Journal of the Korean Society of Civil Engineers
Vol. 38, No. 6: 801-808/ December, 2018
DOI: https://doi.org/10.12652/Ksce.2018.38.6.0801

ISSN 1015-6348 (Print)
ISSN 2287-934X (Online)
www.kscejournal.or.kr

Water Engineering **수공학**

호우피해자료에서의 고차원 자료 및 다중공선성 문제를 해소한 회귀모형 개발

김정환* · 박지현** · 최창현*** · 김형수****
Kim, Jeonghwan*, Park, Jihyun**, Choi, Changhyun***, Kim, Hung Soo*****

Development of Regression Models Resolving High-Dimensional Data and Multicollinearity Problem for Heavy Rain Damage Data

ABSTRACT

The learning of the linear regression model is stable on the assumption that the sample size is sufficiently larger than the number of explanatory variables and there is no serious multicollinearity between explanatory variables. In this study, we investigated the difficulty of model learning when the assumption was violated by analyzing a real heavy rain damage data and we proposed to use a principal component regression model or a ridge regression model after integrating data to overcome the difficulty. We evaluated the predictive performance of the proposed models by using the test data independent from the training data, and confirmed that the proposed methods showed better predictive performances than the linear regression model.

Key words : Heavy rain damage, Linear regression model, Principal component regression, Ridge regression

초 록

선형회귀모형의 학습은 일반적으로 자료의 개수가 설명변수의 개수보다 충분히 크고, 설명변수들 사이에 심각한 다중공선성이 없다는 가정 하에서 안정적으로 이루어진다. 본 연구에서는 이러한 가정이 위배되었을 경우 모형 학습의 어려움을 실제 호우피해자료를 분석함으로써 조명하였고, 이를 해결하기 위해 자료를 통합한 다음 주성분회귀모형 또는 능형회귀모형을 사용할 것을 검토하였다. 모형의 학습에 사용된 자료와 별도의 독립된 자료에서 제안된 모형들의 예측력을 평가하였고, 제안된 방법이 선형회귀모형보다 더 나은 예측력을 보이는 것을 확인하였다.

검색어 : 호우피해, 선형회귀모형, 주성분회귀, 능형회귀

1. 서 론

최근 자연재난의 규모와 발생빈도는 증가하고 있으며, 급격한 도시화로 인해 피해가 심각해지고 있는 실정이다(Kim et al., 2017c). 만약 기존에 발생했던 피해사례와 해당 시점 기상자료의 관계를 파악하여 피해 발생 전에 행정구역별(시군구 별)로 해당 지역의 피해정도를 예측할 수 있다면, 재난으로 인한 피해 복구비를 산정하는데 도움을 줄 수 있을 것이다. 과거에 발생한 자연재난 피해와 기상자료를 기반으로 사전에 피해를 예측하는 연구사례를 찾아보면, 먼저 태풍, 호우, 강풍, 풍랑, 대설 등의 자연재난으로 인한 피해와 이를 직접적으로

* 인하대학교 수자원시스템연구소 박사후연구원 (Inha University · sinkei9456@naver.com)
** 인하대학교 통계학과 석사과정 (Inha University · qkrwlguscj@gmail.com)
*** 정회원 · 인하대학교 토목공학과 박사수료 (Inha University · karesma0cch@naver.com)
**** 종신회원 · 교신저자 · 인하대학교 사회인프라공학과 교수 (**Corresponding Author** · Inha University · sookim@inha.ac.kr)

Received September 4, 2018/ **revised** October 2, 2018/ **accepted** November 13, 2018

유발하는 기상요소와의 관계를 고려하여 선형회귀모형을 개발한 연구들이 있다(Munich, 2002; Lee, 2012; Zhai and Jiang, 2014; Kim et al., 2017a; Choi et al., 2017b). 호우로 인한 피해를 일으키는데 직접적인 영향을 미치는 일강우량 자료나 대설 피해를 일으키는데 직접적인 영향을 미칠 것으로 판단되는 최심신적설량 등의 기상요소만을 사용한 연구의 경우 간편하게 몇 가지 기상요소만을 이용하여 피해를 예측할 수 있는 장점이 있지만, 해당 지역의 특징을 반영하지 못하는 단점이 있다. 따라서 지역별 사회·경제적인 요소들을 반영한 연구들이 진행되었다(Pielke and Downton, 2000; Mendelsohn and Saher, 2011; Jeong and Lee, 2014; Choo et al., 2017; Oh and Chung, 2017). 앞에 살펴본 연구들에서는 자연재난으로 인한 피해와 피해 관련 인자들(기상요소 및 사회·경제적 요소)이 선형적인 관계에 초점을 맞추어져 있는데, 실제 피해액과 피해 관련 인자들 간의 비선형성을 고려하기 위하여 인공신경망과 비선형 회귀식과 같은 모형들이 개발되었다(Mandal et al., 2005; Lee et al., 2016; Choi et al., 2017a; Kwon et al., 2017; Kim et al., 2017b). 최근에는 컴퓨팅 기술이 발전하면서 의사결정나무, 랜덤포레스트 등의 머신러닝 기법을 적용하는 연구들도 진행되었다(Furquim et al., 2016; Choi et al., 2017c; Choi et al., 2018).

