Application of clustering algorithms for dimensionality reduction in infrastructure resilience prediction models

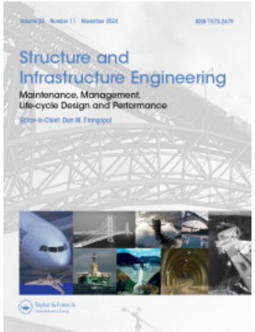
Balakrishnan, S.; Cassottana, Beatrice; Verma, Arun

DOI
10.1080/15732479.2024.2366958

Publication date
2024

Document Version
Final published version

Published in
Structure and Infrastructure Engineering



Taylor & Francis
Taylor & Francis Group

Structure and Infrastructure Engineering

Maintenance, Management, Life-Cycle Design and Performance

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/nsie20

Application of clustering algorithms for dimensionality reduction in infrastructure resilience prediction models

Srijith Balakrishnan, Beatrice Cassottana & Arun Verma

STRUCTURE AND INFRASTRUCTURE ENGINEERING
https://doi.org/10.1080/15732479.2024.2366958

Application of clustering algorithms for dimensionality reduction in infrastructure resilience prediction models

Srijith Balakrishnan[a], Beatrice Cassottana[b] and Arun Verma[c]

[a]Faculty of Technology, Policy and Management, Delft University of Technology, Delft, the Netherlands; [b]Future Resilient Systems, Singapore-ETH Centre at CREATE Singapore, Singapore; [c]School of Computing, National University of Singapore, Singapore

ABSTRACT

Recent studies increasingly adopt simulation-based machine learning (ML) models to analyse critical infrastructure system resilience. For realistic applications, these ML models consider the component-level characteristics that influence the network response during emergencies. However, such an approach could result in a large number of features and cause ML models to suffer from the 'curse of dimensionality'. A clustering-based method is presented that simultaneously minimises the problem of high-dimensionality and improves the prediction accuracy of ML models developed for resilience analysis in large-scale interdependent infrastructure networks. The methodology has three parts: (a) generation of simulation dataset, (b) network component clustering, and (c) dimensionality reduction and development of prediction models. First, an interdependent infrastructure simulation model simulates the network-wide consequences of various disruptive events. The component-level features are extracted from the simulated data. Next, clustering algorithms are used to derive the cluster-level features by grouping component-level features based on their topological and functional characteristics. Finally, ML algorithms are used to develop models that predict the network-wide impacts of disruptive events using the cluster-level features. The applicability of the method is demonstrated using an interdependent power-water-transport testbed. The proposed method can be used to develop decision-support tools for post-disaster recovery of infrastructure networks.

ARTICLE HISTORY
Received 6 November 2023
Revised 21 April 2024
Accepted 24 April 2024

KEYWORDS
High-dimensionality; infrastructure resilience; interdependencies; machine learning; network analysis; network clustering; urban simulation

1. Introduction

The intensifying natural disasters and emergence of new threats, such as cyber-attacks and pandemics, have led to a paradigm shift in infrastructure system management approaches. The physical and functional risks posed by such extreme events to critical infrastructure systems are compounded by climate change, the drastic modifications to the built environment, and the growing interdependence among urban systems. Examples of such infrastructure systems include transportation networks, energy networks, water networks, and financial networks, among others. Therefore, the infrastructure systems are no longer designed for operational efficiency alone; equal emphasis is given to their ability to withstand disasters and minimise the resultant societal and economic impacts from unanticipated service disruptions (Gay & Sinha, 2013). In response, disaster management and resilience enhancement capabilities have become crucial aspects considered in recent efforts towards the development infrastructure analysis models. However, existing domain-specific infrastructure analysis models are computationally intensive and have limited capability in supporting post-disaster decision-making for fast restoration of damaged critical infrastructure systems. While research efforts in this direction have made considerable advancement in surrogate modelling and machine learning-based approaches, issues such as *high-dimensionality of datasets* remain significant obstacles to scaling up these models for the analysis of large-scale interdependent infrastructure networks.

The term 'infrastructure resilience' is widely discussed in the literature and is accompanied by various definitions, dimensions, characteristics, and principles that are frequently employed to describe this concept (Labaka, Hernantes, & Sarriegi, 2016). However, synthesising those definitions, broadly, infrastructure resilience can be defined as the capacity of an infrastructure system to withstand a change or a disruptive event and minimise performance deviations thereafter (Nan & Sansavini, 2017). According to Bruneau et al. (2003), resilience characteristics include robustness, redundancy, resourcefulness, and rapidity. Francis and Bekera (2014) categorised resilience characteristics of infrastructure systems into absorptive, restorative, and adaptive capacities. Recent studies also list preventive, anticipative, and transformative capacities as additional characteristics of infrastructure resilience (Manyena, Machingura, & O'Keefe, 2019).

CONTACT Srijith Balakrishnan s.balakrishnan@tudelft.nl

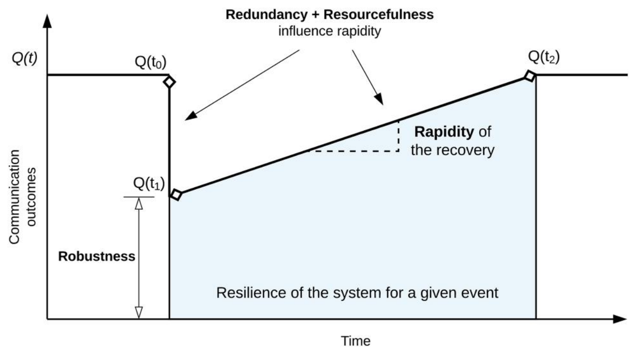


Figure 1. Resilience triangle and its relationship to the four dimensions of resilience (Balakrishnan, 2020).

Regarding the quantification of resilience, the most common approach is the resilience triangle introduced by Bruneau et al. (2003) (Figure 1). Bruneau et al. (2003) defined resilience as the area above the resilience curve, denoting the cumulative loss to system performance due to an event. Based on Figure 1, the resilience is given by $R = \int_{t_0}^{t_2} (Q(t_0) - Q(t))dt$, where t is the time and $Q(\cdot)$ is the system performance as a function of time t. Further modifications to the resilience triangle approach have been introduced by several others, including (Bocchini & Frangopol, 2012; Cimellaro, Tinebra, Renschler, & Fragiadakis, 2016; Domaneschi, Cucuzza, Martinelli, Noori, & Marano, 2024).

Several approaches exist to analyse the resilience capabilities of interdependent infrastructure systems and evaluate the resilience project alternatives. These approaches can be broadly classified into two, namely empirical- and computational approaches (Mitsova, 2021). Empirical approaches rely on datasets, records, and reports based on historical events to identify patterns and severity in physical and functional disruptions to infrastructure systems (Luiijf, Nieuwenhuijs, Klaver, van Eeten, & Cruz, 2009; McDaniels, Chang, Peterson, Mikawoz, & Reed, 2007). Data collected during historical breakdown events are used to characterise interrelationships among different infrastructure systems. On the other hand, computational approaches attempt to replicate the physical-, cyber-, geographic-, and logical dependencies among infrastructure components (hereon, components) using mathematical and logical functions.

