J. Korea Water Resour. Assoc. Vol. 53, No. 12 (2020), pp. 1203-1210
doi: 10.3741/JKWRA.2020.53.12.1203
pISSN 1226-6280
eISSN 2287-6138

# Drought evaluation using unstructured data: a case study for Boryeong area

Jung, Jinhong[a] · Park, Dong-Hyeok[b] · Ahn, Jaehyun[c]*

[a]Graduate Student, Department of Urban Infrastructure and Disaster Prevention Engineering, Seokyeong University, Seoul, Korea
[b]Director of Research Institute, RAON T&C, Anyang, Korea
[c]Professor, Department. of Civil & Architectural Engineering, Seokyeong University, Seoul, Korea

Paper number: 20-108
Received: 10 November 2020; Revised: 20 November 2020; Accepted: 20 November 2020

## Abstract

Drought is caused by a combination of various hydrological or meteorological factor, so it is difficult to accurately assess drought event, but various drought indices have been developed to interpret them quantitatively. However, the drought indexes currently being used are calculated from the lack of a single variable, which is a problem that does not accurately determine the drought event caused by complex causes. Shortage of a single variable may not be a drought, but it is judged to be a drought. On the other hand, research on developing indices using unstructured data, which is widely used in big data analysis, is being carried out in other fields and proven to be superior. Therefore, in this study, we intend to calculate the drought index by combining unstructured data (news data) with weather and hydrologic information (rainfall and dam inflow) that are being used for the existing drought index, and to evaluate the utilization of drought interpretation through verification of the calculated drought index. The Clayton Copula function was used to calculate the joint drought index, and the parameter estimation was used by the calibration method. The analysis showed that the drought index, which combines unstructured data, properly expresses the drought period compared to the existing drought index (SPI, SDI). In addition, ROC scores were calculated higher than existing drought indices, making them more useful in drought interpretation. The joint drought index calculated in this study is considered highly useful in that it complements the analytical limits of the existing single variable drought index and provides excellent utilization of the drought index using unstructured data.

**Keywords:** Drought, Drought index, Big data, Unstructured data

# 비정형 데이터를 활용한 가뭄평가 - 보령지역을 중심으로 -

정진홍[a] · 박동혁[b] · 안재현[c]*

[a]서경대학교 대학원 도시기반방재안전공학과 석사과정, [b](주)라온티앤씨 연구소장, [c]서경대학교 공과대학 토목건축공학과 교수

## 요 지

가뭄은 다양한 수문학적 또는 기상학적 인자들이 복합적으로 작용하여 발생하기 때문에 가뭄의 사상을 정확히 평가하는 것은 어려운 일이나, 이를 정량적으로 해석하기 위해 다양한 가뭄지수들이 개발되어 왔다. 하지만 현재 활용중인 가뭄지수들은 단일변량의 부족량을 통해 산정되며, 복합적인 원인으로 발생하는 가뭄의 사상을 정확히 판단하지 못하는 문제가 있다. 단순 단일변량의 부족을 가뭄이라고 판단하기는 어렵기 때문이다. 최근에는 빅데이터 분석에서 많이 활용되고 있는 비정형 데이터를 활용하여 지수를 개발하는 연구들이 타 분야에서 진행되고 있으며 우수성이 입증되고 있다. 따라서 본 연구에서는 기존 가뭄지수에 활용 중인 기상 및 수문정보(강수량, 댐 유입량)에 각각 비정형 데이터(뉴스데이터)를 결합하여 가뭄지수를 산정하고, 산정된 가뭄지수의 검증을 통해 가뭄해석의 활용성을 평가하고자 한다. 결합가뭄지수 산정을 위해 Clayton Copula 함수를 활용하였으며, 매개변수 추정은 교정방법을 이용하였다. 분석결과, 기존의 가뭄지수(SPI, SDI)보다 비정형 데이터를 결합한 가뭄지수가 가뭄기간을 적절히 재현하는 것으로 나타났다. 또한 Receiver Operating Characteristic (ROC) score가 기존의 가뭄지수들보다 높게 산정되어 가뭄해석에 있어 활용성이 우수하였다. 본 연구에서 산정된 결합가뭄지수는 기존 단일변량 가뭄지수의 해석적 한계를 보완하고 비정형데이터를 활용한 가뭄지수의 활용성이 우수하다는 점에서 활용성이 높다고 판단된다.

**핵심용어:** 가뭄, 가뭄지수, 빅데이터, 비정형 데이터

***Corresponding Author.** Tel: +82-2-940-7770
E-mail: wrr21@naver.com (J. Ahn)

## 1. 서 론

자연적 요인에 의해 발생하는 자연재난 중 하나인 가뭄은 인간과 자연환경에 큰 영향을 미친다. 가뭄은 일정 기간 이상 평균 이하의 강수로 인해 강수량 부족이 장기화되는 현상으로, 홍수와 달리 장기적이고 광범위한 지역에 영향을 미치므로 이로 인한 피해도 크게 나타난다. 특히 우리나라는 대략적으로 2년에 한번마다 가뭄에 의한 수자원공급에 있어서 긴장상태를 경험하고, 심한 경우에는 2년 이상 지속되는 가뭄으로 용수공급에 심각한 차질이 발생하는 것으로 나타나고 있다(Yoo and Ryoo, 2003). 이렇듯 가뭄은 심각한 피해를 유발할 수 있기 때문에 가뭄의 사상 및 특성을 파악하는 것은 가뭄 대응을 위한 중요한 요소이다. 그러나 가뭄은 비가시적이며, 다양한 수문학적 인자(강수량, 증발산량 등)들이 복합적으로 작용하여 발생하기 때문에 가뭄의 사상을 정확히 평가하는 것은 어려운 일이다. 그렇다보니 어떤 요인을 중점적으로 고려하느냐에 따라 기상학적 가뭄, 농업적 가뭄, 수문학적 가뭄, 사회경제적 가뭄으로 구분하고 있으며, 이를 정략적으로 해석하기 위한 다양한 가뭄지수들이 개발되어 왔다.