이러한 기존의 연구들을 참고하여 호우피해액을 추정 가능한 모형을 개발할 수 있다. 먼저 반응변수를 호우피해액으로 하고, 호우피해가 발생한 해당 시점의 기상자료와 사회·경제적인 요소를 설명변수로 하는 호우피해자료를 수집한다. 이후 수집된 자료로부터 회귀모형을 학습하여 향후 호우 피해액 추정에 활용한다. 그런데 모형의 개발 과정에서 간과할 수 있는 2가지 중요한 문제점이 있다. 첫째, 행정구역 별(시군구 별)로 호우피해액을 추정하는 회귀모형을 개발하고자 하는 경우이다. 호우피해사례가 소수인 경우가 대부분이므로 자료의 개수보다 설명변수의 개수가 많은 경우가 빈번히 발생한다. 이러한 자료의 형태를 고차원 자료(high-dimensional data)라고 하는데, 고차원 자료에서 학습된 회귀모형의 경우 일부 설명변수의 회귀계수가 추정되지 않는다. 둘째, 수집된 설명변수들 사이의 강한 상관관계로 인한 다중공선성(multicollinearity) 문제가 심각한 모형이 개발될 수 있다는 점이다. 전술한 2가지 문제점을 포함한 자료에서 개발된 회귀모형은 과적합(over-fitting)될 위험이 있으므로 부정확한 예측값을 제시할 가능성이 크다. 따라서, 본 연구의 목적은 전술한 2가지 문제점을 해소한 회귀모형을 호우피해함수로 제안하는 것이다.

본 연구는 다음과 같이 구성되어있다. 2절에서는 호우피해함수 개발에 사용될 통계적 방법론들을 소개한다. 먼저 기존의 다중회귀모형을 소개하고, 고차원 자료와 다중공선성의 문제점들을 해소하기 위한 통계적 방법들을 소개한다. 3절에서는 호우피해함수 개발에 사용되는 자료에 대해 소개하고, 2절의 통계적 방법론들을 적용한 결과들을 기술한다. 4절에서는 예측력 평가 결과를 요약하여 최종적인 호우피해함수를 제안하고 결과 요약 및 향후 연구방향에 대해 논의한다.

2. 호우피해 예측함수 구축을 위한 모형

2.1 선형회귀모형

선형회귀모형(linear regression model; reg)은 설명변수(predictor variable)들을 사용하여 반응변수(response variable)의 값을 예측하는 대표적인 통계 모형이다. 반응변수를 y, 설명변수집합을 $X=(x_0, x_1, \cdots, x_p)$, $x_0=1$, 그리고 회귀계수를 $\beta=(\beta_0, \beta_1, \cdots \beta_p)$라고 한다면, 반응변수 y는 설명변수와 회귀계수 사이의 선형적 결합에 의해 다음과 같이 표현된다.

$$y=X\beta+\epsilon=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p+\epsilon \tag{1}$$

위 Eq. (1)에서 β_0은 절편을, $\beta_1, \cdots, \beta_p$는 각 설명변수에 대응되는 회귀계수를 의미한다. ϵ는 모형에 의해 설명되지 않는 오차(error)로써 평균이 0이고 분산이 σ^2인 정규분포를 따른다고 가정한다. 회귀계수 $\beta=(\beta_0, \beta_1, \cdots \beta_p)$는 n개의 자료로 이루어지는 다음의 목적함수 $L(\beta)$를 최소화하도록 추정된다.

$$L(\beta)=\sum_{i=1}^{n}\left(y_i-\Sigma_{j=0}^{p}\beta_j x_{ji}\right)^2=\|y-X\beta\|^2 \tag{2}$$

목적함수 $L(\beta)$를 최소화하는 회귀계수의 추정치는 $\hat{\beta}=(X^TX)^{-1}X^Ty$와 같이 구해지고, $\hat{\beta}$로부터 오차분산의 추정량은 $\widehat{\sigma^2}=L(\hat{\beta})/(n-p-1)$와 같이 구해진다. 추정된 개별 회귀계수 $\hat{\beta}_j$는 다른 설명변수들이 고정되었을 때 설명변수 x_j가 한 단위 증가할 때 반응변수의 기댓값의 증가량을 의미하며, 대 표본 하에서 평균이 β_j이고 분산이 $(X^TX)^{-1}_{jj}\times\sigma^2$인 정규분포를 따른다. $\hat{\beta}_j$의 확률분포를 이용하여 귀무가설 $H_0:\beta_j=0$에 대한 가설검정을 실시할 수 있다. 귀무가설이 기각되면 해당 설명변수가 반응변수에 대해 유의한 영향력을 미친다고 말하고, 반대의 경우에는 유의하지 않은 설명변수라고 판단한다.

2.2 고차원 자료에서의 모형 학습의 문제점

선형회귀모형에서의 회귀계수 $\hat{\beta}$은 자료의 개수 n이 설명변수의 개수 p 보다 충분히 크다는 가정 하에 안정적으로 추정된다. 반대의 경우, 즉, $n \leq p$인 자료를 고차원 자료(high-demensional data)라고 하며, 이러한 자료에서는 추정해야 할 회귀계수 중 일부가

추정되지 않는다. 이는 선형 방정식의 해를 구하는 경우와 같은데, p개의 회귀계수를 추정하기 위해 자료의 개수 n이 최소한 p 보다는 더 커야 해가 유일하게 결정될 수 있으며, 추정량의 표준오차까지 안정적으로 계산하기 위해서는 실제로는 $n \gg p$의 조건이 만족되어야 한다(Johnstone and Titterington, 2009). 또한 고차원 자료에서 학습된 회귀모형은 모형의 학습에 사용된 자료를 과적합(over-fitting)하게 되어 미래 시점에서 새로운 설명변수 정보가 주어졌을 때, 과소추정 되거나(under-estimated) 과대추정 되는(over-estimated) 예측값을 제시할 위험이 있다.

행정구역 별(시군구 별)로 호우피해액을 추정하는 회귀모형을 개발하고자 하는 경우, 호우피해사례가 소수인 경우가 대부분이므로 자료의 개수보다 설명변수의 개수가 많은 고차원 자료의 형태를 갖는다. 따라서 이런 경우에는 모형을 학습하기 전에, 인접한 시군구나 비슷한 성격을 갖는 시군구끼리 자료를 통합하는 것이 권장된다. 자료 통합 후 자료의 개수가 설명변수의 개수보다 많게 되면 고차원 자료의 문제가 해소된다.

