치수 특성과 머신러닝의 t-SNE를 이용한 지역 군집화 제고 연구

Improvement of Regional Clustering Using Flood Control Characteristics and t-SNE of Machine Learning

주홍준* · 김종성** · 곽재원*** · 이종소**** · 정재원***** · 김형수*****

Joo, Hongjun*, Kim, Jongseong**, Kwak, Jaewon***, Lee, Jongso****, Jung, Jaewon*****, and Kim, Hung Soo*****

Abstract

The aim of this study was to perform clear clustering according to regional flood control characteristics in order to promote efficient and collective flood control management, which requires structural and non-structural measures to proactively prevent flooding and protect people and property. This approach requires to understand regional flood control characteristics; in particular, clustering based on flood control characteristics that are similar among regions can represent the right direction in establishing a strategy for managing medium to long-term flood control in Korea. Therefore, clustering was performed in this study for 162 Korean regions according to the similarity of their flood control characteristics, which were evaluated by the indicators of the Potential Flood Damage (PFD) suggested by the National Water Resource Plan; a 19-year database (2000-2018) for the indicators was constructed. The t-distributed Stochastic Neighboring Embedding (t-SNE) technique, which can effectively reduce and classify massive data, was utilized. Then, the flood control characteristics for each region were derived and clustered according to individual and integrated indicators. As a result, it was found that regions of similar size and type generally consisted of the same cluster. However, because some regions were composed of the same cluster regardless of size and type, a more detailed analysis is required for each indicator of flood control characteristics, and further research should be conducted to explore the direct and indirect factors affecting the indicators.

Key words : Flood Control Characteristics, Clustering, Machine Learning, t-SNE

요 지

본 연구에서는 효율적이고 일괄적인 치수 관리를 도모하기 위하여 치수 특성에 따른 지역 군집화 제고 방안을 마련하고자 하였다. 치수 관리는 수해로부터 인명과 재산을 사전에 예방하고 보호하기 위한 구조적·비구조적 대책이 요구되며 이를 위해서는 지역별 치수 특성을 파악하는 것이 선행되어야 한다. 특히, 유사한 치수 특성에 기반하여 지역별로 군집화하는 것은 국내의 중장기적인 치수 관리의 전략 수립에 있어서 올바른 방향을 제시해 줄 수 있다. 이에 본 연구에서는 우리나라의 162개 시군을 대상으로 치수 특성의 유사도에 따른 군집화를 수행하였다. 치수 특성 평가 지표는 수자원장기종합계획에서 제시하는 홍수피해잠재능(Potential Flood Damage, PFD)내의 지표를 이용하고, 19개년(2000~2018)에 걸쳐 지표의 DB를 구축하였다. 군집화 기법은 방대한 자료를 효과적으로 축소하고 분류할 수 있는 머신 러닝의 t-distributed Stochastic Neighboring Embedding (t-SNE) 기법을 활용하였으며, 개별 지표 및 통합지표에 따라 지역별 치수 특성을 도출하고 군집화 하였다. 그 결과, 전반적으로 비슷한 규모와 입지 유형을 갖는 지역들이 동일한 군집으로 구성되는 것으로 나타났다. 하지만, 일부 지역은 규모 및 유형과 관계없이 동일한 군집으로 구성되었기 때문에 치수 특성 지표별로 추가적인 세밀한 분석이 필요하며, 지표에 영향을 미치는 직·간접적인 요인을 탐색하는 연구 또한 이루어져야 할 것으로 판단된다.

핵심용어 : 치수 특성, 군집화, 머신러닝, t-SNE

*정회원, 한국건설기술연구원 건설인증센터 박사 후 연구원(E-mail: engineer1026@kict.re.kr)

Member, Postdoctoral Researcher, Construction Certification Center, KICT

**정회원, 인하대학교 사회인프라공학과 박사과정(E-mail: kjs0308@naver.com)

Member, Ph.D Candidate, Department of Civil Engineering, Inha University

***교신저자, 정회원, 낙동강 홍수통제소 예보통제과 시설연구사(Tel: +82-51-603-3326, Fax: +82-51-201-3494, E-mail: firstsword@korea.kr)

Corresponding Author, Member, Senior Officer, Nakdong Flood Control Office, Ministry of Land, Infrastructure and Transport

****정회원, 국토연구원 국토환경·자원연구본부 책임연구원(E-mail: kjs0308@naver.com)

Member, Associate Research Fellow, Department of National Territorial Environment & Resources Research Division, KRIHS

*****정회원, 인하대학교 수자원시스템연구소 박사 후 연구원(E-mail: jungjw89@gmail.com)

Member, Postdoctoral Researcher, Institute of Water Resource System, Inha University

*****정회원, 인하대학교 사회인프라공학과 교수(E-mail: sookim@inha.ac.kr)

Member, Professor, Department of Civil Engineering, Inha University

1. 서 론

기후변화와 도시화로 인해 전 세계적으로 극한 호우의 발생 빈도가 증가하고 있어 치수 대비의 중요성을 인식시키고 있으며 많은 지역 사회는 효율적인 치수 관리를 위한 방안을 강구하고 있다. 우리나라의 경우 최근 10년간 (2009~2018) 발생한 자연 재난 중 홍수 피해는 약 3조 5,000억 원이며, 이는 전체 재해 피해 비중의 88.3%를 차지하고 있다 (MOIS, 2018). 산발적으로 발생하는 홍수 피해를 저감하기 위해서는 지역 현황과 실정에 맞는 치수 관리가 이루어져야 한다. 특히, 지역별 치수 특성의 유사성을 고려한 일괄적인 관리는 국내의 중장기적인 치수 관리의 전략 수립에 있어서 중요한 요소가 될 수 있다.