Several models, such as, network theory-based models (Holden, Val, Burkhard, & Nodwell, 2013; Praks, Kopustinskas, & Masera, 2017), system dynamics models (Powell, DeLand, & Samsa, 2008), agent-based models (Cimellaro, Mahin, & Domaneschi, 2019; Iuliis, Battegazzorre, Domaneschi, Cimellaro, & Bottino, 2023; Thompson et al., 2019) and input-output models (Haimes et al., 2005), have been extensively used for modelling interdependent infrastructure systems and analysing their resilience. More advanced computational models adopt multi-simulation and co-simulation of multiple domain-specific infrastructure simulators instead of a homogeneous method to replicate the collective behaviour of interdependent infrastructure systems and communities (Battegazzorre, Bottino, Domaneschi, & Cimellaro, 2021; Marasco et al., 2021; Wang, Magoua, & Li, 2022).

Recent computational approaches have been increasingly adopting simulation-based machine learning (ML) models to predict the network-wide impacts and use these predictions for the analysis of interdependent infrastructure resilience (Rahimi-Golkhandan, Aslani, & Mohebbi, 2021). Most of the ML models for infrastructure resilience analysis focus on the optimal allocation of resources for improving the absorptive and recovery capabilities of infrastructure systems. Alemzadeh, Talebiyan, Talebi, Duenas-Osorio, and Mesbahi (2020) developed Artificial Neural Network (ANN) models to approximate post-disaster repair sequences for optimal recovery of interdependent infrastructure systems by training simulated data. Sun and Zhang (2020) applied a Deep Q-learning (DQN) algorithm on an interdependent power-water-transport model to predict the optimal repair crew allocation to flood-affected bridges that would minimise the cumulative impact on the network. Dehghani, Jeddi, and Shafieezadeh (2021) applied Deep Reinforcement Learning (DRL) to identify the optimal long-term preventive maintenance strategy that maximises the resilience of power systems.

Although ML models are powerful tools to accurately predict infrastructure system resilience (i.e. network-wide impacts of different disruptive events), they require large training datasets to ensure adequate prediction accuracy. Most of the aforementioned ML models learn the vulnerability and resilience attributes at a component level, leading to a large number of features in the model, also known as the '*curse of dimensionality*' (Turati, Pedroni, & Zio, 2016). When the dimensionality of a problem increases, generalising the trends becomes harder because the training dataset of fixed size can cover only a small fraction of the possible input combinations (Domingos, 2012). While using a larger simulation dataset could be a potential solution to the problem of high-dimensionality, several advanced infrastructure simulation models are computationally intensive and may require considerable time for simulating even a small number of disaster scenarios (Liu, Ferrario, & Zio, 2019; Zou & Chen, 2021). The consequences of high dimensionality could be even more severe when the system response of large-scale interdependent infrastructure networks is to be learned by the ML algorithms.

To mitigate the negative effects of high dimensionality, this study exploits clustering methods to identify similar components in terms of their topological and functional properties and later incorporates that information to enhance the accuracy of resilience prediction models. Several studies in the literature have demonstrated that topological and functional attributes of components influence the infrastructure network vulnerability and resilience characteristics (Balakrishnan & Zhang, 2020; Cadini, Zio, & Petrescu, 2009; Nicholson, Barker, & Ramirez-Marquez, 2016).

The overall objective of this paper is to propose a novel network clustering-based approach for achieving dimensionality reduction in simulation-based machine learning models for infrastructure network analysis. The specific objectives are as follows:

1. Demonstrate how the relationship between network topology and infrastructure resilience, as established in the literature, can be leveraged to incorporate network structure characteristics in ML prediction models.
2. Present a methodological framework to apply network clustering to reduce dimensionality in infrastructure resilience prediction models.
3. Propose unsupervised and supervised methods to find the optimal number of clusters in each infrastructure system in an interdependent network.

The rest of the paper is organised as follows: Section 2 presents the methodological framework adopted in this paper; Section 3 demonstrates the application of the methodology on a synthetic interdependent power-water-transport network; and Section 4 summarises the findings and discusses the scope for further research.

2. Methodology

The methodological framework adopted in the study is outlined in Figure 2. The methodology is implemented in three steps, namely, (a) generation of simulation dataset, (b) infrastructure component clustering, and (c) dimensionality reduction and development of prediction models. In the first step, an interdependent infrastructure model is employed to simulate the infrastructure disruption data required for developing the prediction model. The interdependent simulation model uses power, water, and transportation network simulators, as well as disaster event scenarios, to simulate the functional impacts at the component and network levels under normal and disrupted states. The second step is the application of appropriate clustering algorithms to categorise infrastructure components based on their topological and functional characteristics. In this step, appropriate graph algorithms are employed to extract the component-level topological characteristics of the three infrastructure networks. These topological characteristics are later combined with the corresponding functional characteristics under normal operating conditions. Using the topological and functional characteristics, the infrastructure components are clustered using the K-Means clustering algorithm. Finally, ML models are developed for predicting the network-wide impacts using disaster and recovery-related features. In order to resolve high-dimensionality, cluster-level features corresponding to initial component failures are used instead of component level features. To summarise, a methodological framework that combines infrastructure component clustering with existing machine-learning algorithms is proposed in this study to reduce dimensionality and facilitate infrastructure network resilience prediction simultaneously. In the rest of the section, the various steps in the methodology are discussed in detail.

2.1. Generation of simulation dataset

This step derives the relevant target and predictor features using an interdependent infrastructure simulation model. The step further constitutes two subtasks: interdependent infrastructure simulation and feature extraction.

2.1.1. Interdependent infrastructure simulation

In this study, *InfraRisk*, an open-source Python-based integrated simulation platform, is employed for simulating infrastructure disruptions and subsequent recovery in interdependent power-water-transport networks (Balakrishnan & Cassottana, 2022). *InfraRisk* integrates existing domain-specific infrastructure simulators (*pandapower* for power systems (Thurner et al., 2018), *wntr* for water distribution systems (Klise et al., 2020), and a static traffic assignment