2.3 다중공선성의 해소를 위한 모형

회귀분석을 수행하는데 있어 설명변수들 사이에 상관관계가 강할 경우 다중공선성의 문제가 나타난다. 심각한 수준의 다중공선성은 회귀계수의 추정치와 표준오차를 왜곡시킴으로써 회귀계수의 해석과 가설검정의 결과를 왜곡할 수 있다. 다중공선성을 해소하기 위해서 연구자의 직관에 의존하여 일부 변수만을 선택하거나 혹은 통계적 변수선택절차를 적용할 수도 있으나 실제로 변수선택을 어떻게 할 것인지에 대한 논의가 필요하다. 왜냐면 변수선택 시 중요한 설명변수가 누락될 가능성이 있을 뿐만 아니라 중요한 설명변수들 사이에 상관성이 강한 경우 어떤 변수를 우선적으로 선택해야 하는지에 대해서 명확한 기준이 없기 때문이다. 본 연구에서는 변수선택에 대한 대안으로 주성분회귀모형(principal component regression model; pc-reg) 혹은 능형회귀모형(ridge regression model; ridge-reg)을 제안한다.

2.4 주성분회귀모형

주성분회귀모형은 원래의 설명변수 대신에 주성분들을 설명변수로 사용하는 회귀모형이다. 따라서 주성분회귀모형을 학습하기 이전에 설명변수 집합을 주성분으로 바꾸는 과정이 필요하다. p차원 설명변수들의 행렬을 X라고 하고, 각각의 설명변수들을 각 설명변수들의 평균으로 중심화(centering)한 행렬을 X_c라고 한다면, 설명변수들의 공분산 행렬은 $\Sigma_X = X_c^T X_c/(n-1)$ 와 같다. 여기에 스펙트럼 분해(spectral decomposition)를 적용함으로써 공분산 행렬을 아래와 같은 고유값과 고유벡터의 행렬곱으로 분해할 수 있다.

$$\Sigma_X = \frac{1}{n-1} X_c^T X_c = V D V^T \tag{3}$$

위 Eq. (3)에서 V는 $V^T V = I$ 가 성립하는 고유벡터들의 집합이고, D는 고유값들로 이루어진 대각행렬이다. 위 식의 좌변 및 우변의 앞뒤에 각각 V^T와 V를 곱해주면 아래와 같다.

$$\frac{1}{n-1} V^T X_c^T X_c V = D = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix} \tag{4}$$

여기서, $Z = X_c V$라고 하면 위 Eq. (4)을 Z에 대한 공분산 행렬로 아래와 같이 바꾸어 쓸 수 있다.

$$\Sigma_Z = \frac{1}{n-1} Z^T Z = D \tag{5}$$

위 Eq. (5)에서 Z는 주성분들의 행렬을 의미한다. 주성분들의 공분산 행렬이 대각행렬인 점으로부터 주성분들 사이에는 서로 독립의 관계가 성립하게 됨을 알 수 있다. 따라서 주성분들을 원래의 설명변수 대신에 회귀모형에 사용할 경우 다중공선성 문제를 해결할 수 있게 된다. 또한 주성분들의 분산 사이에는 $\lambda_1 > \cdots > \lambda_p$의 관계가 있고, 주성분들의 분산합 $\Sigma_{i=1}^{p} \lambda_i$은 전체 설명변수의 변동으로 해석될 수 있다. 주성분들을 모두 다 사용하는 경우보다는 일부 주성분들만을 선택해서 사용하는 경우가 많다. 주성분을 선택하는 기준은 여러 가지가 있는데 대표적으로 2가지의 방법이 널리 사용된다. 첫째, 주성분의 분산값이 1이상인 주성분들만을 선택하거나 둘째, 처음 몇 개의 주성분들의 누적 분산이 전체 주성분들의 분산 합의 90%가 되는 기준으로 일부 주성분들만을 선택한다. 선택된 주성분들만을 원래의 설명변수 대신에 사용함으로써 설명변수의 차원축소(dimension reduction)를 달성할 수 있다.

주성분회귀모형을 학습하기 위한 과정을 요약하면 다음과 같다. 먼저 설명변수 집합에 대한 중심화와 설명변수들의 공분산 행렬에 대한 스펙트럼 분해를 통해 주성분 $Z = X_c V$를 계산한다. 이후 일부의 주성분들만을 선택하여 원래의 설명변수 대신에 회귀모형에 사용한다. 본 연구에서는 전체 설명변수의 약 90%의 변동을 설명하는 기준으로 주성분들을 선택하여 회귀모형의 학습에 사용하였다.

2.5 능형회귀모형

설명변수들 간에 강한 상관관계가 있는 경우 $X^T X$의 행렬식이 0에 가까워지게 되어 회귀계수 추정치의 분산이 매우 커지게 되는 문제가 발생한다. 이런 문제점을 해결하기 위해 Hoerl and Kennard

(1970)는 다음과 같은 목적함수를 최소화하는 회귀계수 β를 구하는 것을 제안하였다.

$$D(\beta) = \|y - X\beta\|^2 + \lambda\|\beta\|^2 \quad (6)$$

위 목적함수는 원래의 회귀모형의 목적함수 $L(\beta)$에 벌점항 $\|\beta\|^2$과 조절모수(tuning parameter) λ가 추가된 형태이며, 위의 목적함수를 최소화하는 회귀계수의 추정치는 $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$와 같이 구해진다. λ는 0을 포함하는 양수의 값으로써, $\lambda = 0$인 경우 능형회귀모형은 선형회귀모형과 일치한다. λ는 모형 학습에 사용되는 자료에서 교차검증(cross-validation)을 통해 모형의 예측력을 극대화하는 값으로 결정되기 때문에, 능형회귀모형의 예측력은 선형회귀모형과 같거나 더 우수하다. 단, 능형회귀모형의 추정치는 편향추정량(biased estimator)으로 개별 회귀계수에 대한 가설검정 $H_0 : \beta_j = 0$ 등을 위해서는 별도의 기법이 필요하다. 하지만 능형회귀모형을 통해 가설검정을 하고자 하는 경우는 거의 없고, 선형회귀모형보다 예측력 높은 회귀모형을 개발하기 위해 사용되는 경우가 대부분이다.