치수 특성이란 어떠한 단위 구역 내의 홍수 특성을 의미하며 수문-지형, 사회-경제, 홍수 방어, 기상 등 여러 요인에 의하여 결정된다. 먼저, 치수 특성을 평가하는 연구를 살펴보면, 국내의 경우 Choi and Kim (2006)은 인명, 재산, 침수면적에 대한 피해밀도의 자료를 이용하여 전국의 치수 특성을 제시하였고, Park et al. (2009)은 과거 홍수피해자료를 이용하여 홍수피해 발생빈도와 홍수피해액의 무차원 상관관계 해석을 통하여 전국의 치수피해특성을 분석하였다. Jang and Kim (2009)은 홍수피해위험 저감 및 예측을 위해 한강유역을 대상으로 다수의 강우특성 변수들과 홍수피해액에 대한 상관성을 분석하였다. Yeo et al. (2011)은 홍수 위험에 영향을 미치는 3가지 요인과 저감성에 영향을 미치는 2가지 요인들을 먼저 선정하고, 이를 대표할 수 있는 12개 인자를 선정한 후 치수 특성을 파악할 수 있는 중첩 지역안전도평가 모형을 제안하였다. Hong et al. (2017)은 홍수피해잠재능 (Potential Flood Damage, PFD) 및 홍수 위험도 지표(Flood Risk Assessment, FRA)의 개선안을 제시하여 한강 중권역에 대한 치수 특성을 정의한 바 있다. 또한, Joo et al. (2019a)은 6개의 기존 홍수 위험도 지수를 통해 대표 지표를 선정하고, 베이지안 네트워크를 이용하여 통합홍수위험평가지수 (Integrated index for Flood Risk Assessment, InFRA)를 개발한 바 있다.

국외의 경우, FEMA (2003)와 NOAA (2007)에서는 홍수 취약성 및 치수 특성 평가 프로그램을 개발하여 홍수 피해 규모를 추정하여 활용하고 있다. Munich Re Group (2004)에서는 자연적, 기술적, 사회 정치적, 경제적 측면에서의 4가지 요인에 대한 홍수 발생 시나리오와 직간접적 피해결과를 산정기준으로 평가한 후 치수 특성을 4단계로 분류한바 있다. 또한, Brooks (2003)는 치수 특성에 따라 사회적, 생물학적 취약성을 구분하고 취약성별 적응력 향상을 위한 기본 프레임워크를 제시하였다. Rygel et al. (2006)은 치수 특성 평가에서 가장 중요한 것은 적합한 지표들을 선정하는 것에 있음을 주장하였다. 따라서, 취약성 노출과 사회학적 요인으로 구분하여 지표를 수집하고 Pareto ranking process를 이용하여

치수 특성을 평가하는 방법을 제안하였다. Chang and Huang (2015)은 타이완의 도시지역을 대상으로 잠재영향 지표 (Potential Index, PI)를 선정하고, 이를 적응능력 지표 (Adaptation Index, AI)와 종합하여 평가하였다. 그리고 Kablan et al. (2017)에서는 기후변화 취약성 평가이론을 기본으로 재난위험관리 및 기후변화적응과 관련된 대리변수를 통한 노출성, 민감도, 적응능력의 3가지 평가지표 (Exposure, Sensitivity, Adaptation)를 이용하여 치수 평가 모형을 개발하였다.

한편, 앞서 언급했듯이 일괄적인 치수 관리를 위해서는 지역별 치수 특성의 유사성에 기반한 군집화가 필요하며, 이에 대한 방법론으로 대부분 군집 분석(Cluster Analysis)을 사용하고 있다. 통계 분석의 군집 분석은 유사성을 지니는 대상들을 모아서 하나의 군집으로 구성하는 분석으로 각 군집의 특성을 획일화할 수 있다. 치수 측면을 포함한 수문 분야에서의 군집 분석은 시계열 자료에 기반하고 있다(Unal et al., 2003; Jingyi and Hall, 2004; Lin and Chen, 2005; Ramachandra Rao and Srinivas, 2006; Kyoung et al., 2007; Isik and Singh, 2008; Kahya et al., 2008; Ouarda et al., 2008; Rhee et al., 2008; Ouyang et al., 2010; Corduas, 2011; Tercek et al., 2012; Iyigun et al., 2013; Archfield et al., 2014; Boscarello et al., 2015; Kuentz et al., 2017). 이들의 연구는 대부분 비슷한 군집끼리 결합해가면서 최종적으로 하나의 그룹이 될 때까지 군집을 수행하는 계층적 군집화 (Hierarchical Clustering)를 이용하여 연구를 진행하였다.

군집 분석을 위한 많은 기법이 존재하며 대표적인 것으로는 k-평균(k-means), 계층적(Hierarchical), 함수적(Functional), 가우시안 혼합 모델(Gaussian Mixture Model) 군집 분석 등이 존재한다. 이들 연구에서 사용한 군집 분석은 군집의 개수를 미리 결정을 해야 하기 때문에 군집 개수를 결정하기 위해서 사전실험을 해야 한다는 단점이 있다. 뿐만 아니라, 시계열 데이터를 요인분석(Factor Analysis)으로 차원축소한 다음 독립적으로 군집 분석을 수행하기 때문에 차원축소 방법과 군집 분석 사이의 상관관계에 대한 분석이 필요하다 (Min, 2018).

따라서, 본 연구에서는 치수 특성을 결정짓는 지표를 고려한 지역별 군집화 제고 방안에 초점을 두었다. 이를 위해 먼저 홍수피해잠재능을 구성하는 지표를 이용하여 우리나라 162개 시군을 대상으로 지표별 DB를 구축하였다. 많은 자료를 처리하기 위해 머신러닝의 차원감소법의 종류인 t-Stochastic Neighbour Embedding (t-SNE) 기법을 이용하여 지역별-지표별의 유사성에 따른 군집화를 수행하고 비교 및 검토하였다. 또한, 모든 지표를 동시에 고려하여 지역별 군집화를 수행하고 우리나라의 전반적인 치수 특성에 대하여 고찰하였다.