국내 가뭄해석에 주로 활용되는 가뭄지수로는 Palmer (1965)의 PDSI (Palmer Drought Severity Index), Mckee *et al*. (1993)의 SPI (Standardized Precipitation Index), Kwon *et al*. (2006)의 MSWSI (Modified Surface Water Supply Index), Choe and Go (2006)의 SMI (Soil Moisture Index) 등이 있으며, 이 지수들은 가뭄해석 및 가뭄 위기경보 수준(관심-주의-경계-심각) 판단기준에 활용되고 있다. 하지만 현재 활용중인 가뭄지수들은 단일변량의 부족량을 통해 산정되며, 이는 가뭄의 사상을 정확히 판단하지 못하는 문제가 있다. 단순 단일변량의 부족이 가뭄이 아닐 수 있으나, 가뭄이라고 판단하기 때문이다. 따라서 가뭄의 판단은 단일변량이 아닌 하나의 통합된 정보로 나타내어 가뭄을 해석할 수 있는 기술이 요구된다. 한편 국외에서는 다변량을 결합한 가뭄지수 개발에 관한 연구를 수행한 바 있는데, NDMC (National Drought Mitigation Center)에서는 6가지 지수에 대하여 가중치를 고려하여 미국 가뭄 감시정보(U.S Drought Monitor Information)를 생산 및 제공하고 있다(NDMC, 2002). Keyantash and Draucup (2004)은 기상학적 가뭄인자(강수량), 수문학적 가뭄인자(유출량), 농업적 가뭄인자(토양수분량)를 결합하여 ADI (Aggregate Drought Index) 가뭄지수를 개발하였다. 국내의 경우, Kim *et al*. (2012)은 지속시간(1, 2, 3 ~ 12개월)별 표준강수지수(SPI)를 결합한 JDI (Joint Drought Index)를 개발하고 국내에 적용한 바 있으며, So *et al*. (2014)은 강수량과 토양수분량을 활용한 이변량 결합가뭄지수를 산정하고 국내 가뭄해석의 활용성을 평가하였다.

이처럼 국내·외적으로 변량을 결합한 연구들은 많이 진행되어왔다. 그러나 수치화된 정형데이터만을 활용하여 가뭄지수들이 개발되어 왔으며 최근에는 빅데이터 분석에 많이 사용되고 있는 비정형 데이터를 활용하여 가뭄정보를 생산하거나 지수를 개발하는 연구들이 진행되고 있다. 국가가뭄정보포털에서는 시도별 가뭄관련 뉴스기사의 빈도를 산출하여 국민들에게 가뭄정보를 제공하고 있으며, Lee *et al*. (2015)은 SNS 데이터를 활용하여 가뭄지수를 산정하고 가뭄분석의 새로운

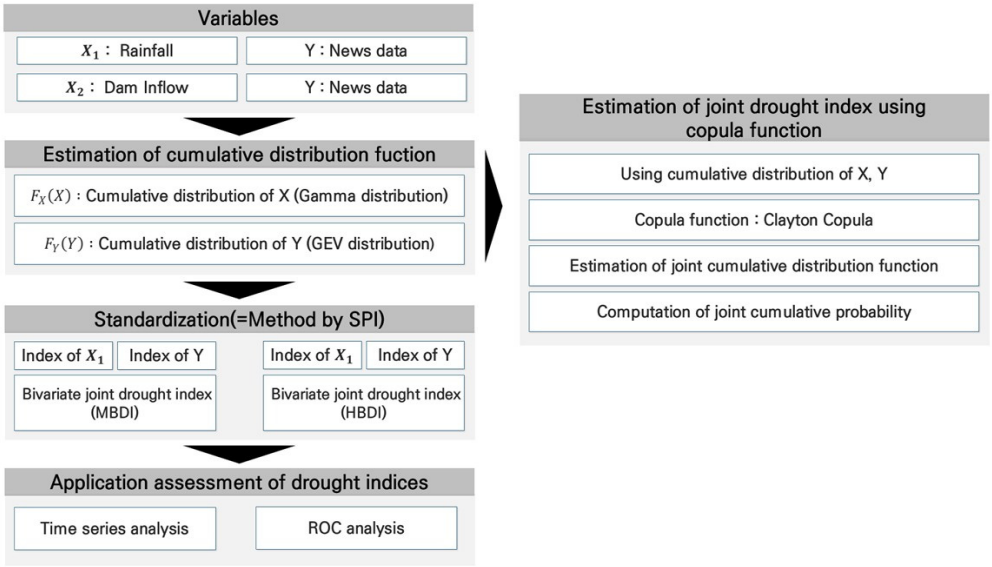



**Fig. 1.** Procedure for the study

접근법을 제시한 바 있다. 또한 타 분야에서는 Park *et al.* (2018)이 뉴스 데이터를 활용하여 해운뉴스지수를 개발하고 예측회귀분석을 통해 해운뉴스지수의 유용성을 평가하였다. 이와 같이 최근에는 비정형 데이터를 활용한 연구들이 수행되고 있으며 우수성이 입증되고 있다. 뉴스 데이터 혹은 SNS 데이터와 같은 비정형 데이터의 장점은 가뭄이 발생하거나 피해를 입으면 많은 기사들이 쏟아지고, SNS에 관련 글이 업로드 되기 때문에 사람들이 실제로 체감하는 데이터라는 것이다.