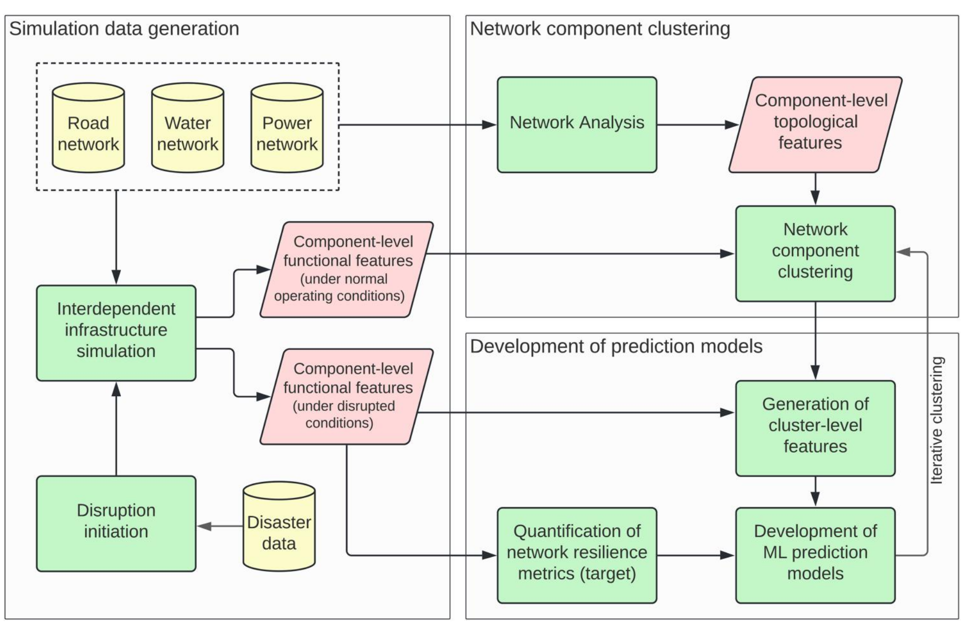


Figure 2. Framework for developing ML-based infrastructure resilience models. (a) Original network with no clustering. (b) Clustered network.

package for transport network (Boyles, Lownes, & Unnikrishnan, 2020)) *via* an object-oriented interface to perform the interdependent simulations. All the three infrastructure simulators assign flows (power, water, or traffic) in the respective networks by minimising the total loss subject to network-specific constraints and dependencies with other infrastructure systems.

Consider an interdependent infrastructure network $\mathbb{K}$ with constituent infrastructure systems $k : k \in K$. An infrastructure system k can be represented as a graph $G(V_k, E_k)$, where V_k is the set of nodes and E_k the set of links connecting the nodes. In addition, K also includes the dependencies between infrastructure systems, which are represented as links. The set of consumers who are dependent on K is denoted by N, and the resource supply from k to each consumer $n \in N$ at simulation time $t \in T$ under normal operating conditions is represented by $s_n^{k,0}(t) \in S^0$.

To implement the simulation, it is required to define the disaster scenario and the recovery sequence for infrastructure system k. Suppose $h : h \in H$ is the disaster scenario which results in the failure of infrastructure nodes $V_k^h \subseteq V_k$ and links $E_k^h \subseteq E_k$. On the other hand, the recovery sequence p_k^h indicates the order in which the damaged components are to be restored in an infrastructure system k. The repair sequence is determined by prioritising damaged components based on three measures: betweenness centrality, zoning, and maximum daily flow that are handled during normal operating conditions. Once the repair sequence is finalised, the *InfraRisk* model has a recovery module that takes into consideration various factors, such as the availability of repair crews, component repair times, and accessibility by road to the failed components, to determine the start time and end time of each repair action. Normally, each system has its own set of repair crews and restoration strategies to expedite recovery actions.

The consequences of initial infrastructure disruptions due to the hazard, the network-wide functional disruptions due to infrastructure interdependencies, and the subsequent recovery efforts are reflected in the resource supplied to the consumers during the disrupted conditions $(s_n^{k,h}(t) \in S^h)$ and is captured by *InfraRisk*. To summarise, the interdependent infrastructure model can be represented as follows:

$$S^h = \mathbf{\Gamma}\left(K, h, P^h\right), \tag{1}$$

where $\mathbf{\Gamma}(\cdot)$ is the simulation model and $P^h = \{p_k^h : k \in K\}$.

2.1.2. Feature extraction

The simulation model provides three types of data required to develop the prediction models.

a. Timeline of consumer-level and system-level functional performance under normal and disrupted network conditions.
b. Topological characteristics of the infrastructure systems.
c. System disruption and recovery characteristics, such as the list of initially failed components and the repair sequencing strategy used.

Since the objective of the ML model is to predict the network-level resilience, the consumer-level resource supply values generated by the simulation model are converted to network-level resilience metrics. In this study, the Prioritised Consumer Serviceability (PCS) is used as the measure of performance (MOP) for tracking the performance of the networks (Balakrishnan & Cassottana, 2022), defined as:

$$\mathrm{PCS}^{k,h}(t) = \frac{\sum_{\forall n: s_n^{k,0}(t)>0} s_n^{k,h}(t)}{\sum_{\forall n: s_n^{k,0}(t)>0} s_n^{k,0}(t)}, \text{where } 0 \leq s_n^{k,h}(t) \leq s_n^{k,0}(t). \tag{2}$$

The resource supply to consumers under normal operating conditions $s_n^k(t) \in S$ is computed by performing the interdependent infrastructure simulation without failing any infrastructure components $(S = \mathbf{\Gamma}(K))$. The resilience of infrastructure systems is quantified using the concept of equivalent outage hours (EOH). The EOH corresponding to each infrastructure network for a given disaster scenario γ_k is:

$$\gamma_k = \frac{1}{3600} \int_{t_0}^{t_{\max}} [1 - PCS_k(t)]dt, \tag{3}$$

where t_0 is the time of occurrence of the disaster event in the simulation and t_{max} is the maximum simulation time (both in seconds). Mathematically, $[1 - PCS_k(t)]$ is the mean unmet demand (slack demand) in k at time t and γ_k is the area of the resilience triangle (Bruneau et al., 2003) formed by the PCS curve. The unit of γ_k is system performance-hours.

The target variable for the ML model is the resilience metric of the interdependent infrastructure network, computed as the weighted equivalent outage hours ($\bar{\gamma}$) of individual infrastructure systems as follows:

$$\bar{\gamma} = \sum_{k \in K} w_k \gamma_k, \tag{4}$$

where w_k is the weight assigned to system k. Calibrating the weights w_k in the model can be done in several ways. For residential buildings, it depends on the consumption characteristics of households. To accurately compute the weights, surveys need to be conducted to understand the impact of water and power outages on households with different social, demographic, and economic characteristics. In the case of non-residential buildings, the impact of outages depends on the type of economic activity. There are two dominant methods to determine the effect of water and power outages in non-residential buildings. The first method relies on technical coefficients derived from input-output (IO) tables to measure the initial impact on economic sectors caused by infrastructure disruptions (Haimes et al., 2005; Santos, 2006). Alternatively, the second approach involves conducting surveys to quantify the functional dependencies of businesses on infrastructure services (Chang, Seligson, & Eguchi, 1996; Kajitani & Tatano, 2009). In addition to the model inputs (V_k^h, E_k^h, and P_k^h) and outputs (S, S^h, and $\bar{\gamma}$), the topological features of components, such as, centrality values, are also extracted from the infrastructure network to aid network clustering.

2.2. Infrastructure component clustering

Clustering is an unsupervised learning method that attempts to identify the most natural way of partitioning a dataset based on similarity and dissimilarity among observations (Xu & Tian, 2015). Clustering could reveal the underlying structure of the dataset, which could be used to build supervised learning models with simple features and a better prediction accuracy (Trivedi, Pardos, & Heffernan, 2015). In this study, clustering is proposed to identify components that are similar in their vulnerability and resilience characteristics. For this purpose, the applicability of topological and functional features of components that are identified as indicators of infrastructure vulnerability, criticality, and resilience in the literature is also investigated.