2. 지역별 치수 특성에 따른 군집화 방법론

2.1 치수 특성 평가를 위한 지표 선정

우리나라는 2000년대에 들어서 치수 특성에 대한 평가의 중요성이 요구되었으며, 그 이후로 홍수 취약성 및 위험도 모형을 통해 지역별 치수 특성을 분석하는 연구가 여러 기관에서 진행되어 왔다(MLTM, 2001, 2010; KRIHS, 2005; SI, 2006; KEI, 2008; KICT, 2009; NDMI, 2011; JTHINK, 2010). 그 중, 수자원 최상위 계획인 수자원장기종합계획(MLTM, 2001, 2010)에서 제시하는 홍수피해잠재능(Potential Flood Damage, PFD)과 홍수 위험도 지표(Flood Risk Index, FRI)가 있으며, 본 연구에서는 홍수피해잠재능을 구성하는 지표를 선정하여 지역별로 치수 특성을 군집화하였다.

홍수피해잠재능은 치수 특성을 파악하기 위해 개발된 방법이며 치수단위구역의 잠재적인 홍수피해 취약 정도를 나타내는 지수다. 홍수방어계획 수립에 있어 중요한 기준이 되는 항목은 그 지역의 잠재 피해 정도 및 발생확률과 이에 대한 방어능력으로 구분할 수 있다(Joo et al., 2018a). 따라서, 홍수피해잠재능은 이에 입각한 홍수취약성 평가 방법이며 크게 잠재성 요소와 위험성 요소로 나누어 항목을 산정하게 된다(Table 1).

Table 1. Components of PFD

Components	Sub-Components	
Potential, F_{PO}		• Population Density, F_{PD}
		• Property, F_{PP}
		• Urbanization, F_{URB}
		• Social Overhead Capital (Railroad, Bridge, Road), F_{SOC}
Risk, F_{RI}	Likelihood	• Flood Damage Cost, F_{FDC}
		• Probability Rainfall (100 year), F_{PR}
	Protection	• River improvements, F_{RI}
		• Dam and Reservoir, F_{DAM}
		• Pumping Station, F_{PUMP}

잠재성 요소는 동일조건(위치, 표고, 강우량 등)하에서 홍수가 발생하였을 때 발생할 수 있는 잠재적인 피해 정도를 나타내는 요소로 인구밀도, 자산, 도시화율, 사회기반시설 등의 지표로 구성된다. 그리고, 위험성 요소는 어떤 지역에 대해 홍수피해가 발생할 가능성 및 홍수피해 발생시 이를 저지할 수 있는 방어능력 정도를 나타내며 홍수피해액, 확률 강우량, 하천 개수율, 댐 및 저수지, 배수펌프장 현황으로 구성된다(Table 1).

2.2 머신러닝의 t-SNE 기법

일반적으로 치수 특성의 경향성을 파악하기 위해 다년도 의 자료가 요구되며 본 연구의 목적인 치수 특성에 의한 지역별 군집화도 마찬가지다. 따라서, 각 지표의 자료는 매우 방대하며 고차원의 구조를 이룬다. 이러한 자료를 축소 하고 군집하기 위해서는 머신러닝의 비지도 학습(Unsupervised Learning)이 요구되며, 본 연구에서는 고차원의 자료를 효과적으로 축소할 수 있는 t-분포 확률적 임베딩(t-distributed Stochastic Neighboring Embedding, t-SNE) 기법을 선정하였다. t-SNE는 차원 축소 및 군집 분류의 시각화에 강력한 성능을 지니고 있기 때문에 텍스트, 이미지, 음성 인식 및 생화학을 비롯한 의학 분야에서 널리 사용되고 있으며 여러 가지 사물을 최종적으로 2차원으로 시각화하여 표현한다. t-SNE의 기본 개념과 실제 적용 사례를 나타낸 것은 Figs. 1 및 2와 같다.

t-SNE에서 확률적 임베딩(Stochastic Neighbor Embedding, SNE)은 차원축소(Dimensionality Reduction) 기법으로 고차원에서의 거리와 축소된 차원에서의 데이터 사이의 거리 혹은 유사도가 호환이 되도록 고차원 데이터를 축소하는 것이다(Hinton and Roweis, 2002; van der Maaten and Hinton, 2008; Min, 2018). SNE에서는 유사도를 측정할 때, 유클리디안 거리(Euclidean Distance) 등의 일반적인 방식을 이용하지 않고 다음과 같이 고차원(p_{ji}) 및 저차원(q_{ji})을 조건부 확률로 유사도를 표현한다(Eqs. 1, 2).

$$p_{ji} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)} \quad (1)$$

$$q_{ji} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)} \quad (2)$$

여기서 x_i 와 x_j 는 고차원 데이터, y_i 와 y_j 는 저차원 데이터, σ_j 는 x_i 를 중심으로 하는 정규 분포의 표준편차를 말한다. 그 다음 단계로 저차원 공간상에서의 확률 분포가 고차원 공간상에서의 확률 분포를 잘 대변하는지를 판단하기 위해 Kullback (1959)이 제시한 쿨백-라이블러 발산(Kullback-Leibler divergence) 식을 사용한다(Eq. 3).

$$H(p, q) - H(p) \quad (3)$$

여기서 $H(p)$ 는 원래 분포가 가지는 엔트로피, $H(p, q)$ 는 p 대신 q 를 사용할 때의 교차 엔트로피(cross entropy)를 의미하며 본래의 엔트로피 식으로 유도하면 Eq. 4와 같다.

$$C = \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{ji} \log \frac{p_{ji}}{q_{ji}} \quad (4)$$

Fig. 1. t-SNE concept (van der Maaten, 2018)

Fig. 2. Text recognition using t-SNE technique (Derksen, 2019)