3. 실제자료 분석

3.1 자료수집 및 예측모형 개발과정

본 연구에서는 서울시를 대상지역으로 선정하여 호우피해함수 개발하고자 하였다. 반응변수로는 행정안전부(구 국민안전처)에서 제공하는 재해연보의 2005년부터 2016년 재해기간 별 호우피해액(단위: 천원)을 사용하였고, 과거의 화폐가치와 현재의 화폐가치가 상이하기 때문에 생산자 물가지수를 이용하여 현재(2016년)의 가치로 환산하였다. 설명변수로는 직접적인 호우피해를 유발할 것으로 기대되는 기상요소와 지역적 특징을 반영할 수 있는 사회·경제적 요소를 고려하였다. 우선 기상요소의 경우'기상자료개방포털'에서 제공하는 기상청 방재기상관측장비(Automatic Weather System; AWS)의 시강우 자료를 가공하여 재해기간 별 총 강우량, 선행강우량(1~7일), 지속시간별 최대강우량(1~24시간), 재해일수 자료를 구축하였다. 사회·경제적 요소로는 해당 시군구의 면적, 재정자립도, 지역내총생산(Gross regional domestic product; GRDP), 인구수, 취약인구수를 고려하였다.

피해액은 0원 이상의 값만을 포함하고 있으므로 학습된 회귀모형 또한 반드시 0원 이상의 예측값 만을 제시하여야 한다. 따라서 반응변수를 로그 변환하여 회귀모형을 학습하였다. 이 경우 모형에 의한 예측값은 로그 피해액이므로, 지수변환을 통해 원래의 피해액 단위로 환산해주어야 한다. 또한 모형의 실질적인 예측력을 평가하기 위해 2005년부터 2012년까지의 반응변수와 설명변수 자료만을 이용하여 모형을 학습한 다음, 학습된 모형에 2013년부터 2016년까지의 설명변수 자료를 입력하여 해당 기간 동안의 예측 피해액을 계산하였다. 예측 피해액과 실제 피해액을 비교함으로써 예측의 정확정도를 정량화 할 수 있으며 PRMSE (Predictive Root Mean Square Error)가 사용되었다. PRMSE는 실제값과 예측값 사이의 평균오차제곱근에 해당하는 값으로 값이 작을수록 모형의 상대적인 예측력이 더 높음을 의미하며 아래와 같은 형태로 주어진다.

$$\text{PRMSE} = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / n} \quad (7)$$

위 Eq. (7)에서의 y는 2013년부터 2016년 동안의 실제 피해액이며, $\hat{y}$는 실제 피해액에 대응되는 예측 피해액으로 학습된 모형에 의해 계산된다. 본 연구에서는 동일한 자료로부터 2개 이상의 모형을 학습한 다음, 모형 별 PRMSE를 비교하여 가장 작은 PRMSE를 갖는 모형을 최종모형으로 선정하는 방식으로 호우피해함수를 개발하였다.

3.2 서울시 개별 시군구 자료에 대한 선형회귀모형

행정구역 별(시군구 별)로 호우피해액을 추정하는 회귀모형을 개발할 수 있다. 시군구 별 자료에서의 반응변수는 해당 시군구의 로그 호우피해액($\log(y)$)이고, 38개의 설명변수는 해당 시군구에서 호우피해액의 동일 시점에 대응되는 면적($area$), 재해일수($ndays$), 재해기간별 총 강우량(tot), 1~7일 선행강우량($d_1, \cdots, d_7$), 1~24시간 지속시간별 최대강우량($x_1, \cdots, x_{24}$), 그리고 사회경제적 요인인 재정자립도, GRDP, 인구수, 취약인구수($s_1, \cdots, s_4$)이다. 반응변수와 설명변수로부터 Eq. (8)과 같은 회귀모형을 학습한다.

$$\begin{aligned} y = {} & \alpha_0 + \alpha_1 area + \alpha_2 nday + \alpha_3 tot \\ & + \sum_{j=1}^{7}\beta_j d_j + \sum_{k=1}^{24}\gamma_k x_k + \sum_{\ell=1}^{4}\delta_\ell s_\ell + \epsilon \\ & j = 1,\cdots,7,\ k = 1,\cdots,24,\ \ell = 1,\cdots,4 \end{aligned} \quad (8)$$

시군구 별 자료의 개수를 정리하면 Fig. 1과 같다. Fig. 1에서 x축은 자료의 개수를 의미하며, y축은 해당 자료의 개수를 갖는 시군구의 개수를 나타낸다. Fig. 1에서 모든 시군구의 자료의 개수는 10개 미만으로 설명변수의 개수인 38개 보다 적다. 따라서 모든 시군구는 자료의 개수보다 설명변수의 개수가 많은 고차원 자료의 형태로써 학습된 회귀모형이 과소추정 되거나 과대추정 되는 예측값을 제시할 위험이 있다.

각 시군구 마다 학습된 다중회귀모형에 대해 2013년부터 2015년까지의 설명변수 자료를 입력하여 해당 기간 동안의 예측 피해액을 계산한 후 해당 기간의 실제값과 비교하여 PRMSE를 계산하고자 하였으나, 일부 매우 크게 추정되는 예측 피해액으로 인해 값이