Figure 3 illustrates the clustering concept introduced in the study for partitioning infrastructure systems based on their underlying graph structures. Any infrastructure network can be treated as a set of graph elements, i.e. nodes and links, where nodes represent producers, consumers, or intermediate transfer points, whereas the links denote the connections and interdependencies among the infrastructure nodes (Svendsen & Wolthusen, 2008). Therefore, each component in the system can be assigned a group (or cluster) based on its topological and functional properties (color-coded in Figure 3(b)). However, it shall be noted that the components in one cluster need not be adjacent to each other in the network.

A number of algorithms are available for performing clustering of datasets, such as, K-Means, K-Medoids, agglomerative propagation, and DBSCAN, depending upon the types of the datasets. For systematic reviews of clustering algorithms and their applications, see Rodriguez et al. (2019).

Consider the infrastructure system $k \in K$ which can be represented as a graph $G(V_k, E_k)$. Without clustering, a characteristic associated with the components (for example, post-disaster functional status) can be incorporated in two different ways as follows:

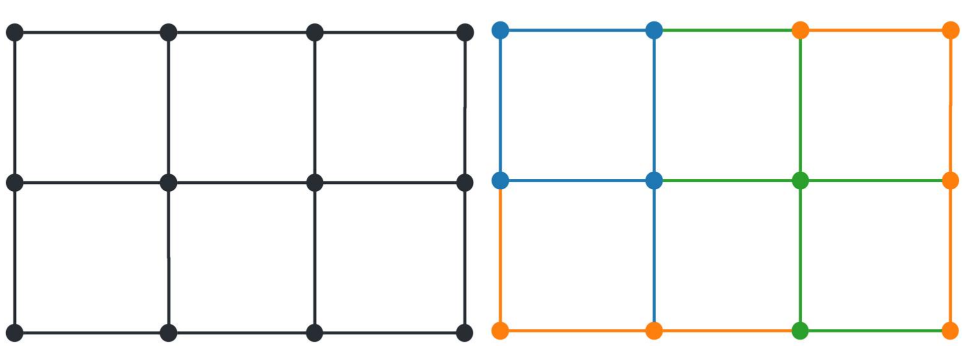
1. One feature for each component so that there will be $|V_k| + |E_k|$ additional features corresponding to each infrastructure system k in the ML prediction model.
2. A single feature which aggregates the component-level values using mathematical operations, such as summation, count, average, maximum, and minimum.

The above two approaches have their own advantages and limitations. While the first approach may effectively learn from the spatial and structural aspects of the characteristic to be considered and ensure a high level of prediction accuracy, it leads to the issue of high-dimensionality. For a fully connected graph with m nodes, $m + m(m-1)/2$ features for every component-level characteristic need to be constructed in this approach if both node- and link-level information are to be used in the ML model. In the second approach, the number of features could be considerably lower than in the first approach; however, aggregation would lead to the loss of useful information related to components, leading to a low prediction accuracy. The clustering approach aims at finding a middle ground between the above two extreme approaches. By employing appropriate clustering algorithms, $\ell_k^V : 1 \leq \ell_k^V \leq |V_k|$ node clusters and $\ell_k^E : 1 \leq \ell_k^E \leq |E_k|$ link clusters with similar functional and topological properties in the infrastructure system $k \in K$ are identified.

Several studies have demonstrated that centrality measures (Balakrishnan & Zhang, 2020; Cadini et al., 2009; Dunn, Fu, Wilkinson, & Dawson, 2013), such as degree centrality, betweenness centrality, and eigenvector centrality, could capture the vulnerability and resilience of components. In addition, many studies have shown that specific functional properties, such as flow rates under normal operating conditions, could serve as indicators of the vulnerability and importance of the components in a system (Nicholson et al., 2016). To incorporate node-specific and link-specific characteristics, clustering of nodes and links are done separately. Table 1 enlists the potential topological and functional features that are considered in the current study for component clustering.

2.3. Dimensionality reduction and development of prediction models

In this study, dimensionality reduction is achieved using an iterative clustering algorithm introduced in this study. The iterative clustering algorithm combines clustering methods with regression algorithms to produce concise infrastructure resilience prediction models. This step consists of two



(a) Original network with no clustering (b) Clustered network

Figure 3. Illustrative example of clustering infrastructure components.

Table 1. Topological and functional features used for infrastructure component clustering.

Feature	Node element (i)	Link element (i, j)
Degree centrality	$C_d(i) = \sum_{v\in V} a_{i,v}$	$C_d(i) = \sum_{v\in V} a_{i,v}$; $C_d(j) = \sum_{v\in V} a_{j,v}$
Betweenness centrality	$C_b(i) = \frac{\sigma_{st}(i)}{\sigma_{st}}$	$C_b(ij) = \frac{\sigma_{st}(i,j)}{\sigma_{st}}$
Eigenvector centrality	$C_e(i) = \frac{1}{\lambda}\sum_{v\in V} a_{i,v}C_e(v)$	$C_e(i) = \frac{1}{\lambda}\sum_{v\in V} a_{i,v}C_e(v)$; $C_e(j) = \frac{1}{\lambda}\sum_{v\in V} a_{j,v}C_e(v)$
Closeness centrality	$C_c(i) = \frac{\lvert V\rvert - 1}{\sum_{v\in V} d(i,v)}$	$C_c(i) = \frac{\lvert V\rvert - 1}{\sum_{v\in V} d(i,v)}$; $C_c(j) = \frac{\lvert V\rvert - 1}{\sum_{v\in V} d(j,v)}$
Flow-rate	$Q_i = \sum_{j\in M(i)} \text{abs}(q_{ji})$	$Q_{ij} = \text{abs}(q_{ij})$
Weighted flow-rate	$\bar{Q}_i = \sum_{j\in M(i)} C_b(ij)\text{abs}(q_{ji})$	$\bar{Q}_{ij} = C_b(ij)\text{abs}(q_{ij})$

Notes: $i, j, v \in V$ are nodes in network G. $(i,j) \in E$ are links. $a_{i,n} = 1$ if $(i,n) \in E$ else 0.
σ_{st} is the number of all shortest paths in G. $\sigma_{st}(i)$ and $\sigma_{st}(i,j)$ are number of shortest paths passing through i and (i, j), respectively.
λ is a constant equal to the largest positive element in the eigenvector.
d(i, v) is the shortest distance between i and v nodes if a path exists between them.
q_{ji} is the maximum daily flow-rate from j to i during normal operation. $M(i)$ is the set of neighbour nodes of i.

subtasks: construction of cluster-level features and development of ML models.