P_i 는 주어진 대상 x_i 에 대한 나머지 모든 대상들의 조건부 확률을 의미하며, Q_i 는 주어진 대상 y_i 에 대한 나머지 모든 대상들의 조건부 확률 분포를 의미한다(Maaten and Hinton, 2008; Joo et al., 2019b, 2019c). 즉, 저차원과 고차원이 비슷할수록 0에 가까운 값을 가지게 된다. 하지만 SNE는 비대칭적인 조건부 확률을 사용한다는 점과 충분히 이격된 고차원 데이터를 저차원에서는 구현하지 못하는 과밀 문제(Crowding Problem)를 갖고 있다. 따라서, 이에 대한 대안으로 SNE의 조건부 확률 p_{ji} 와 q_{ji} 를 일부 수정한 최종적인 t-SNE는 아래와 같이 나타낼 수 있으며, Fig. 1에서 제시한 고차원을 수정된 조건부 확률로 표현한 것임을 알 수 있다 (Eqs. 5, 6).

$$p_{ij} = \frac{p_{ji} + p_{ilj}}{2} \quad (5)$$

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_i - y_l\|^2)^{-1}} \quad (6)$$

t-SNE는 저차원에서 자료의 특징을 잘 구분해 주기 위해 분포의 범위를 높여주는 스튜던트 t 분포(Student's t-distribution)를 유사도로 사용하며, SNE와 마찬가지로 쿨백-라이블러 발산을 최소화하는 방식으로 진행된다.

3. 치수 특성 지표의 자료 구축

3.1 치수 특성 지표의 DB 구축

치수 특성 지표의 정확한 군집을 위해서는 일정한 입지 유형이 아닌 다양한 입지 유형의 지표 DB가 요구된다. 또한, 우리나라의 치수를 비롯한 방재에 대한 계획 및 투자 대책은 행정구역별로 이루어지고 있다(Joo et al., 2018b). 이를 고려하여 본 연구에서는 우리나라의 전국 162개 시군을 대상으로 치수 특성 지표의 DB를 구축하였다. 군집 분석을 위한 연기간으로는 다년도 자료를 고려하기 위해 19개년(2000~2018)의 자료를 이용하였다. 즉, 162개 시군, 9개 지표, 19개년을 고려한 자료의 총 개수는 27,702개가 된다($162 \times 9 \times 19 = 27,702$).

각 지표의 DB 구축 방식을 살펴보면 먼저 잠재성 요소는

Table 2. DB Construction for Flood Control Characteristic Indicators (ex. Population Density)

District	2000	2001	2002	...	...	2018
Seoul	16,965	16,886	16,794	...	...	16,034
Busan	4,909	4,875	4,823	...	...	4,416
Daegu	2,855	2,856	2,856	...	...	2,773
⋮	...	...	...	...	...	...
⋮	...	...	...	...	...	...
Jeju	385	391	397	...	...	353
Seogwipo	187	186	183	...	...	2,165

Table 3. Results of Pre-processing for DB of Flood Control Characteristic Indicators (ex. Population Density)

District	2000	2001	2002	...	...	2018
Seoul	0.9999991	0.9999992	0.9999992	...	...	0.9999997
Busan	0.9167885	0.9289658	0.9313488	...	...	0.9781425
Daegu	0.7960962	0.8034125	0.8136808	...	...	0.9020415
⋮	...	...	...	...	...	...
⋮	...	...	...	...	...	...
Jeju	0.4297843	0.4368137	0.4368137	...	...	0.6129335
Seogwipo	0.376265	0.3945574	0.3945574	...	...	0.5234198

국가통계포털(Korean Statistical Information Service, KOSIS) 및 지자체 통계 연보를 통해 구축할 수 있다. 인구밀도는 치수 특성 평가에서 자주 쓰이고 있는 지표이며 인구 밀도가 높은 지역일수록 인명피해가 높고 자산, 사회기반시설은 홍수 발생시 홍수 피해 규모 측면에서 좀 더 취약하다고 가정할 수 있다. 도시화율은 수치가 높을수록 불투수율의 증가로 홍수 피해가 증가할 것이며 시군별 전체 면적에 도시 면적을 고려하여 산정하였다.

위험성 요소의 홍수피해액은 재해연보상의 11개의 재해 유형 중에 홍수와 관련된 태풍, 호우, 호우태풍, 태풍호우 4개 유형의 피해액을 고려하였으며 2018년도 소비자물가지수를 반영하여 환산한 후 적용하였다. 확률강우량은 전국 65개 관측소를 대상으로 기준년도 최근 30년의 강우량을 이용하였으며, 강우 티센망을 작성한 후 행정구역별로 유도하였다. 개수율, 댐 및 저수지 지수, 배수펌프장 개소는 KOSIS, 도시기본계획, 하천기본계획 등을 참고하여 산정하였다. Table 2는 치수 특성 지표 중 인구 밀도 지표의 DB 구축 결과를 요약하여 나타낸 것이다.

3.2 DB의 전처리

치수 특성 지표의 DB를 구축할 때 통계 조사 착오 및 측정 기기 오류 등으로 인해 이상치를 비롯한 모순된 결과가 기록될 수 있다. 또한, 시군별, 연도별, 지표별에 따른 정량적인 수치의 심한 변동폭 및 편차로 인해 분석 결과가 왜곡될 수 있다. 예를 들어 홍수피해액은 0에서 1,000억 단위를 가지고 있고 도시화율은 0에서 100단위를 갖는다. 따라서

분석의 목적과 무관한 요소를 미리 데이터에서 제거하고 수치의 큰 편차가 분석 결과에 영향을 미치는 것을 방지하기 위해 전처리가 수행되어야 한다. 전처리 기법 중에서 정규화 과정은 데이터를 동일한 규모로 처리할 수 있다는 장점이 있기 때문에 머신러닝 연구에서 널리 사용되고 있다. 따라서 본 논문에서는 모든 시군을 포함하는 각 지표별 전체 연도의 평균과 표준편차를 이용하여 정규화 과정을 수행하였다 (Table 3).

전처리 후 최종적인 DB 구축 결과를 보면 잠재성 요소의 지표는 대도시, 중소도시 및 일부 촌락 지역에서 높은 경향을 보였으나 위험성 요소의 지표는 지역 유형에 상관없는 결과가 도출되었다. 하지만 전처리를 수행함에도 불구하고 지역간의 유사성은 다소 불확실하고 명확한 특성을 찾기가 어려웠다. 이는 다년도에 걸쳐 많은 자료가 혼재에 있고 전처리 과정을 상쇄할만큼 각각의 수치적인 편차가 큰 것에 기인한다.