따라서 본 연구에서는 기존 가뭄지수에 활용 중인 기상 및 수문정보(강수량, 댐 유입량)에 각각 비정형 데이터(뉴스 데이터)를 결합하여 기상학적 빅데이터 가뭄지수(Meteological Bigdata Drout Index, MBDI)와 수문학적 빅데이터 가뭄지수(Hydrological Bigdata Drought Index, HBDI)를 산정하고, 산정된 가뭄지수의 검증을 통해 국내 가뭄해석의 활용성을 평가하고자 한다. Fig. 1은 연구절차를 설명한 것이다.

## 2. 연구방법

### 2.1 웹 크롤링(Web Crawling)

웹 크롤링(Web Crawling)이란 비정형 데이터를 수집하는 방법으로써, 인터넷 상에 있는 텍스트 자료를 수집하고 데이터 분석이 용이하도록 가공하는 작업을 의미한다. 웹 크롤링은 Python을 기반으로 개발되었으며, BeautifulSoup와 Selenium 방식이 있다.

BeautifulSoup 방법은 웹 페이지의 HTML, XML 파일의 정보를 추출해내는 Python 기반의 라이브러리로, Python 내장 모듈(Module)인 requests 혹은 urllib을 이용해 HTML을 다운로드 받고, BeautifulSoup으로 데이터를 추출하는 방식이다. 본 방식은 HTML을 다운받기 때문에 서버사이드렌더링(Server-Side-Rendering)을 사용하지 않는 사이트, 즉 HTML을 제공하지 않는 사이트의 경우 크롤링 하기 어려운 단점이 있으나, 사용성이 간편하고 데이터를 수집하는 속도가 빠르기 때문에 많이 활용되고 있다.

Selenium 방법은 인터넷 브라우저(Browser)를 통해 크롤링을 하는 개념으로써, 실제 보여지는 웹페이지의 데이터를 수집하는 방식이다. 본 방식은 크롤링하는 작업을 직관적으로 관찰 할 수 있기 때문에 디버깅(Deburging)에 유리하며, 웹페이지에서 Javascript 렌더링을 통해 생성되는 데이터들을 손쉽게 가져올 수 있는 장점이 있으나, 웹 페이지를 실제로 실행시키는 방식이기 때문에 크롤링하는 속도가 느리고, 수집된 데이터의 용량이 크다는 단점이 있다.

웹 크롤링을 통해 비정형 데이터가 수집이 되면 텍스트 분석을 통한 정제작업이 수행된다. 수집된 자료를 대상으로 Python 패키지 koNLPy를 통해 형태소 분석을 실시 한 뒤, 특수문자, 숫자, 기호 등 필요하지 않은 텍스트는 제거하고 명사만을 추출하였다. 추출된 명사를 통해 특정 키워드(Keyword)를 입력하여 수집된 기사가 가뭄관련 기사 및 연구대상지점의 기사인지를 파악하였다.

### 2.2 Copula 함수 및 매개변수 추정

Copula 함수는 여러 변수들 사이의 종속성 구조를 고려하면서 누적분포함수를 추정하는데 유용한 방법으로 1959년 Sklar에 의해 처음으로 제시되었다. Copula 함수는 각 변량의 분포 특성이 다를 경우, 개별적 분포의 특성을 결합 또는 분리가 용이하여 각 변량들의 분포특성을 효과적으로 반영한다고 알려져 있다(Sklar, 1959). 그동안 국내 가뭄관련 연구에서도 Copula 함수를 활용한 연구들이 수행되어 왔다(Kim *et al.*, 2012; Yoo *et al.*, 2013; So *et al.*, 2014).

Copula 함수를 활용하기 위해 Sklar의 정리를 살펴보면 다음과 같다. 확률변수 $X_1,\cdots,X_n$의 주변분포함수가 $F_1(x_1),\cdots,F_{(n)}(x_n)$일 때, 확률변수들의 결합분포함수 $F(x_1,\cdots,x_n)$에 대하여 n차원의 Copula 함수 $C$가 존재하며 다음이 성립한다.

$$F(x_1,\cdots,x_2) = C(F_1(x_1),\cdots F_n(x_n)) \tag{1}$$

만약, 확률변수 $U_1,\cdots,U_n$이 구간 [0, 1]에서의 균일분포를 따르고, $F_1(x),\cdots,F_n(x_n)$이 연속형인 경우에 Copula 함수 C는 유일하게 존재하면 Copula 함수는 다음과 같이 주어진다.

$$\begin{aligned} C(u_1,\cdots,u_n) &= P(U_1 \le u_1,\cdots,U_n \le u_n) \\ &= C(F^{-1}(u_1),\cdots,F_n^{-1}(u_n)) \end{aligned} \tag{2}$$

여기서 $u_i \in [0, 1]$이고 $F_i^{-1}$는 $F_i$의 역행렬이다($i = 1,\cdots,n$).