2.3.1. Construction of cluster-level features from component-level features

Once the infrastructure components are categorised into different clusters, the next step is to derive the cluster-level features from the simulation dataset. In the dataset generated using *InfraRisk*, the component-level information to be incorporated in the ML model is their initial functional states (disrupted or operational) after the occurrence of a disaster event. Let a node cluster and a link cluster for infrastructure system k generated by the clustering algorithm are denoted by $\ell_k^v : \ell_k^v \subseteq V_k$ and $\ell_k^e : \ell_k^e \subseteq E_k$, respectively and the set of all clusters by L. Then the cluster-level features corresponding to the initial functional states of components are to be derived. For this, it is required to assign the cluster to each component as in the following equation:

$$\delta(i, \ell_k^v) = \begin{cases} 1 & \text{if } i \in \ell_k^v, \\ 0 & \text{otherwise} \end{cases}; \quad \delta(i, \ell_k^e) = \begin{cases} 1 & \text{if } i \in \ell_k^e, \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\delta(i, \ell_k^v)$ and $\delta(i, \ell_k^e)$ are the indicator variables denoting whether a component i belongs to a node cluster ℓ_k^v or a link cluster ℓ_k^e, respectively.

Now, the cluster-level features can be derived. Let F_ℓ denote a cluster-level feature (node- or link-level) corresponding to the cluster $\ell \in L$. The cluster-level feature F_ℓ which represents the initial impacts to the cluster due to hazard scenarios $H \ni h$ is computed as the total number of failed components that belong to that cluster under different hazard scenarios as follows:

$$F_\ell = \left\{ \sum_{i\in V_k^h \cup E_k^h} \delta(i, \ell) \right\}_{\forall h \in H} \tag{6}$$

2.3.2. Development of ML prediction models

The last step of the proposed methodology is to develop the prediction models by employing state-of-the-art ML algorithms. Since the target variable (weighted EOH) in this study is continuous, only regression algorithms are considered. Commonly used regression algorithms include multiple linear regression (MLR), support vector regression (SVR), decision tree (DT), Random Forest (RF), and neural network regression (NN). For detailed discussions and applications of the major ML regression algorithms, the readers to are referred to Berk (2008) and James, Witten, Hastie, and Tibshirani (2013).

Finally, the ML models are built using cluster-level features and recovery strategy as the predictors as follows:

$$\bar{\gamma} = \Phi(\mathbb{F}, \mathbb{P}), \tag{7}$$

where $\mathbb{F} = \{F_\ell : \ell \in L\}$, $\mathbb{P} = \{P_h : h \in H\}$, and $\Phi(\cdot)$ is the ML algorithm. To obtain robust and accurate models, cross-validation and hyper-parameter tuning are performed.

Two performance metrics are used to evaluate the goodness-of-fit of the ML models. The first metric is the coefficient of determination (R^2), which is defined as the proportion of the variation in the target variable captured by the prediction model.

The second performance metric is the root mean square error (*RMSE*), which is the standard deviation of the prediction errors in the model. The goal is to have a high R^2 and low *RMSE* values for the final prediction model. The unit of *RMSE* in the this study is system performance-hours.

A major aspect that is not yet resolved in the method is determining the optimal number of clusters in each infrastructure system for building the final ML model. Since an increase in the number of clusters would enhance the capability of the model to capture the spatial and network structure characteristics, an improvement in the quality of model prediction is expected. The optimal cluster count in each infrastructure system is determined using the *elbow* method (Yuan & Yang, 2019) and the proposed iterative clustering method.

The elbow method is an unsupervised method in which the sum of squared distances between observations and the centroids of the clusters they belong to is used as the performance measure for evaluating the consistency of clusters. The ‘elbow’ of the curve connecting the sum of squared distances and total cluster count is determined, and the corresponding number of clusters in each infrastructure system is identified.

The second method proposed in this study to identify the optimal number of clusters in each infrastructure network is an iterative clustering algorithm (Figure 4). It is a supervised method in which the performance of the ML model on test dataset is used as the performance measure. In this method, setting the ML model corresponding to the elbow method as the base model, an equal number of clusters are increased or decreased in all infrastructure systems to develop additional ML models. Once an adequate number of cluster combinations and corresponding ML models are constructed, the improvement to the model goodness-of-fit (R^2) due to an increase in the number of clusters is quantified. The most efficient model, i.e. the ML model with the highest R^2, is then adopted.

Consider $\{l_k : k \in K\}$ are the cluster counts corresponding to the infrastructure systems in the interdependent network. If the cluster counts determined by the elbow method are $\{l_k^{elb} : k \in K\}$, the maximum number of clusters that can be removed simultaneously from each network is $b^{(-)} = \min\{l_k^{elb} - 1 : k \in K\}$, and the maximum number of clusters that can be added is $b^{(+)} = \min\{|V_k| + |E_k| - l_k^{elb} : k \in K\}$. The iterative clustering algorithm is initialised by building the initial model with cluster counts $\{l_k^{elb} - b^{(-)} : k \in K\}$ and evaluating the model train and test performance metrics. In the subsequent iterations, the above procedure is repeated after updating the cluster counts $\{l_k^{elb} + b : k \in K\}$, where $b \in \{-b^{(-)}, \ldots, b^{(+)}\}$. Finally, the optimal cluster counts in infrastructure networks are obtained by finding the 'knee' of the curve between the test dataset R^2 and the total cluster count. *Kneedle* algorithm proposed by Satopää, Albrecht, Irwin, and Raghavan (2011) is employed for this purpose.

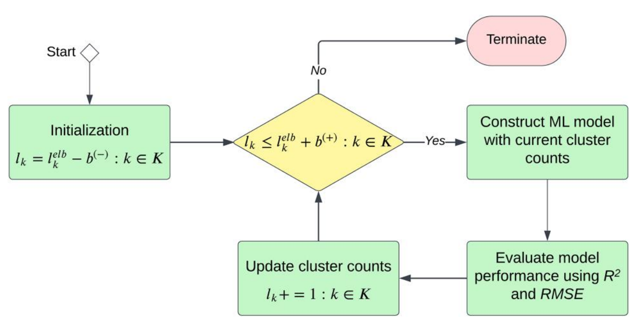


Figure 4. Iterative clustering method.

3. Case study

The proposed methodology is implemented to develop a resilience prediction model for the Micropolis interdependent infrastructure network (Figure 5). Micropolis is a virtual city designed for 5000 inhabitants with water, power, and road networks (Brumbelow, Torres, Guikema, Bristow, & Kanta, 2007).