4. t-SNE를 이용한 지역별 치수 특성 군집화

4.1 지표별-지역별 군집화

본 연구에서는 머신 러닝 기법 중 t-SNE 분석을 이용하여 시군별로 군집분석을 수행하였다. 치수 특성 지표는 총 19개 년으로 구성되어 있기 때문에 차원 측면에서는 19차원을 의미한다. t-SNE의 시각화를 위해서는 2차원(vector X, vector Y)의 지도로 축소해야하며 주요 매개변수인 혼란도 (Perplexity), 학습율(Learning Rate), 최적화를 위한 반복횟수(Number of Iteration)를 정의해야 한다. 혼란도는 일반적

Fig 3. Clustering by Indicator-region using t-SNE

으로 5에서 50사이를 권고하기 때문에 본 연구에서는 그 중에서도 일반적으로 적용하는 30으로 설정하고, 학습율은 200, 반복횟수는 1,000회로 설정하였다(Maaten and Hilton, 2008). Fig. 3은 t-SNE에 의한 각 지표별 군집 결과를 시각화한 것이며, Fig. 4는 군집 결과를 우리나라 지역별로 매핑(Mapping)하여 보여준 것이다.

2차원으로 시각화한 결과, 9개의 치수 특성 지표는 지역별로 군집화가 적절하게 이루어지는 것으로 나타났으며 각 지표는 총 10개의 군집으로 구성되었다. 군집내의 자료들은 전반적으로 유사성이 높음에 따라 밀도가 높은 결과로 나타났다. 인구밀도와 댐 및 저수지 지표 등에서 나타나는 일부 선형 형태의 군집은 vector X(고차원 및 실제 차원)와 vector Y(저차원 및 축소된 차원)에서 다소 차이가 있음을 알 수 있다. 즉, 선형 군집의 형태에 따라서 횡 방향은 차원이 축소될 때 유사성이 높아지고, 종 방향은 차원이 축소될 때 유사성이 상대적으로 낮아지는 것을 의미한다. 배수펌프장 지표에서는 같은 군집임에도 불구하고 거리가 다소 이격

되고 고차원과 저차원의 범위가 각각 넓은 것으로 나타났다. 이는 실제 자료간의 수치적인 편차가 큰 것에 기인하며 실제로 지역별 배수펌프장 개수의 편차가 크다는 결론에 도달할 수 있다.

다만, t-SNE에서 동일한 군집 안에서의 자료의 거리는 서로의 유사성을 확인하는 척도가 될 수 있으나, 군집간의 거리 정도는 유사성의 척도로 활용될 수 없다. 예를 들어, 군집 A가 B보다 상대적으로 C와 근접하다고 해서 군집 A와 C가 유사성이 좀 더 있다고 할 수 없다는 것이다. t-SNE 결과는 혼란도의 수치에 따라 군집간의 거리가 변할 수 있으며, Neighbour Embedding (NE)의 개념에 따라 실제 거리와는 상이할 수 있다. 따라서, t-SNE는 어디까지나 지역적 유사성을 강화한 명확한 군집 구성에 초점을 두고 있기 때문에 이를 고려한 해석이 필요할 것이다.

군집에 따른 지역별 매핑 결과를 보면, 먼저 인구밀도는 대도시, 중소도시 및 촌락지역의 군집화가 뚜렷하게 나뉘는 양극화 현상을 보였으며 현재 우리나라의 인구 밀집 현황을

Fig 4. Clustering Map by Indicator-region using t-SNE

보면 타당한 결과라 판단된다. 자산, 도시화율, 사회기반시설의 전반적인 군집 양상은 인구 밀도와 비슷하나 군집간 비율의 편차는 다소 낮은 것으로 나타났다. 특히, 사회기반시설은 지역 유형이 상이함에도 불구하고 동일한 군집으로 구성되는 경우가 다수 존재했다(울산, 제주 등). 이는 지역 유형과는 별개로 지역의 특색을 고려한 여러 가지 공업, 생활, 관광 인프라 시설이 위치함에 기인된 것으로 판단된다.

홍수 피해액은 대도시, 제주도를 포함한 해안 도시, 강원도 일부 지역이 동일한 군집으로 형성되었다. 해당 지역의 특징은 실제로 홍수 피해가 많은 지역으로 각각 대도시의 높은 자산, 해안 도시의 복합 범람, 강원도 일부 지역의 도시

계획 미흡 등에 따른 결과인 것으로 판단된다. 확률 강수량은 주로 해안 도시가 동일한 군집으로 형성되었으며 해당 지역은 기온이 높고 저기압이 발달함에 따라 많은 강수량이 발생하는 특징이 있다. 하천 정비율의 군집 정도는 지역별로 다소 불확실하고 명확한 특성을 찾기가 어려웠다. 마지막으로 댐 및 저수지 지수 및 배수 펌프장 개소는 162개 지역 중 각각 46%, 41%에 해당하는 지역이 0에 근접하게 산정됨에 따라 동일한 군집으로 형성되었다. 앞선 분석 내용을 토대로 지표별-군집별에 따른 지역 개수 및 비율은 Table 4와 같으며, 여기서 (a)-(i)는 각 치수 특성 지표를 의미한다 (Fig. 4 참조).