Copula 함수는 Archimedean Copula, Clayton Copula, Gumbel Copula, Kernel Copula 등이 있으며 이 중, Clayton Copula는 극소값(Smallest) 추정 및 자료의 꼬리 구조를 잘 반영하는 것으로 알려져 국내 · 외 가뭄 연구에 활용되어 왔다(Shiau *et al.*, 2007; Kwak *et al.*, 2013; So *et al.*, 2014). 따라서 본 연구에서는 Clayton Copula 함수를 활용하였으며, 식은 다음과 같다.

$$\begin{aligned} C_\theta(u,v) &= \max[(u^{-\theta}+v^{-\theta}-1)^{-\frac{1}{\theta}},0] \\ \phi_\theta(t) &= \frac{1}{\theta}(t^{-\theta}-1) \end{aligned} \tag{3}$$

Clayton Copula 함수의 매개변수 추정을 위해서 교정방법(Calibration method by using sample dependence measure)을 활용하였다. 이 방법은 Kendall의 순위 상관계수 $\alpha$를 이용해 계산된다. 계산방법은 두 개의 확률변수를 크기에 따라 순위를 부여하고 증감의 경향을 파악한다. 증감의 값에 합을 계산 한 후 순위 상관계수($\alpha$)를 산정한다. 산정된 계수($\alpha$)는 Copula 함수의 매개변수 추정에 활용되며 산정방법은 다음과 같다.

$$\theta = \frac{2\alpha}{1-\alpha} \tag{4}$$

본 연구에서는 기상학적 가뭄인자(강수량), 수문학적 가뭄인자(댐 유입량)을 각각 비정형데이터(뉴스데이터)와 결합하여 결합 누가확률값을 산정하였다.

### 2.3 가뭄지수 산정 및 활용성 평가

본 연구에서는 Clayton Copula 함수로부터 산정된 결합 누가확률값을 가뭄지수로 변환하기 위해 Mckee *et al.* (1993)이 표준강수지수(SPI)에서 적용한 방법을 활용하였다. 이 방법을 통해 가뭄지수(Z)는 이변량 값에 따른 결합 누가확률 P1을 산정한 후 표준정규분포 상에서 동일한 누가확률 P2에 해당하는 X축 값이 가뭄지수가 된다(Fig. 2).

결합 누가확률값에 따라 지수로 변환되는 식은 표준강수지수 계산에서 활용되고 있는 Eqs. (5) and (6)을 사용하였다.

$$Z = -(t - \frac{c_0 + c_1 t + c_2 t}{1 + d_0 t + d_1 t^2 + d_2 t^3}), \text{ for } 0 < C(u,v) \le 0.5$$

$$Z = (t - \frac{c_0 + c_1 t + c_2 t}{1 + d_0 t + d_1 t^2 + d_2 t^3}), \text{ for } 0.5 < C(u,v) \le 1.0 \tag{5}$$

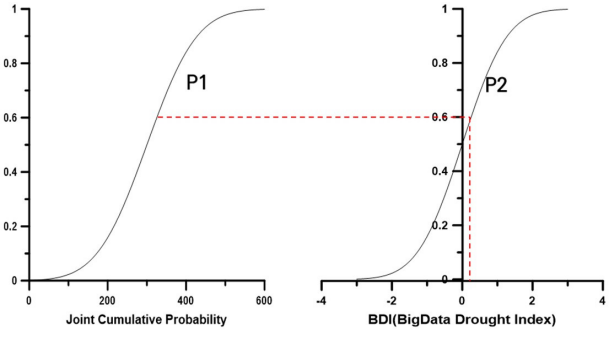



**Fig. 2.** Drought index calculation method

$$t = \sqrt{\lim(\frac{1}{(C(u,v))^2})}, \text{ for } 0 < (u,v) \le 0.5$$

$$t = \sqrt{\lim(\frac{1}{(1.0 - C(u,v))^2})}, \text{ for } 0.5 < (u,v) \le 1.0 \tag{6}$$

여기서, $C(u, v)$는 이변량 결합 누가확률값이며, $c_0 = 2.515517$, $c_1 = 0.802853$, $c_2 = 0.010328$, $d_0 = 1.432788$, $d_1 = 0.189267$, $d_2 = 0.001308$이다.

생산된 가뭄지수는 시계열분석과 ROC (Receiver Operating Characteristic)분석을 통해 활용성을 평가하였다. 시계열분석은 가뭄지수를 시계열로 도시한 후, 가뭄지수가 가뭄사상을 적절히 반영하는지에 관한 분석 방법이며, 본 연구에서는 표준강수지수와 개발된 가뭄지수를 비교·평가하였다. 그러나 시계열분석은 연구자의 주관이 개입될 수 있기 때문에 본 연구에서는 평가의 객관성을 확보하기 위해 ROC 분석을 활용하였다. ROC 분석은 가뭄사례와 가뭄지수의 가뭄발생 유무에 대한 상호비교를 통해 적중률과 비적중률을 산정하고 ROC score를 계산하여 가뭄지수의 정확도를 평가하는 방법이다. ROC Sore는 ROC Curve의 밑면적을 계산한 값인 Area Under the Curve (AUC)를 통해 산출되며, 값이 1.0인 경우 가뭄지수가 가뭄사례를 정확히 재현하였음을 의미한다.

## 3. 적용 및 결과

### 3.1 연구범위 및 연구자료

본 연구의 목적은 뉴스데이터와 강수량, 뉴스데이터와 댐 유입량을 각각 결합하여 가뭄지수를 산정하고, 산정된 가뭄지수의 활용성을 평가하고자 하는 것이다.

연구 대상지점은 2014 ~ 2015년 충남서북부가뭄 지역 중 가장 큰 피해를 입었던 보령지역으로 선정하였으며 Fig. 3과 같다. 강수량은 기상청에서 제공하는 보령지점의 강우자료를 활용하였으며, 댐 유입량은 국가수자원관리종합시스템(WAMIS)에서 제공하는 보령댐 유입량을 사용하였다.

뉴스데이터는 웹 크롤링(Web Crawling)을 통해 네이버 뉴스의 기사를 수집하였다. 네이버 뉴스는 453개 이상의 언론사, 67개의 매체로 구성되어 있어 다양하고 많은 기사를 수집할 수 있다는 장점이 있다. 뉴스데이터는 키워드(Keyword)를 통해 자료가 수집되며, 본 연구에서는 국가가뭄정보분석센터에서 빅데이터 가뭄분석에 활용하고 있는 '빅데이터 가뭄분석 뉴스 키워드'를 제공받아 참고하였으며 아래와 같다.