The hazard module in *InfraRisk* is used to generate synthetic flood events and fail components randomly based on the disaster intensity. To generate the floods using realistic flood models, a flood hazard profile is required. However, since there is no real historical disaster data available for the network, the authors adopted simpler assumptions to derive the flood risks to components. It was assumed that the floodplains (regions exposed to floods) are located within a certain distance from the centreline of the main water stream. Multiple flood intensities were considered, each with corresponding probabilities of occurrence. The conditional probability of exposure was determined based on a distance from centreline. The conditional component failure probabilities for the generated floods in the study were determined based

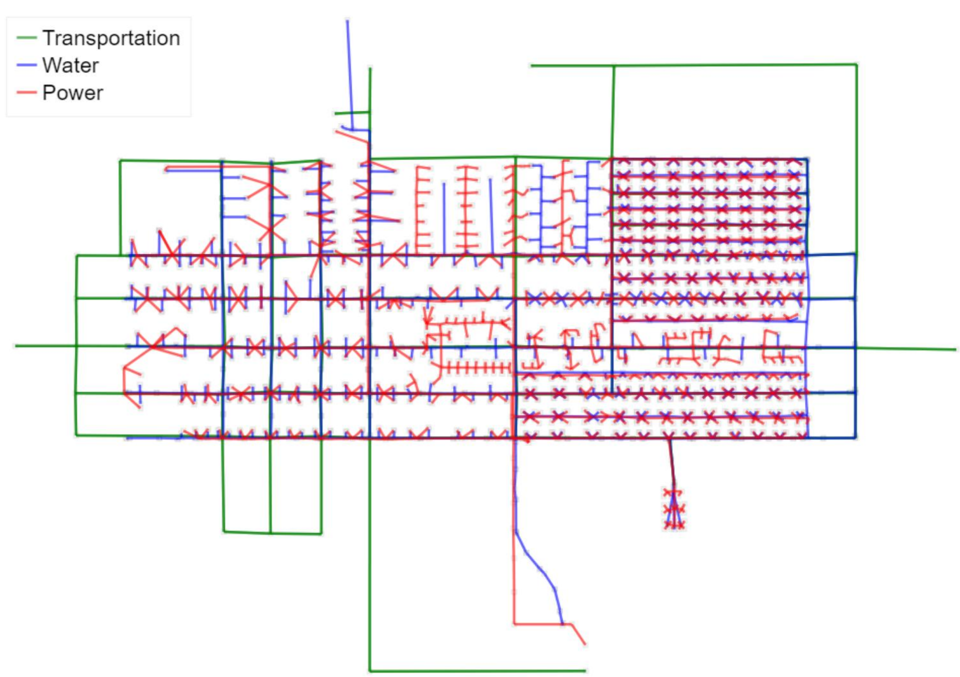


Figure 5. Micropolis interdependent infrastructure network.

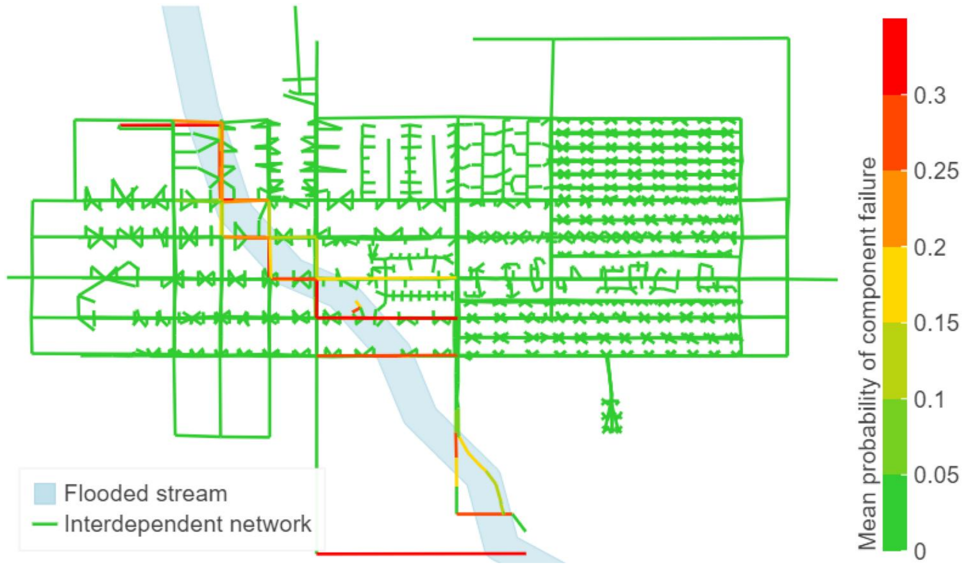


Figure 6. Micropolis network exposed to floods. (a) Physically disrupted components and crew locations. (b) System recovery curves.

on the intensity of the events. The failure modelling approach has limited impact on the performance of the machine learning model developed in subsequent sections, as the purpose of this stage is to create a dataset with diverse combinations of component failures.

A total of 325 flood scenarios are generated, assuming infrastructure components closer to the Micropolis stream are more likely to fail from a flood (Figure 6). Each disaster scenario results in the failure of a specific set of components in the interdependent infrastructure network. For this case study, only water mains, power lines, and road links are considered for failure as they are the most critical to the functioning of the respective infrastructure systems. It is found that 52 water mains, 22 power lines, and 17 road links along the water stream are either located or traversing through the regions exposed to the simulated floods. The maximum number of failures in each flood scenario is limited to 35 components to reduce the computational effort required for the case study. The disruptions to water links are modelled as leaks/pipe breaks, whereas that of power lines and road links are modelled by isolating them from the network.

It is assumed that each failed pipeline is remotely isolated by a predefined set of shutoff valves 10 min after the disaster occurred (considering sensing and actuation times). By isolating the leaking pipelines, the loss of water is minimised; however, isolating some segments of the water system would cutoff consumers located within the isolated regions. Once a water pipe is repaired, the corresponding isolating valves are opened, conditional upon whether that would interfere with the remaining repair actions. Similarly, when a power line is fully repaired, the corresponding circuit breakers are closed to allow electric power to flow through the line. In the case of damaged road links, each link is added back to the network after repair, and then the traffic assignment model recomputes the traffic flows.

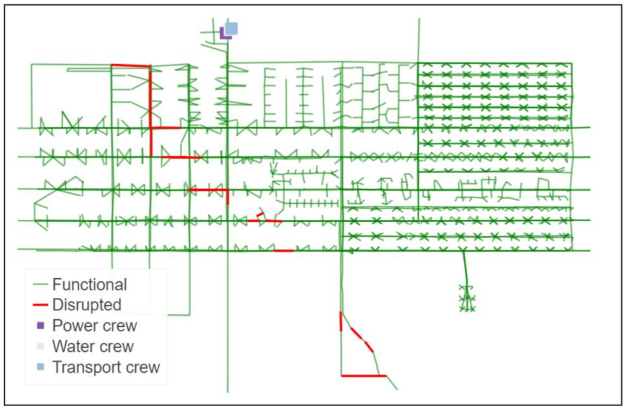
In this case study, each infrastructure system is assigned a repair crew for performing the post-disaster recovery. For implementing network recovery by component repair, three repair strategies are considered as follows:

- Betweenness centrality-based: Those components with a higher value of betweenness centrality are repaired first.
- Maximum flow-based: Those components that handle larger resource flow rates are repaired first.
- Zone-based: Components are repaired based on the zone in which they are located. The zones are prioritised in the order of central business district, industrial, and residential areas.