Table 4. Number and Proportion of Clustered Regions according to Each Indicator-region

Cluster	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
	Number of regions (%)								
1	2 (1)	1 (1)	2 (1)	4 (2)	16 (10)	26 (16)	16 (10)	15 (9)	1 (1)
2	5 (3)	2 (1)	8 (5)	6 (4)	4 (2)	25 (15)	28 (17)	22 (14)	2 (1)
3	9 (6)	12 (7)	8 (5)	35 (22)	43 (27)	10 (6)	16 (10)	14 (9)	14 (9)
4	9 (6)	9 (6)	7 (4)	6 (4)	26 (16)	17 (10)	29 (18)	74 (46)	15 (9)
5	38 (23)	16 (10)	30 (19)	14 (9)	33 (20)	18 (11)	20 (12)	10 (6)	4 (2)
6	3 (2)	9 (6)	6 (4)	34 (21)	20 (12)	12 (7)	14 (9)	9 (6)	15 (9)
7	2 (1)	27 (17)	17 (10)	16 (10)	12 (7)	17 (10)	16 (10)	4 (2)	67 (41)
8	4 (2)	31 (19)	22 (14)	13 (8)	1 (1)	12 (7)	11 (7)	10 (6)	9 (6)
9	25 (15)	22 (14)	27 (17)	9 (6)	2 (1)	19 (12)	11 (7)	3 (2)	21 (13)
10	65 (40)	33 (20)	35 (22)	25 (15)	5 (3)	6 (4)	1 (1)	1 (1)	14 (9)

Fig 5. Regional Clustering for Integrated Indicators using t-SNE

4.2 통합 지표를 반영한 지역별 군집화

본 연구에서는 선정한 치수 특성 지표를 모두 반영하여 군집화를 수행하였다. 9개의 치수 특성 지표는 각각 2차원이며, 이를 모두 고려한 통합 지표는 총 18차원으로 구성될 수 있다. t-SNE 기법을 적용하기 위한 매개변수는 4.1절의 군집화 과정과 동일하게 설정하였으며, 분석 결과를 2차원 그래프와 우리나라 지역별 지도에 시각화하였다(Figs. 5, 6).

통합 지표의 군집화 결과, 9개를 통합한 치수 특성 지표는 지역별로 군집화가 적절하게 이루어지는 것으로 나타났으며 총 10개의 군집으로 구성되었다. 지표별로 군집을 수행할 때보다 자료의 유사성이 다소 낮음에 따라 군집내 밀도 또한 낮은 결과로 나타났다. 하지만 고차원 및 저차원의 범위는 다소 좁은 것으로 분석됨에 따라 자료 간의 편차 범위는 크지 않는 것으로 확인되었다. 이는 지표별-지역별 군집 분석 단계에서 1차적으로 차원 축소가 선행되었기

때문인 것으로 판단된다.

매핑된 결과를 통한 각 군집에서의 주요 지역을 보면, 군집 1은 서울시만 포함되는 것으로 나타났다. 따라서, 서울시는 다른 지역과는 달리 고유의 특성에 맞는 치수 관리가 이루어져야 할 것이다. 군집 2는 6대 광역시를 포함한 제주시, 청주시 등으로 구성되었으며, 군집 3, 9, 10은 주로 춘천도시로 구성되었다. 군집 4는 수도권 내의 중·대규모 도시(수원시, 성남시, 안양시 등), 군집 5는 경기도 동부(가평, 양평 등)와 강원도 일부 지역(춘천 등)으로 구성되었다. 또한, 군집 6은 충청도 내의 중소도시(아산시, 서산시, 당진시, 계룡시 등), 군집 7은 경기 북부 및 남부 지역(화성시, 양주시, 용인시 등), 군집 8은 주로 중소도시(공주시, 익산시 등)와 포항시로 구성되었다(Table 5).

통합 지표에 의한 지역별 군집 결과는 비슷한 규모와 유형을 가진 지역들이 대부분 동일한 군집으로 구성되었다.

Fig 6. Regional Clustering Map for Integrated Indicators using t-SNE

Table 5. Number and Proportion of Clustered Regions according to Integrated Flood Control Characteristic Indicators

Cluster	Number of Regions (%) for Integrated Index	Major Region
1	1 (1)	Seoul
2	9 (6)	6 metropolitan cities, Jeju, Cheongju et al.
3	24 (15)	Rural area
4	11 (7)	Suwon, Seongnam, Anyang et al.
5	17 (10)	Gapyeong, Yangpyeong, Chuncheon et al.
6	29 (18)	Rural area
7	15 (9)	hwaseong, Yangju, Yongin et al.
8	13 (8)	Gongju, Iksan et al.
9	9 (6)	Rural area
10	34 (21)	Rural area

이는 규모 및 유형에 따라서 여타 지역적 특성이 부합하는 경우가 많고 사회과학 및 자연과학 지표로 이루어진 치수 특성 지표에도 전이될 확률이 높은 것에 기인한다. 하지만 일부 군집은 지역 규모와 유형적인 측면으로만 해석하기에는 다소 부족한 것으로 나타났다. 예를 들어, 군집 3에 세종시와 신안군, 군집 7은 용인시와 철원군이 같이 포함되어 있다. 따라서, 지역별 치수 특성 지표 자료에 의거한 세밀한 분석이 이루어짐과 동시에 지표에 영향을 미치는 직·간접적인 요인을 탐색하는 등의 다각도적인 연구가 추가적으로 필요

하다고 판단된다.

일괄적인 치수 관리를 도모하고 거시적인 치수 관리의 전략 수립을 위해서는 치수 특성에 따른 군집화의 정확성 제고가 필요할 것이다. 따라서, 본 연구는 효율적인 지역 치수 관리를 위한 참고 자료로 참고될 수 있을 것이다. 향후, 여러 가지 과학적인 방법론을 적용하여 지역별 치수 특성을 군집화하고 이들의 분석 기법들을 비교 및 검토하는 연구가 계속적으로 이루어져야 할 것이다.

5. 결 론

본 연구에서는 효율적인 치수 관리를 위하여 치수 특성에 따라 우리나라 162개 시군을 대상으로 군집 분석을 수행하고 지표별 및 통합지표에 의한 지역별 군집 결과를 고찰하였다. 치수 특성을 평가하기 위해 홍수피해잠재능을 구성하는 9개의 지표를 선정하였으며, 군집 분석을 위한 기법으로는 방대한 자료를 효과적으로 축소하고 분류할 수 있는 머신러닝의 t-SNE 기법을 이용하였다. 치수 특성 지표의 DB는 19개년(2000~2018) 기간의 자료로 구성하고 전처리를 수행하였다. t-SNE에 의하여 지역별 군집이 구성되었으며, 본 연구의 주요 결과를 요약하면 다음과 같다.

