand the complex dependence structure between strong winds and heavy precipitation, the copula links a set of univariate distributions to a multivariate distribution and has been actively studied in the field of hydrology. In this study, we carried out analysis using data of wind speed and precipitation collected from the weather stations in Busan and Jeju. Log-Normal, Gamma, and Weibull distributions were considered to explain marginal distributions of the copula. Kolmogorov-Smirnov, Cramer-von-Mises, and Anderson-Darling test statistics were employed for testing the goodness-of-fit of marginal distribution. Observed pseudo data were calculated through inverse transformation method for establishing the copula. Elliptical, archimedean, and extreme copula were considered to explain the dependence structure between strong winds and heavy precipitation. In selecting the best copula, we employed the Cramer-von-Mises test and cross-validation. In Busan, precipitation according to average wind speed followed t copula and precipitation just as maximum wind speed adopted Clayton copula. In Jeju, precipitation according to maximum wind speed complied Normal copula and average wind speed as stated in precipitation followed Frank copula and maximum wind speed according to precipitation observed Husler-Reiss copula.

Keywords: Copula, extreme value, k-fold cross validation

[1] Master's course, Department of Statistics, Daegu University, Gyeongbuk 38453, Republic of Korea.
[2] Corresponding author: Assistant professor, Department of Statistics and Computer Science, Daegu University, Gyeongbuk 38453, Republic of Korea. E-mail: statstar@daegu.ac.kr