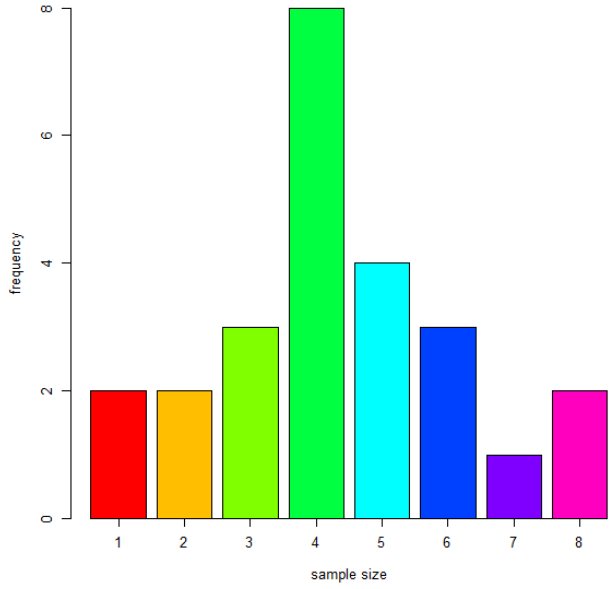

Fig. 1. Sample Sizes of Sigungu

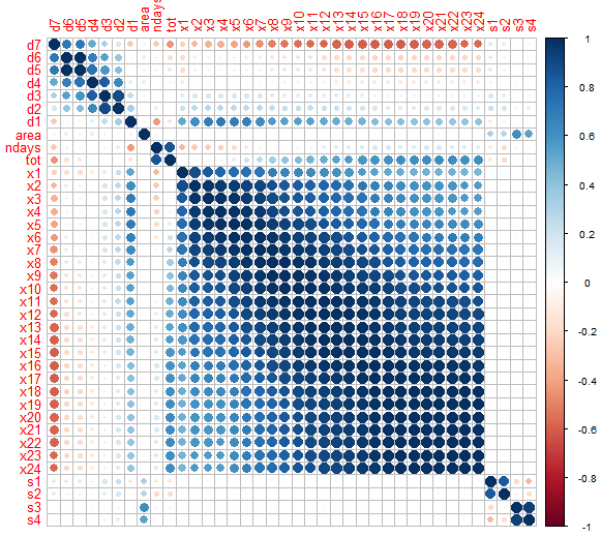

Fig. 2. Correlation Among Explanatory Variables

무한대로 표시되어 정확한 값을 알 수가 없었다. 무한대의 값을 갖는 PRMSE로부터 모형이 과대적합되었다고 판단할 수 있었으며, 고차원 자료인 행정구역 별(시군구 별) 자료에서 회귀모형을 학습하기 보다는 여러 행정구역을 통합하여 고차원 자료의 문제점을 해소한 다음 모형을 개발하는 방향으로 연구를 진행하였다.

3.3 서울시 전체 자료에 대한 선형회귀모형

고차원 자료의 문제점을 완전히 해소하기 위해 서울시에 속하는 모든 시군구의 자료를 통합하여 서울시 전체 자료를 이용하여 모형을 개발할 수 있다. 개발하고자 하는 모형의 형태는 Eq. (8)과 동일하며, 이때 학습에 사용되는 자료의 개수는 108개이다. 이는 설명변수의 개수 38개와 비교하였을 때 약 3배에 해당하는 크기이므로, 고차원 자료의 문제는 더 이상 발생하지 않는다. 모형의 예측력 평가는 동일한 방식으로 계산되었으며, PRMSE는 383,339 (천원)을 기록하였다. 이 값은 시군구 별로 학습된 모형에서 무한대의 값을 갖는 PRMSE에 비해서 훨씬 작은 값으로써, 고차원의 문제가 해소되어 예측모형 개발이 가능해졌다는 것을 시사한다. 이후에는 통합된 자료인 서울시 전체에 대해서 2.4절의 주성분회귀모형과 2.5절의 능형회귀모형을 적용하여 더 나은 예측모형을 개발하는 방향으로 연구를 진행하였다.

3.4 서울시 전체 자료에 대한 주성분회귀모형

서울시 전체의 자료에 대해 다중회귀모형을 적합하는 과정에서 p개의 설명변수를 대체할 수 있는 $m(\ll p)$ 개의 새로운 설명변수인 주성분을 원래의 설명변수 대신에 사용하고자 한다. Fig. 2는 시군구 면적($area$), 재해일수($ndays$), 재해기간별 총 강우량(tot), 1일~7일 선행강우량($d_1, \cdots, d_7$), 1~24시간 지속시간별 최대강우량($x_1, \cdots, x_{24}$), 그리고 사회경제적 요인인 재정자립도, GRDP, 인구수, 취약인구수($s_1, \cdots, s_4$)의 상관관계를 나타낸 것이다.

Fig. 2에서 짙고 큰 원일수록 강한 상관관계가 강함을 의미하며 여기서 거의 대부분의 상관계수는 양의 상관관계를 의미한다. Fig. 2에서 일부 설명변수의 부분집합 내에서의 강한 상관관계를 관측할 수 있다. 즉, 1일~7일 선행강우량 사이에 강한 상관성이 관측되며, 1~24시간 지속시간별 최대강우량에도 비슷한 패턴이 나타난다. 또한 사회경제적 요인인 재정자립도, GRDP, 인구수, 취약인구수 사이에도 강한 상관관계가 관측되었다.

Fig. 2를 참고하여 다음과 같이 주성분을 선택하였다. 원래 설명변수 집합의 분산의 약 90%를 설명할 수 있는 수준만큼의 소수의 주성분으로 차원을 축소하되 설명변수 부분집합 마다의 강한 상관관계를 고려하여, 1~7일 선행강우량($d_1, \cdots, d_7$)에서 2개의 주성분($P_1^{(d)}, P_2^{(d)}$)을, 1~24시간 지속시간별 최대강우량($x_1, \cdots, x_{24}$)에서 2개의 주성분($P_1^{(x)}, P_2^{(x)}$)을, 그리고 사회경제적 요소($s_1, \cdots, s_4$)에서 2개의 주성분($P_1^{(s)}, P_2^{(s)}$)를 선택하였다. 각 주성분은 주성분을 계산되는 데 사용된 원래의 설명변수들을 표준화하여 주성분 계수(principal loading)을 곱한 값과 같다. Table 1은 주성분을 계산하는 데 사용된 주성분 계수를 정리한 것으로 이로부터 주성분을 계산하는데 사용된 설명변수들의 영향력의 정도를 파악할 수 있다.