- (1) 지표별-지역별 군집화 결과, 인구 밀도, 자산, 도시화율은 도시와 비도시간의 군집 양상의 양극화를 보였으나 사회기반시설은 도시와 비도시가 동일한 군집으로 구성되는 경우가 많았다. 이는 지역 유형과는 별개로 지역의 특색을 고려한 여러 가지 공업, 생활, 관광 인프라 시설이 위치하여 나타난 결과라 판단된다. 전반적으로 지표는 지역 유형에 맞게 군집이 구성되었으나, 하천 개수율 지표는 다소 불확실하고 명확한 특성을 찾기가 어려웠다.
- (2) 통합 지표를 반영한 지역별 군집화 결과, 전반적으로 비슷한 규모와 유형을 가진 지역들이 동일한 군집으로 구성되었다. 이는 규모 및 유형에 따라서 여타 지역적 특성이 부합하는 경우가 많고 사회과학 및 자연과학 지표로 이루어진 치수 특성 지표에도 전이될 확률이 높기 때문일 것이다. 하지만 세종시와 신안군, 용인시와 철원군이 각각 동일한 군집으로 포함된 것으로 미루어보아 지역 규모와 유형적인 측면으로만 군집 양상을 단정짓기에는 어려운 것으로 나타났다. 따라서, 치수 특성 지표별로 세밀한 분석이 이루어져야 하며, 지표에 영향을 미치는 직·간접적인 요인을 탐색하는 연구가 추가적으로 필요할 것으로 판단된다.
- (3) 일괄적인 치수 관리를 도모하고 거시적인 치수 관리의 전략 수립을 위해서는 치수 특성에 따른 군집화의 정확성 제고가 선행되어야 한다. 따라서, 본 연구는 효율적인 지역 치수 관리를 위한 참고 자료로 참고될 수 있을

것이다. 향후, 다양한 군집 방법을 도입하여 치수 특성에 따라 군집화하고, 상이한 군집 결과를 비교 및 검토하는 연구를 통해 치수 관리를 도모해야 할 것이다.

감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2017R1A2B3005695).

References

- Archfield, S.A., Kennen, J.G., Carlisle, D.M., and Wolock, D.M. (2014). An objective and parsimonious approach for classifying natural flow regimes at a continental scale. *River Res. Appl.*, Vol. 30, No. 9, pp. 1166-1183.
- Boscarello, L., Ravazzani, G., Cislighi, A., and Mancini, M. (2015). Regionalization of flow-duration curves through catchment classification with streamflow signatures and physiographic-climate indices. *J. Hydrol. Eng.*, Vol. 21, No. 3. doi:10.1061/(ASCE)HE.1943-5584.0001307
- Brooks, N. (2003). *Vulnerability, risk and adaptation: A conceptual framework*. Tyndall Centre Working Paper No. 38.
- Chang, L.F., and Huang, S.L. (2015). Assessing urban flooding vulnerability with an emergy approach. *Landsc. Urban. Plan.*, Vol. 143, pp. 11-24.
- Choi, C.I., and Kim, W. (2006). Implications of the flood risk assessment in urban and regional planning. *Journal of the Korea Planners Association*, Vol. 41, No. 4, pp. 143-155.
- Corduas, M. (2011). Clustering streamflow time series for regional classification. *J. Hydrol.*, Vol. 407, No. 1-4, pp. 73-80.
- Derksen, L. (2019). *Visualising high-dimensional datasets using PCA and t-SNE in Python*. Retrieved from <https://towardsdatascience.com>
- Federal Emergency Management Agency (FEMA). (2003). *Flood information tool user manual (Rev. 7)*.
- Hinton, G.E., and Roweis, S.T. (2002). Stochastic neighbor embedding. *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 833-840.
- Hong, S.J., Joo, H.J., and Kim, K.T. (2017). Study on the improvement method of flood risk assessment by flood damage area. *Journal of Wetlands Research*, Vol. 19, No. 4, pp. 459-469.
- Isik, S., and Singh, V.P. (2008). Hydrologic regionalization of watersheds in Turkey. *J. Hydrol. Eng.*, Vol. 13, No. 9, pp. 824-834.
- Iyigun, C., Türkeş, M., Batmaz, İ., Yozgatligil, C., Purutçuoğlu, V., Koç, E.K., et al. (2013). Clustering current climate regions of Turkey by using a multivariate statistical method. *Theor. Appl. Climatol.*, Vol. 114, No. 1-2, pp. 95-106.
- Jang, O.J., and Kim, Y.O. (2009). Flood risk estimation using regional regression analysis. *J. Korean Soc. Hazard Mitig.*, Vol. 9, No. 4, pp. 71-80.
- Jingyi, Z., and Hall, M.J. (2004). Regional flood frequency analysis for the Gan-Ming River basin in China. *J. Hydrol.*, Vol. 296, No. 1-4, pp. 98-117.
- Joo, H.J., Choi, C.H., Kim, J.W., Kim, D.H., Kim, S.J., and Kim, H.S. (2019a). A bayesian network-based integrated for flood risk assessment (InFRA). *Sustainability*, Vol. 11, No. 13, 3733. doi:10.3390/su11133733
- Joo, H.J., Jun, H.D., Lee, J.H., and Kim, H.S. (2019b). Assessment of a stream gauge network using upstream and downstream runoff characteristics and entropy. *Entropy*, Vol. 21, No. 7, 673. doi:10.3390/e21070673
- Joo, H.J., Kim, S.J., Lee, M.J., and Kim, H.S. (2018a). A study on determination of investment priority of flood control considering flood vulnerability. *J. Korean Soc. Hazard Mitig.*, Vol. 18, No. 2, pp. 417-429.
- Joo, H.J., Lee, J.H., Jun, H.D., Kim, K.T., Hong, S.J., Kim, J.W., et al. (2019c). Optimal stream gauge network design using entropy theory and importance of stream gauge stations. *Entropy*, Vol. 21, No. 10, 991. doi:10.3390/e21100991
- Joo, H.J., Lee, M.J., Choi, C.H., Kim, S.J., and Kim, H.S. (2018b). A study on the selection of representative indicators of flood vulnerability assessment. *J. Korean Soc. Hazard Mitig.*, Vol. 18, No. 6, pp. 335-346.
- Jeonbuk Development Institute (JTHINK). (2010). *Improving flood plans against climate change in Jeollabuk-do*. Jthink 2010-PR-10.
- Kablan, M.K.A., Dongo, K., and Coulibaly, M. (2017). Assessment of social vulnerability to flood in urban Côte d'Ivoire using the MOVE Framework. *Water*, Vol. 9, No. 4, 292. doi:10.3390/w9040292