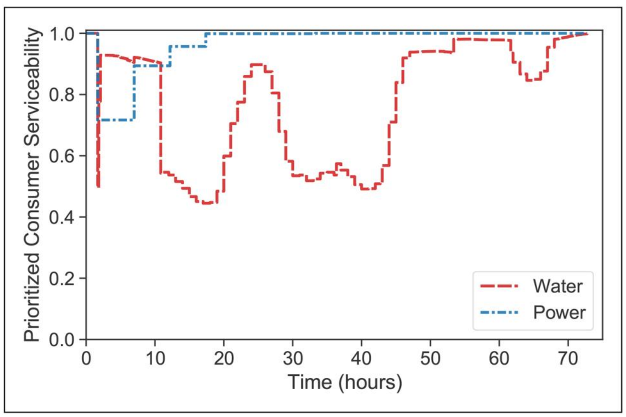
The recovery model in *InfraRisk* also takes the accessibility to disrupted components into consideration and dynamically modifies repair sequences during the simulation.

Figure 7 presents the simulation results corresponding to one of the 325 simulated flood events. The flood event resulted in the failure of 14 water mains, six power lines, and four road links (Figure 7(a)). Figure 7(b) shows the performance of the Micropolis water and power network during the flood event when the capacity-based strategy is chosen for the network recovery. The infrastructure system performance is measured using the MOP in Equation (2). Both performance curves follow the typical resilience triangle used to characterise system resilience. The power system is restored to the pre-disaster state in approximately 33 h, whereas the water crew takes approximately 68 h to complete all the repair actions. The road network is fully restored in approximately 51 h. Using Equation (3), the equivalent outage duration (in hours) corresponding to the disaster scenario in the water network is estimated to be 17.55 system performance-hours, and that in the power network is 2.92 system performance-hours.

The consumer-level outages in power- and water utility services during the flood event are shown in Figure 8. Even though the consumers in the adjacent areas along the flooded stream are affected by water outages the most,

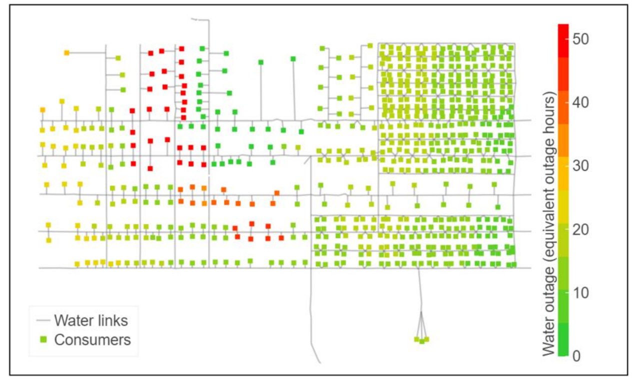


(a) Physically disrupted components and crew locations

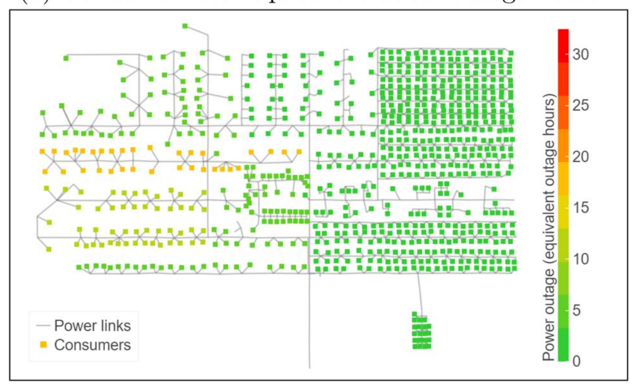


(b) System recovery curves

Figure 7. Simulation results from a simulated flood event.



(a) Consumer-level equivalent water outage hours



(b) Consumer-level equivalent power outage hours

Figure 8. Consumer-level impacts due to the simulated flood event.

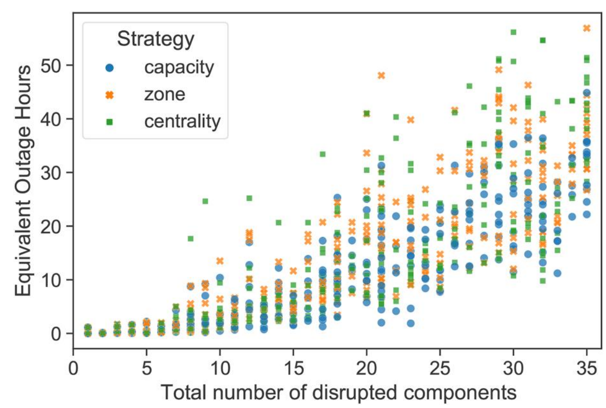


Figure 9. Weighted equivalent outage hours versus failure count.

consumers in other parts of the network (especially those in the western and central Micropolis) are also affected by water outages (Figure 8(a)). The leakage through failed pipes resulted in a reduced water head in the tank. The closure of shutoff valves isolated many consumers in other regions even though they were not directly affected by the flood event. On the other hand, the power outage is less severe compared to the water outage and is limited to the western side of the flooded stream (Figure 8(b)). The drop in resilience values is mainly attributed to the consumers in downstream of failed power lines or opened circuit breakers who are disconnected from the rest of the power network.

Next, the network resilience metrics (weighted equivalent outage hours) corresponding to all the disaster scenarios are calculated by assigning equal weights of 0.5 in Equation (4). Only water and power systems are considered for evaluating the network resilience. Figure 9 presents the distribution of the resilience metrics and their relationship with the recovery strategy adopted and the number of physically disrupted components. The results show a positive correlation between the weighted equivalent outage hours and the number of initially failed components.

For developing cluster-level features, partitioning of the components belonging the three infrastructure systems needs to be performed. For this study, K-Means clustering algorithm (Hartigan & Wong, 1979) is adopted. K-Means clustering is an unsupervised learning algorithm used for partitioning datasets into clusters based on their similarity. The objective of K-Means clustering is to minimise the within-cluster object distances. In the context of this study, the purpose of clustering is to identify components that share similar topological and functional characteristics. K-Means clustering offers a straightforward and interpretable approach for effectively partitioning datasets with similar properties.

Next, ML models to predict the resilience (in terms of weighted equivalent outage hours) are developed as in Equation (7). Random Forest algorithm is used for this case study because of its simplicity and robustness. In addition, Random Forest algorithm has been effective in several critical infrastructure network applications, such as fault/failure