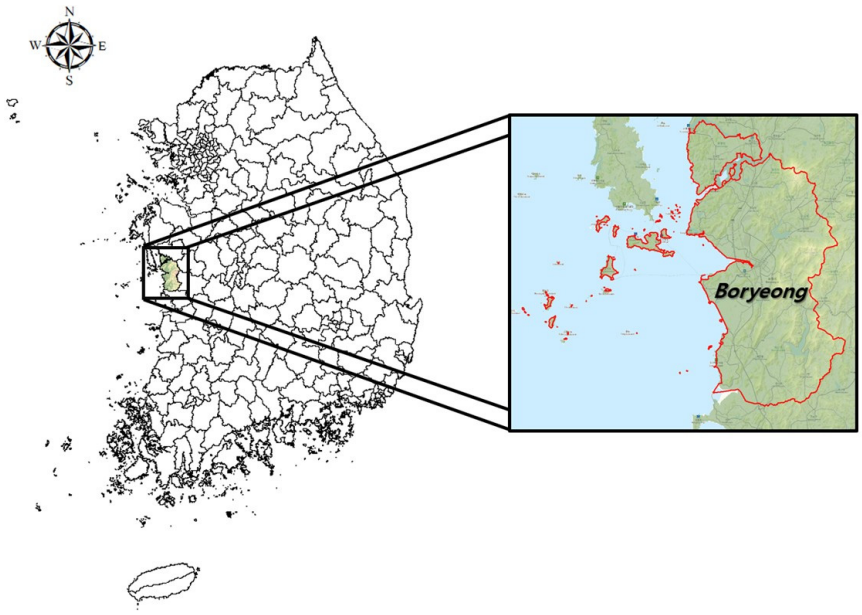


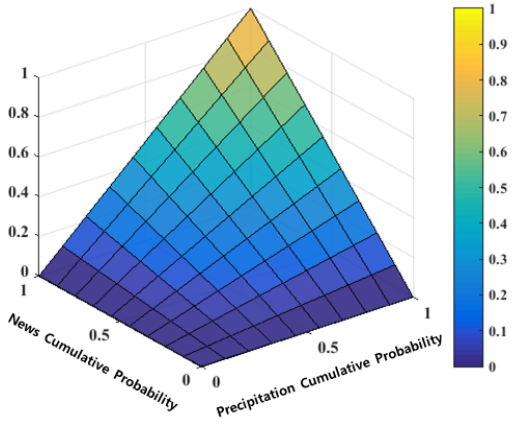

**Fig. 3.** Spatial range of study

**Table 1.** Calculation of monthly frequency

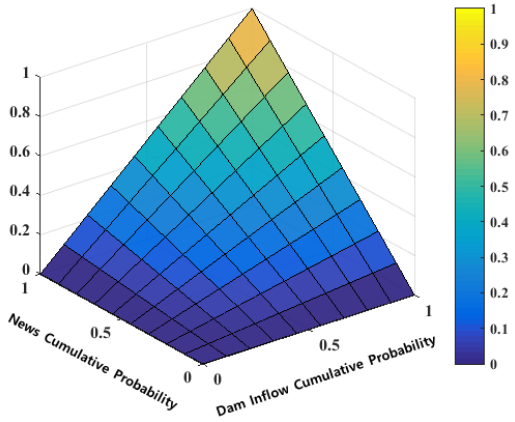
| Time (Month-year) | Frequency | Time (Month-year) | Frequency | Time (Month-year) | Frequency | Time (Month-year) | Frequency |
|---|---|---|---|---|---|---|---|
| Jan-13 | 3 | Jan-14 | 9 | Jan-15 | 6 | Jan-16 | 263 |
| Feb-13 | 2 | Feb-14 | 10 | Feb-15 | 3 | Feb-16 | 252 |
| Mar-13 | 4 | Mar-14 | 41 | Mar-15 | 10 | Mar-16 | 46 |
| Apr-13 | 8 | Apr-14 | 18 | Apr-15 | 3 | Apr-16 | 27 |
| May-13 | 0 | May-14 | 13 | May-15 | 6 | May-16 | 13 |
| Jun-13 | 0 | Jun-14 | 36 | Jun-15 | 312 | Jun-16 | 7 |
| Jul-13 | 3 | Jul-14 | 256 | Jul-15 | 357 | Jul-16 | 16 |
| Aug-13 | 7 | Aug-14 | 51 | Aug-15 | 256 | Aug-16 | 72 |
| Sep-13 | 4 | Sep-14 | 8 | Sep-15 | 322 | Sep-16 | 35 |
| Oct-13 | 6 | Oct-14 | 21 | Oct-15 | 839 | Oct-16 | 16 |
| Nov-13 | 1 | Nov-14 | 17 | Nov-15 | 483 | Nov-16 | 17 |
| Dec-13 | 7 | Dec-14 | 21 | Dec-15 | 284 | Dec-16 | 10 |
| Sum | 45 | Sum | 501 | Sum | 2,881 | Sum | 774 |

가뭄징조(4) : 강수량부족, 갈수, 저수율, 하천수위 저하
가뭄발생(6) : 가뭄, 물부족, 저수지고갈, 제한급수, 식수부족, 농작물 피해
가뭄대응(6) : 급수차, 급수지원, 물차, 자율급수, 물절약, 도수로
가뭄영향(4) : 지하수 고갈, 모내기 지연, 대체작물, 물 분쟁

위의 키워드를 활용하여 뉴스기사를 수집하였으며, 수집된 자료는 형태소 분석을 실시하여 명사만을 추출하였다. 추출된 명사를 기반으로 '보령', '대천', '충남' 이라는 연구 대상 지점의 키워드가 없는 기사, 가뭄관련 기사여부를 파악하기 위해 '가뭄' 키워드가 없는 기사는 제거하였다. 그 결과 4,201개의 뉴스기사를 수집하였으며, '가뭄발생' 관련 뉴스 2,513건(59.8%), '가뭄대응' 뉴스 1,215건(28.9%), '가뭄징조' 뉴스 286건(6.8%), '가뭄영향' 뉴스 187건(4.5%)으로 나타났다. 각 범주별로 수집된 뉴스기사는 하나의 통합된 월별 빈도로 산출하였으며(Table 1), 월별빈도는 뉴스데이터의 변량으로 활용하여 추후 변량 결합에 활용하였다.