선행강우량($d_1, \cdots, d_7$)이 표준화되었다고 가정할 때, Table 1의 주성분 계수로부터 주성분 $P_1^{(d)}$은 $P_1^{(d)} = .0443 \times d_1 + .3526 \times d_2 + .4184 \times d_3 + \cdots + .4349 \times d_6 + .3370 \times d_7$로 계산된다. 선택된 총 6개의 주성분을 차원축소에 사용된 총 35개의 설명변수

대신에 회귀모형에 사용하였다. 주성분들을 대신 사용하여 학습되는 모형의 회귀식은 Eq. (9)와 같다. Eq. (9)는 Eq. (8)과 동일하게 설명변수 시군구 면적($area$), 재해일수($ndays$), 재해기간별 총 강우량(tot)를 사용하되, 1일~7일 선행강우량($d_1, \cdots, d_7$), 1~24시간 지속시간별 최대강우량($x_1, \cdots, x_{24}$), 그리고 사회경제적 요인인 재정자립도, GRDP, 인구수, 취약인구수($s_1, \cdots, s_4$) 대신에 주성분 $P_1^{(d)}, P_2^{(d)}, P_1^{(x)}, P_2^{(x)}, P_1^{(s)}, P_2^{(s)}$을 설명변수로 사용하는 회귀모형이다.

$$y = \tau_0 + \tau_1 area + \tau_2 nday + \tau_3 tot + \sum_{j=1}^{2}\lambda_j P_j^{(d)} + \sum_{k=1}^{2}\phi_k P_k^{(x)} + \sum_{\ell=1}^{2}\nu_\ell P_\ell^{(s)} + \epsilon \quad (9)$$

$$j = 1,2,\ k = 1,2,\ \ell = 1,2$$

원래의 회귀모형이 총 38개의 설명변수를 사용하는데 반해서 주성분을 사용한 Eq. (9)의 회귀모형은 반응변수와 주성분을 통한 차원축소를 통해 총 9개의 설명변수만을 사용한다. 즉, 서울시 전체의 자료에 대해 Eq. (9)의 회귀모형을 학습한다. 이후 2013년부터 2015년까지의 설명변수 자료를 입력하여 해당 기간 동안의 예측 피해액을 계산한 후, 해당 기간의 실제값과 비교하여 PRMSE를 계산한 결과 123,259(천원)을 기록하였고, 선형회귀모형의 PRMSE인 383,339(천원) 대비해서 훨씬 작은 값으로 예측력이 대폭 개선되었음을 의미한다. 이후로는 능형회귀모형을 적용하여 더 나은 예측모형 개발이 가능한지 여부를 확인하였다.

Table 1. Principal Loadings

Predictor variable	Principal component		Predictor variable	Principal component	
	1st	2nd		1st	2nd
d1	.0443	-.6071	x12	-.2185	.0132
d2	.3526	-.4475	x13	-.2180	.0348
d3	.4184	-.3300	x14	-.2174	.0670
d4	.4427	-.0938	x15	-.2161	.0988
d5	.4472	.2308	x16	-.2144	.1341
d6	.4349	.2604	x17	-.2118	.1638
d7	.3370	.4387	x18	-.2084	.1935
x1	-.1475	-.3060	x19	-.2061	.2088
x2	-.1814	-.3151	x20	-.2052	.2155
x3	-.1834	-.3246	x21	-.2046	.2200
x4	-.1839	-.3251	x22	-.2037	.2249
x5	-.1930	-.2811	x23	-.2018	.2342
x6	-.2014	-.2331	x24	-.2008	.2382
x7	-.2071	-.1889	s1	-.4977	.4854
x8	-.2107	-.1348	s2	-.4254	.5814
x9	-.2137	-.0878	s3	.5092	.5022
x10	-.2167	-.0481	s4	.5586	.4174
x11	-.2184	-.0174			

3.5 서울시 전체 자료에 대한 능형회귀모형

서울시 전체의 자료에 대해 능형회귀모형을 학습하고자 한다. 모형의 회귀식은 Eq. (8)과 동일하지만, 회귀계수는 교차검증(cross-validation)을 통해 결정되는 조절모수 λ 하에서 추정되는 것으로 선형회귀모형와는 다른 값을 갖는다. Fig. 3은 교차검증과정에서 MSE (Mean Square Error)를 최소화하는 조절모수 λ의 최적값을 찾는 과정을 나타낸 것이다.

Fig. 3에서 세로축은 교차검증된 MSE값을 의미하며, 세로축은 λ의 로그변환된 값이다. 교차검증된 MSE값을 최소화하는 λ의 값을 선택하여 능형회귀모형의 학습에 적용한다. 교차검증에 의해 선택된 λ의 값은 약 7.7386이다. 학습된 능형회귀모형의 예측력 평가는 동일한 방식으로 계산되어 PRMSE 값은 122,920(천원)으로 기록되었다. 이는 현재까지의 예측모형 중에서 가장 낮은 PRMSE를 기록한 것으로 예측력이 가장 우수하다는 것을 의미한다.