- Kahya, E., Demirel, M.C., and Bég, O.A. (2008). Hydrologic homogeneous regions using monthly streamflow in Turkey. *Earth Sci. Res. J.*, Vol. 12, No. 2, pp. 181-193.
- Korea Environment Institute (KEI). (2008). *Development and introduction of indicators to assess vulnerability of climate change*.
- Korea Institute of Civil Engineering and Building Technology (KICT). (2009). *Strengthen facility standards against excess climate*.
- Korea Research Institute for Human Settlements (KRIHS). (2005). *Analysis of flood damage characteristics development of flood damage index*.
- Kuentz, A., Arheimer, B., Hundecha, Y., and Wagener, T. (2017). Understanding hydrologic variability across Europe through catchment classification. *Hydrol. Earth Syst. Sci.*, Vol. 21, pp. 2863-2879.
- Kullback, S. (1959). *Information theory and statistics*. John Wiley & Sons. Republished by Dover Publications in 1968; reprinted in 1978: ISBN 0-8446-5625-9.
- Kyoung, M.S., Kim, S.D., Kim, B.K., and Kim, H.S. (2007). Construction of hydrological drought severity-area-duration curves using cluster analysis. *J. Korean Soc. Civ. Eng.*, Vol. 27, No. 3B, pp. 267-276.
- Lin, G.F., and Chen, L.H. (2005). Time series forecasting by combining the radial basis function network and the self-organizing map. *Hydrol. Process.*, Vol. 19, No. 10, pp. 1925-1937.
- van der Maaten, L. (2018). Do's and don'ts of using t-SNE to understand vision models. *Interpretable Machine Learning for Computer Vision Workshop*.
- van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, Vol. 9, pp. 2579-2605.
- Min, Y.H. (2018). Cluster analysis of daily electricity demand with t-SNE. *Journal of the Korea Society of Computer and Information*, Vol. 23, No. 5, pp. 9-14.
- Ministry of Land, Transport and Maritime Affairs (MLTM). (2001). *National water resource plan*.
- Ministry of Land, Transport and Maritime Affairs (MLTM). (2010). *National water resource plan (2010-2020)*.
- Ministry of the Interior and Safety (MOIS). (2018). *Statistical year book of natural disaster*. Registration No. 11-1741000-000002-11.
- Munich Re Group. (2004). *Annual report 2004: Advancing innovation*.
- National Disaster Management Research Institute (NDMI). (2011). *Development of community-based flood disaster risk reduction index*.
- National Oceanic and Atmospheric Administration (NOAA). (2007). *Risk vulnerability assessment tool*.
- Ouarda, T.B.M.J., Bâ, K.M., Diaz-Delgado, C., Cârsteanu, A., Chokmani, K., Gingras, H., et al. (2008). Inter-comparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. *J. Hydrol.*, Vol. 348, No. 1-2, pp. 40-58.
- Ouyang, R., Ren, L., Cheng, W., and Zhou, C. (2010). Similarity search and pattern discovery in hydrological time series data mining. *Hydrol. Process.*, Vol. 24, No. 9, pp. 1198-1210.
- Park, T.S., Choi, M.H., Yeo, C.G., and Lee, S.O. (2009). Analysis of regional flood damage Characteristics using relationship between flood frequency and damages. *J. Korean Soc. Hazard Mitig.*, Vol. 9, No. 5, pp. 87-92.
- Ramachandra Rao, A., and Srinivas, V.V. (2006). Regionalization of watersheds by hybrid-cluster analysis. *J. Hydrol.*, Vol. 318, No. 1-4, pp. 37-56.
- Rhee, J., Im, J., Carbone, G.J., and Jensen, J.R. (2008). Delineation of climate regions using *in-situ* and remotely-sensed data for the Carolinas. *Remote Sens. Environ.*, Vol. 112, No. 6, pp. 3099-3111.
- Rygel, L., O'Sullivan, D., and Yarnal, B. (2006). A method for constructing a social vulnerability index: An application to hurricane storm surges in a developed country. *Mitigation and Adaptation Strategies for Global Change*, Vol. 11, pp. 741-764.
- Tercek, M.T., Gray, S.T., and Nicholson, C.M. (2012). Climate zone delineation: Evaluating approaches for use in natural resource management. *Environ. Manage.*, Vol. 49, No. 5, pp. 1076-1091.
- The Seoul Institute (SI). (2006). *Development of the regional safety assessment model in Seoul: Focusing on flood*.
- Unal, Y., Kindap, T., and Karaca, M. (2003). Redefining the climate zones of Turkey using cluster analysis. *Int. J. Climatol.*, Vol. 23, No. 9, pp. 1045-1055.
- Yeo, C.G., Seo, G.S., and Song, J.W. (2011). Regional safety assesment due to urban flood using GIS. *Journal of the Korean Association of Geographic Information Studies*, Vol. 14, No. 3, pp. 68-77.

Received	May 13, 2020
Revised	May 13, 2020
Accepted	May 29, 2020