(a) The join cumulative probability curve of precipitation and news



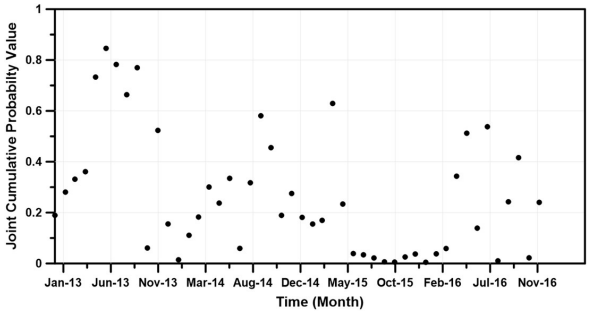

(b) The join cumulative probability curve of dam inflow and news

**Fig. 4.** The joint cumulative probability curve

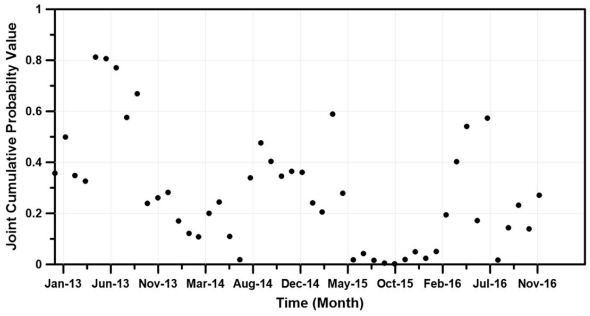
### 3.2 이변량 결합 누가확률값 산정 및 가뭄지수 개발

이변량 결합 누가확률값을 산정하기 위해 강수량과 댐 유입량의 확률분포형은 Mckee *et al.* (1993)가 제시한 Gamma분포, 뉴스데이터는 Xi-Squared 검정을 통해 유의수준 1%를 만족하는 GEV (Generalized Extreme Value)분포를 활용하여 누가분포함수를 산정하였다. 결합누가확률분포는 Clayton Copula 함수를 활용하였으며, 매개변수 추정에는 교정방법을 사용하였다.

Fig. 4는 결합 누가확률분포함수곡선을 3차원으로 도식화한 것이며, X축은 각각, 강수량과 댐 유입량, Y축은 뉴스데이터의 누가확률값을 의미한다(Fig. 4(a) and (b)). 결합누가확률분포 함수의 곡선은 Copula 함수의 매개변수로부터 그 형상이 결정되며, 이를 통해 결합누가확률값을 추정한다. Fig. 5(a)는 강우량과 뉴스데이터의 결합 누가확률값, Fig. 5(b)는 댐 유입량과 뉴스데이터의 결합 누가확률값을 나타내며, 결합 누가확률값은 표준화 과정을 거쳐 기상학적 빅데이터 가뭄지수(MBDI)와 수문학적 빅데이터 가뭄지수(HBDI)로 변환된다.




(a) Precipitation and news data




(b) Dam inflow and news data

**Fig. 5.** Estimation of joint cumulative probability value

**Table 2.** Classification of drought severity

| Drought Index (Z) | Drought Category |
|---|---|
| 2.00 ≤ Z | Extremely wet |
| 1.99 ~ 1.50 | Very wet |
| 1.49 ~ 1.00 | Moderately wet |
| 0.99 ~ −0.99 | Near normal |
| −1.00 ~ −1.49 | Moderate Drought |
| −1.50 ~ −1.99 | Severe Drought |
| −2.00 ≥ Z | Extremely Drought |

### 3.3 가뭄지수 활용성 평가

시계열에 따른 MBDI, HBDI의 거동 특성을 분석하기 위해 기존 가뭄지수(SPI, SDI)를 평균하여 2013 ~ 2016년까지 나열한 후, 당시 가뭄사상과 비교·검토를 수행하였다. 가뭄의 심도는 Table 2와 같이 SPI와 동일하게 구분하여 가뭄해석에 활용하였으며, 가뭄지수가 -1 이하일 때를 가뭄으로 판단하였다.

가뭄기록조사보고서(2015), 뉴스기사(4,201건), 가뭄 위기경보 발령 사례에 따르면, 보령은 2015년 7월 처음으로 가뭄 위기경보 '주의' 단계가 발령되었으며, '주의' 단계는 국지적 가뭄이 실제로 발생했을 경우 발령된다. 그 이후로 2015년 8, 9월에 각각 '경계', '심각' 단계가 발령되었고 2016년 3월 저수율이 회복되면서 위기경보가 해제되었다. 그러나 2016년 4월을 기준으로 제한급수가 해제되었으며, 일부 영농지에는