3.6 예측력 평가 결과 및 최종모형 선택

현재까지 개발된 모형의 예측력 평가 결과가 Table 2에 정리되었다. 시군구 별 학습된 회귀모형은 고차원 자료의 문제점으로 인해 비정상적인 PRMSE값을 기록하였으므로 Table 2에 포함시키지

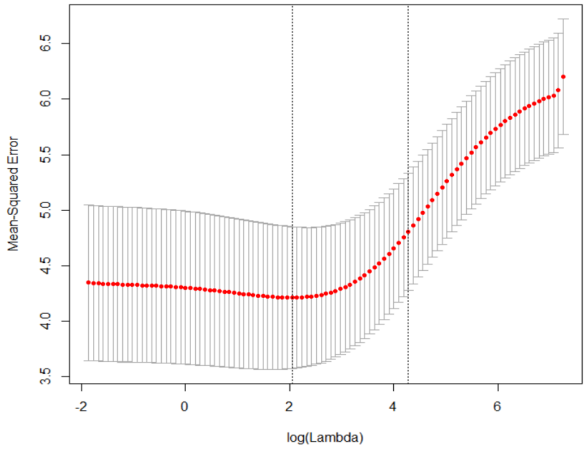

Fig. 3. Cross-Validation Plot for λ

Table 2. Predictive Performances (Unit: 1,000 KW)

Method	PRMSE
Linear regression model	383,339
Principal component regression model	123,259
Ridge regression model	122,920

Table 3. Estimated Regression Coefficients

Predictor variable	Linear regression model	Ridge regression model	Predictor variable	Principal component regression
Intercept	8.448738234	7.674429763	Intercept	10.433932371
area	-.038371098	-.000169025	area	-.015795790
ndays	.133174460	-.002006190	ndays	-.093279343
tot	-.001604802	.000054187	tot	.003747003
d1	.029029535	.006263781	$P_1^{(d)}$	-.089953129
d2	-.030648520	-.000580888	$P_2^{(d)}$	-.344404861
d3	.039525218	.000698674	$P_1^{(x)}$	-.163817324
d4	.006050075	.001308319	$P_2^{(x)}$	-.249858733
d5	-.028836930	-.001270523	$P_1^{(s)}$	.190212448
d6	.016025912	-.001379701	$P_2^{(s)}$	.185085271
d7	-.010384421	-.001294058		
x1	-.021400472	.004308308		
x2	.177109578	.002828289		
x3	-.195359988	.001361159		
x4	.032760696	.001286892		
x5	.099171649	.001213482		
x6	-.049857958	.001132881		
x7	.032226244	.001026318		
x8	.016961162	.000859155		
x9	.090694011	.000792136		
x10	-.175721269	.000668446		
x11	-.054024046	.000619365		
x12	.129018115	.000621070		
x13	.092769483	.000557138		
x14	-.313479987	.000495882		
x15	.228326684	.000513175		
x16	-.033462037	.000600890		
x17	-.119699424	.000562833		
x18	.004263287	.000532446		
x19	.376688648	.000500144		
x20	-.470017618	.000483605		
x21	-.151865326	.000495479		
x22	.325417820	.000491904		
x23	.219104099	.000462173		
x24	-.199545316	.000449602		
s1	.025371678	-.000534223		
s2	-.000000034	-.000000003		
s3	.000009444	.000000543		
s4	-.000049288	.000001674		

않았다. Table 2의 모형들은 고차원 자료의 문제가 없는 서울시 전체 자료에서 학습된 모형들의 결과이며, 이 중 능형회귀모형의 PRMSE 값이 가장 작으므로 최종모형으로 선택하여 서울시에 대한 호우피해함수로 제안할 수 있다.

Table 3은 각 모형들의 추정된 회귀계수를 정리한 것이다. Table 3에서 2열과 3열은 각각 선형회귀모형과 능형회귀모형의 회귀계수 결과로 Eq. (8)의 결과이다. 단, 능형회귀모형의 경우 $\lambda = 7.7386$ 하에서의 회귀계수를 추정한 것으로 선형회귀모형의 결과와는 조금 차이가 있다. Table 3의 5열은 Eq. (9)의 주성분회귀모형의 회귀계수 추정치이다. 주성분회귀모형의 경우 차원축소 효과로 인해 추정된 회귀계수의 개수가 훨씬 적은 것을 확인할 수 있다.

4. 결 론

본 연구는 회귀모형을 이용하여 호우피해함수를 개발함에 있어서 자료의 형태가 고차원 자료이거나 설명변수 사이에 다중공선성이 존재하는 경우 모형 개발의 어려움을 부각하고 이를 개선하기 위한 방법으로 자료 통합 및 주성분회귀모형과 능형회귀모형을 제시하여 최종 호우피해함수를 개발하는 과정을 소개하였다. 연구결과를 요약하면 다음과 같다.

(1) 피해함수개발을 위한 대상지역은 서울시를 선택하였고, 반응변수로는 행정안전부(구 국민안전처)에서 제공하는 재해연보의 2005년부터 2016년 재해기간 별 호우피해액(단위: 천원)을 사용하였다. 설명변수로는 선행강우량, 지속시간별 최대강우량, 총 강우량, 재해일수, 지역 면적, 재정자립도, 지역내총생산, 인구수, 취약인구수를 고려하였다.

(2) 행정구역 별(시군구 별)로 호우피해액을 추정하는 회귀모형을 개발했을 때, 자료의 개수보다 설명변수의 개수가 많은 고차원 자료의 문제가 나타났고, 실제로 학습된 회귀모형이 과대추정되는 것을 확인하였다. 따라서 서울시에 속하는 모든 시군구의 자료를 통합하여 자료의 개수를 늘림으로써 고차원 자료의 문제를 해소하였다. 이후 연구는 서울시 전체 자료로부터 모형을 학습하는 방향으로 진행하였다.

(3) 서울시 전체 자료에 대해 선형회귀모형, 주성분회귀모형, 능형회귀모형을 학습하여 모형들의 예측력을 비교하였다. 즉, 2005년부터 2012년까지의 자료만을 이용하여 모형을 학습한 다음, 2013년부터 2016년까지의 자료를 이용하여 학습된 모형의 예측력을 PRMSE로 평가하였다. 여러 모형의 PRMSE를 비교하여 가장 낮은 PRMSE 값을 갖는 모형을 최종 호우피해함수를 선정하였다.