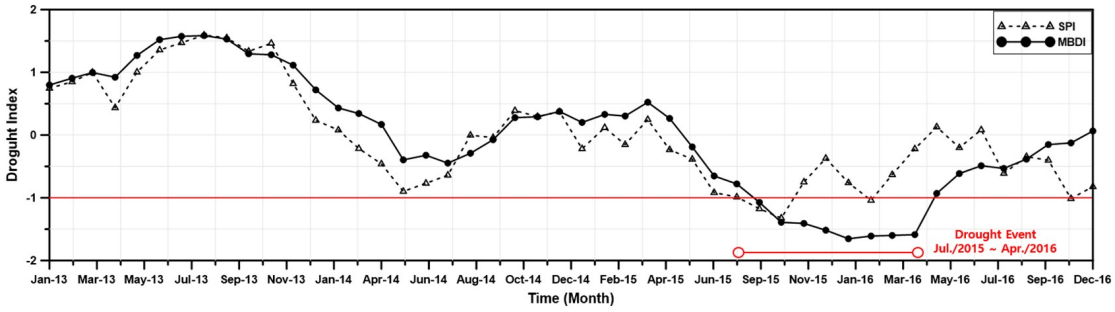



**Fig. 6.** Time series analysis of SPI and MBDI

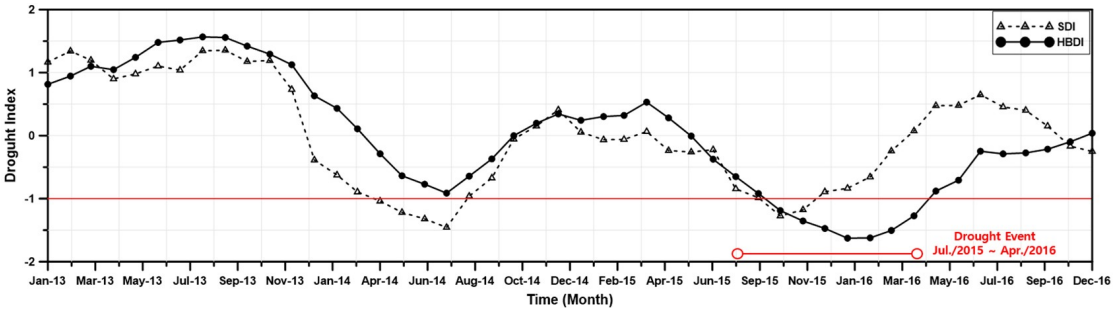



**Fig. 7.** Time series analysis of SDI and HBDI

가뭄피해가 발생하였다는 사례로 보아 당시 보령지역의 가뭄기간은 2015년 7월 ~ 2016년 4월(10개월)까지로 볼 수 있다.

이를 토대로 SPI와 SDI 검토 결과, 각각 2015년 9, 10월에 처음으로 가뭄을 감지하였으며. 두 지수 모두 초기가뭄 감지에는 한계가 있는 것으로 파악되었다. 또한 SPI의 경우, 2015년 9월 ~ 2015년 10월까지를 가뭄으로 보고 있으며, 2015년 11월부터는 가뭄이 해갈되는 것으로 나타났다. SDI는 가뭄의 전이현상으로 가뭄의 시작이 한달 정도 늦은 2015년 10월 ~ 2015년 11월까지를 가뭄으로 판단하였으며, 2015년 12월부터 가뭄이 해소되는 현상이 나타났다. 이를 통해 기존 가뭄지수(SPI, SDI)들이 가뭄기간(2015년 7월 ~ 2016년 4월, 10개월)을 적절히 반영하지 못하는 것으로 확인되었다. 반면에 MBDI, HBDI의 경우, 각각 2015년 9월, 10월에 가뭄을 감지하고 2016년 4월까지를 가뭄으로 보고 있으며, 2016년 5월에 가뭄이 해소되었다(Figs. 6 and 7). 이는 당시 보령지역의 가뭄기간을 기존의 가뭄지수보다 적절히 재현한 것으로 나타났다.

또한, 본 연구에서는 평가의 객관성을 확보하고자 ROC분석을 수행했으며, Fig. 8은 SPI, SDI, MBDI, HBDI 가뭄지수에 대한 ROC 분석 결과이다. 가뭄지수로부터 산정된 ROC socre를 살펴보면 SPI 0.68 SDI 0.58, MBDI 0.90, HBDI 0.84로 기존의 가뭄지수보다 빅데이터 가뭄지수들이 더 높게 산정된 것으로 나타났다. 따라서 본 연구에서 산정된 결합 가뭄지수는 가뭄해석에 있어 활용성이 높다고 판단된다.

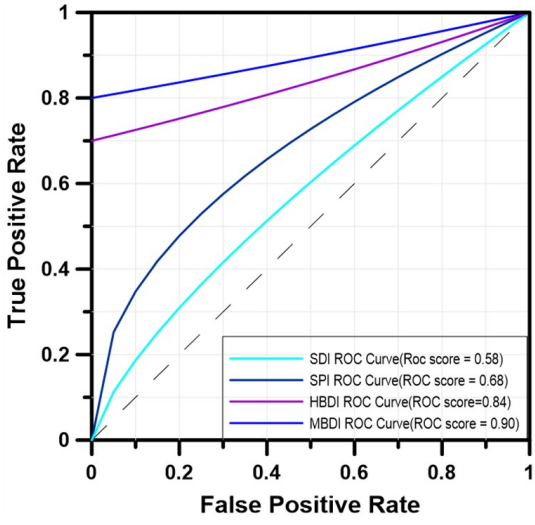



**Fig. 8.** ROC score using ROC analysis

## 4. 결 론

가뭄의 평가 및 위기경보 판단기준에 사용하고 있는 가뭄지수들은 단일변량 부족을 근거로 산정되고 있으며, 이는 가뭄을 정확히 판단하지 못하는 문제가 있다. 이를 보완하기 위해 변량을 결합한 가뭄지수들이 개발되고 있으나 선행연구들은 수치화된 정형 데이터만을 활용하였다. 하지만 최근에는

비정형 데이터를 활용하여 가뭄정보를 생산하거나 지수를 개발하는 연구들이 진행되고 있다. 따라서 본 연구에서는 기존 가뭄지수에 활용 중인 기상 및 수문정보에 비정형 데이터를 결합한 가뭄지수를 개발하고, 개발된 가뭄지수의 검증을 통해 가뭄해석의 활용성을 평가하였다. 본 연구의 주요내용 및 결과를 요약하면 다음과 같다.