본 연구에서는 고차원 자료의 문제를 해소하기 위한 방법으로 서울시에 속하는 모든 시군구의 자료를 통합하는 것을 제시하였으나 이것이 유일한 해결책은 아니다. 알려진 사전지식 혹은 적절한 군집분석을 통해 군집 단위로 자료를 통합하여, 군집 별 자료에서 모형을 개발할 수도 있을 것이다. 그리고 본 연구에서 능형회귀모형의 PRMSE 값이 가장 작긴 하지만, 주성분 회귀모형의 PRMSE와 큰 차이가 난다고 보기는 어렵다. 따라서 서울시가 아닌 지역에 대해서는 두 모형을 모두 적용하여 보다 더 나은 예측력 평가 결과를 제시하는 모형을 선택적으로 사용하는 방안도 고려해볼 수 있을 것으로 사료된다.

감사의 글

본 연구는 정부(행정안전부)의 재원으로 재난안전기술개발사업단의 지원을 받아 수행된 연구임(MOIS-재난-2015-05).

References

Choi, C. H., Kim, J. H., Kim, J. S., Kim, D. H., Bae, Y. H. and Kim, H. S. (2018). “Development of heavy rain damage prediction model using machine learning based on big data.” *Journal of Advances in Meteorology*, Vol. 2018, Article ID 5024930.

Choi, C. H., Kim, J. S., Kim, J. H., Kim, H. Y., Lee, W. J. and Kim, H. S. (2017a). “Development of heavy rain damage prediction function using statistical methodology.” *Journal of the Korean Society of Hazard Mitigation*, Vol. 17, No. 3, pp. 331-338 (in Korean).

Choi, C. H., Kim, J. S., Lee, M. J., Kim, J. H., Lee, W. J. and Kim, H. S. (2017b). “Development of heavy rain damage prediction function using principal component analysis and logistic regression model.” *Journal of the Korean Society of Hazard Mitigation*, Vol. 17, No. 6, pp. 159-166 (in Korean).

Choi, C. H., Park, K. H., Park, H. K., Lee, M. J., Kim, J. S. and Kim, H. S. (2017c). “Development of heavy rain damage prediction function for public facility using machin learning.” *Journal of the Korean Society of Hazard Mitigation*, Vol. 17, No. 6, pp. 443-450 (in Korean).

Choo, T. H., Kwak, K. S., Ahn, S. H., Yang, D. U. and Son, J. K. (2017). “Development for the function of wind wave damage estimation at the western coastal zone based on disaster statistics.” *Journal of the Korea Academia-Industrial cooperation Society*, Vol. 18, No. 2, pp. 14-22 (in Korean).

Furquim, G., Pessin, G., Faiçal, B. S., Mendiondo, E. M. and Ueyama, J. (2016). “Improving the accuracy of a flood forecasting model by means of machine learning and chaos theory.” *Neural computing and applications*, Vol. 27, No. 5, pp. 1129-1141.

Hoerl, A. E. and Kennard, R. W. (1970). “Ridge regression: biased estimation for nonorthogonal problems.” *Technometrics*, Vol. 12, No. 1, pp. 55-67.

Jeong, J. H. and Lee, S. H. (2014). “Estimating the direct economic damages from heavy snowfall in Korea.” *Journal of climate research*, Vol. 9, No. 2, pp. 125-139.

Johnstone, I. M. and Titterington, D. M. (2009). “Statistical challenges of high-dimensional data.” *Philos Trans A Math Phys Eng Sci*, Vol. 367, No. 1906, pp. 4237-4253.

Kim, J. H. Kim, T. G. and Lee, B. R. (2017a). “An analysis of typhoon damage pattern type and development of typhoon damage forecasting function.” *Journal of the Korean Society of Hazard Mitigation*, Vol. 17, No. 2, pp. 339-347 (in Korean).

Kim, J. S., Choi, C. H., Kim, D. H., Lee, M. J. and Kim, H. S. (2017b). “Development of heavy rain damage prediction function using artificial neural network and multiple regression model.” *Journal of the Korean Society of Hazard Mitigation*, Vol. 17, No. 6, pp. 73-80 (in Korean).

Kim, J. S., Choi, C. H., Lee, J. S. and Kim, H. S. (2017c). “Damage prediction using heavy rain risk assessment : (2) Development of heavy rain damage prediction function.” *Journal of Korean Society of Hazard Mitigation*, Vol. 17, No. 2, pp. 371-379 (in Korean).

Kwon, S. H., Lee, J. W. and Chung, G. H. (2017). “Snow damages estimation using artificial neural network and multiple regression analysis.” *Journal of the Korean Society of Hazard Mitigation*, Vol. 17, No. 2, pp. 315-325 (in Korean).

Lee, J. S., Eo, G., Choi, C. H., Jung, J. W. and Kim, H. S. (2016). “Development of rainfall-flood damage estimation function using nonlinear regression equation.” *Journal of the Korean Society of Disaster Information*, Vol. 12, No. 1, pp. 74-88 (in Korean).

Lee, S. I. (2012). “A study on damage scale prediction by rainfall and wind velocity with typhoon. master's thesis.” *Sunchon National University*.

Mandal, S., Saha, D. and Banerjee, T. (2005). “A neural network based prediction model for flood in a disaster management system with sensor networks.” *In Intelligent sensing and information processing, Proc. of 2005 international conference*, pp. 78-82.

Mendelsohn, R. and Saher, G. (2011) “The global impact of climate change on extreme events.” *World Bank*.

Munich, R. (2002). “Winter storms in europe: analysis of 1990 losses and future loss potentials.”

Oh, Y. R. and Chung, G. H. (2017). “Estimation of snow damage and proposal of snow damage threshold based on historical disaster data.” *Journal of the Korean Society of Civil Engineers*, Vol. 37, No. 2, pp. 325-331.

Pielke, R. A. and Downton, M. W. (2000). “Precipitation and damaging floods: trends in the united states, 1932-97.” *Journal of Climate*, Vol. 13, No. 20, pp. 3625-3637.

Zhai, A. R. and Jiang, J. H. (2014). “Dependence of US hurricane economic loss on maximum wind speed and storm size.” *Environmental Research Letters*, Vol. 9, No. 6, pp. 1-9.