1) 결합가뭄지수 산정을 위해 Clayton Copula 함수를 활용하였으며, 매개변수 추정은 교정방법을 이용하였다. 입력변수인 기상학적 가뭄인자(강수량), 수문학적 가뭄인자(댐 유입량)를 각각 비정정형 데이터(뉴스 데이터)와 결합하여 MBDI, HBDI를 산정하였다. 산정된 가뭄지수를 2015년 충남서북부가뭄 지역 중 가장 큰 피해를 입었던 보령지역에 적용하여 기존 가뭄지수(SPI, SDI)들과 시계열 분석을 통해 비교·평가하였다. 분석결과, SPI, SDI, MBDI, HBDI 모두 초기 가뭄감지에는 한계가 있는 것으로 파악되었다. 그러나 MBDI와 HBDI의 경우, 기존 가뭄지수들에 비해 가뭄 지속시간을 잘 나타냈었으며, 가뭄기간을 적절히 재현하였다.
2) 산정된 가뭄지수의 객관적 평가를 하고자 ROC분석을 수행하였다. 분석결과, ROC socre는 SPI 0.68, SDI 0.58, MBDI 0.90, HBDI 0.84로 기존의 가뭄지수보다 본 연구를 통해 산정된 가뭄지수들이 더 높게 나타났으며, 강수량과 뉴스 데이터를 결합한 MBDI가 ROC score 0.90으로 가장 높게 산정되었다. 따라서 본 연구에서 산정된 결합가뭄지수는 가뭄해석에 있어 활용성이 높다고 판단된다.

본 연구에서 산정한 결합가뭄지수는 기존 정형 데이터를 활용한 가뭄지수의 해석적 한계를 보완하고 비정형데이터를 활용한 가뭄지수의 활용성이 우수하다는 점에서 그 가치가 높다고 판단된다. 앞으로의 국내 가뭄해석은 본 연구에서와 비정형 데이터를 활용하여 기존의 가뭄지수의 활용성을 극대화하는 연구가 필요할 것으로 사료된다.

## 감사의 글

이 논문은 행정안전부 재난안전취약핵심역량 도약기술개발사업의 지원을 받아 수행된 연구임(2019-MOIS33-006).

## References

Choe, J.Y., and Go, Y,S. (2006). “Development of soil moisture index.” *Water of Future*, Vol. 39, No. 3, pp. 24-28.

Keyantash, J., and Dracup, J.A. (2002). “The quantification of drought: An evaluation of drought indices.” *American Meteorological Society*, Vol. 83, No. 8, pp. 1167-1180.

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*, KOSHAM, Vol. 12, No. 1, pp. 223-230.

Kwak, J.W., Lee, S.D., Kim, Y.S., and Kim, H.S. (2013). “Return period estimation of droughts using drought variables from standardized precipitation index.” *Journal of Korea Water Resources Association*, KWRA, Vol. 46, No. 8, pp. 795-805.

Kwon, H.J., Park, H.J., Hong, D.O., and Kim, S.J. (2006). “A study on semi-distributed hydrologic drought assessment modifying SWSI.” *Journal of Korea Water Resources Association*, KWRA, Vol. 39, No. 8, pp. 645-658.

Lee, B.R., Bae, B.G., and Choi, S.H. (2015). “Drought analysis using comparison of standardized precipitation index and social bigdata.” *Journal of Computing Science and Engineering*, Vol. 6, No. 6, pp. 16-18.

Mckee, T.B., Doesken, N.J., and Kleist, J. (1993). “The relationship of drought frequency and duration to times cales.” *8th Conference on Applied Climatology*. Anaheim, CA, U.S.

National Drought Mitigation Center (NDMC) (2002). *Three years and counting: What’s new with the drought monitor, drought mitigation center faculty publication 4*. NE, U.S.

Palmer, W.C. (1965). *Meteorological drought, research paper*. No. 45, U.S. Weather Bureau, Silver Spring, MD, U.S.

Park, S.H., Kwon, J.H., and Kim, T.I. (2018). “Developing the maritime news index using news bigdata analysis.” *Marin Poilcy Reserch*, Vol. 33, No. 1, pp.281-301.

Shiau, J.T., Feng, S., and Nadarajah, S. (2007). “Assessment of hydrological droughts for the Yellow river, China, using copulas.” *Hydrological Processes*, Vol. 21, No. 16, pp. 2157-2163.

Sklar, K. (1959). “Fontions de reprartition an dimensionset leurs marges.” Publications de l’Institut Statistique de l’Université de Paris 8, pp. 229-231.

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*, KWRA, Vol. 47, No. 2, pp. 171-182.

Yoo, C.S., and Ryoo, S.R. (2003). “Analysis of drought return and duration characteristics at seoul.” *Journal of Korea Water Resources Association*, KWRA, Vol. 36, No. 4, pp. 561-573.

Yoo, J.Y., Shin, J.Y., Kim, D.H., and Kim, T.W. (2013). “Drought risk analysis using stochastic rainfall generation model and copula functions.” *Journal of Korean Water Resources Association*, KWRA, Vol. 46, No. 4, pp. 425-437.