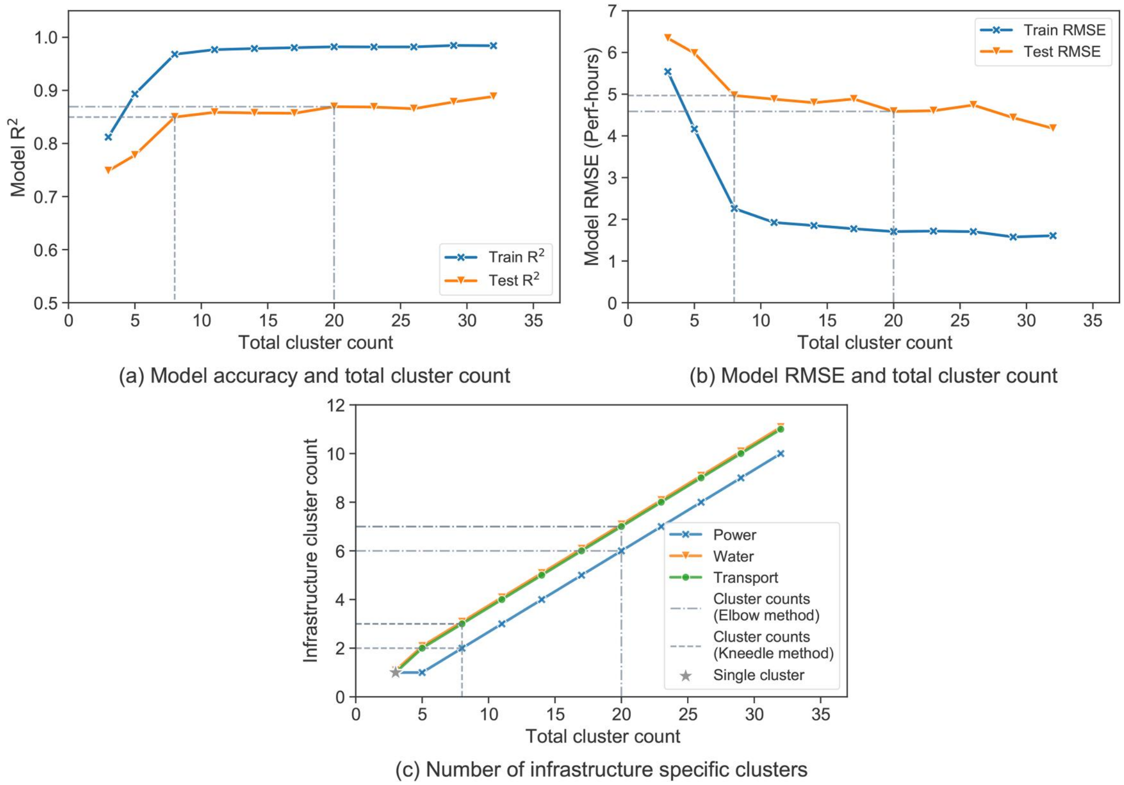


Figure 10. Relationship between model performance and number of infrastructure clusters.

Table 2. Comparison of performance metrics of prediction models.

Clustering method	Total clusters	Train R^2	Test R^2	Train *RMSE*	Test *RMSE*	Relative change[a] Test R^2	Test *RMSE*
Single cluster	3	0.8121	0.7485	5.54	6.34	–	–
Elbow method	20	0.9821	0.8693	1.71	4.59	+16.13%	−27.67%
Kneedle method	8	0.9680	0.8497	2.26	4.96	+13.52%	−21.65%

[a]All percentages are relative to the test dataset metrics obtained in the single cluster method.

detection (Haggag, Yorsi, El-Dakhakhni, & Hassini, 2021), anomaly detection (Farnaaz & Jabbar, 2016), and resilience analysis and prediction (Cassottana et al., 2022; Goforth, Yosri, El-Dakhakhni, & Wiebe, 2022). The ML models are built using 75% of the data (training dataset), and the rest 25% of the data (test dataset) is used for validation. The hyperparameters corresponding to the maximum depth and the total number of trees in the Random Forest algorithm are tuned for each model using three-fold cross-validation.

To identify the optimal clusters, the elbow method and the iterative clustering method combined with the *kneedle* algorithm are used. Figure 10 shows the results from the machine learning models developed using these methods. In the case of the elbow method, the optimal number of clusters in water, power, and transport systems are found to be six, seven, and seven, respectively. The train- and test R^2 corresponding to the optimal cluster counts are 0.98 and 0.87, respectively. The corresponding *RMSE* values are 1.71 and 4.59 system performance-hours.

The machine learning models developed based on the iterative clustering algorithm reveal that the increase in the number of clusters in infrastructure systems initially leads to a noteworthy improvement in the model prediction. However, subsequent increases in the number of clusters only lead to marginal improvements in the same metrics. When the kneedle algorithm is used, the optimal number of clusters based on the iterative clustering method is found to be eight (two clusters in the power system and three each in the water system and the transport system). The corresponding Random Forest model has a train R^2 of 0.97 and a test R^2 of 0.85. The train *RMSE* is 2.26 system performance-hours, and the test *RMSE* is 4.97 system performance-hours.

Both elbow and the iterative clustering method resulted in improved prediction models compared to that of the model with the single cluster-level feature for each infrastructure system. In the case of the elbow method, the relative improvement observed with the elbow method in test R^2 is 16.13%, whereas that using the iterative clustering method is 13.52%. At the same time, the models identified using the elbow method and the iterative clustering method resulted in significant reductions of 27.67% and 21.65% in test *RMSE*, respectively. The results show that the iterative clustering method resulted in a model with considerably fewer cluster features than the elbow method (eight cluster features compared to 20 cluster features) without compromising too much on the prediction accuracy. The summary of the models developed in this study is presented in Table 2.

4. Conclusions

In this study, component clustering methods are introduced to generate concise infrastructure resilience prediction models with fewer features than the traditional models, thereby resolving the problem of high dimensionality. Disaster scenarios and resultant impacts on interdependent infrastructure networks are simulated using an interdependent infrastructure model. The disaster impacts on infrastructure are quantified using well-established resilience metrics. Prediction models are developed by applying machine learning algorithms. The component-level features are categorised into cluster-level features to reduce the number of features in the models (dimensionality reduction). The clusters are identified by partitioning infrastructure components with similar topological and functional characteristics as indicators of component vulnerability and importance. Finally, algorithms are proposed for determining the optimal number of component clusters based on elbow- and iterative clustering methods.

The clustering approach is a simple transformation technique that reduces the number of features (dimensionality reduction) and improves the model performance simultaneously. Since the clustering technique reduces the number of features in the model, improved prediction accuracy could be achieved with smaller simulation datasets. Therefore, the methodology can be adopted when simulation models' data generation is computationally expensive and time-consuming.

The methodology could be further improved by considering the following aspects to produce more accurate prediction models.

- Along with topological and functional characteristics, simulation data may also be used to improve the quality of network partitioning.
- Clustering algorithms based on graph neural networks (such as, Graph Convolutional Networks) that implicitly learn the infrastructure network structure could produce more relevant clusters for resilience prediction.
- Interdependencies are currently not considered for clustering of infrastructure components.
- Additional component-level features relevant to resilience (for example, repair times) could be used in the clustering process to enhance the quality of clustering.

Though this paper only focused on interdependent infrastructure systems, the presented methodology can be used to design efficient ML models for any network problem where the characteristics of vertices or edges are treated as features. The proposed methodology could find applications in network problems in various fields, such as chemistry, medicine, finance, and social science.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

Data availability statement

The *InfraRisk* package can be downloaded from GitHub. Documentation and codes for sample simulations are available in the *InfraRisk* package. This paper has been published on the pre-print server of ArXiv (Balakrishnan, Cassottana, & Verma, 2022).

References

Alemzadeh, S., Talebiyan, H., Talebi, S., Duenas-Osorio, L., & Mesbahi, M. (2020). Resource allocation for infrastructure resilience using artificial neural networks. In Miltos Alamaniotis and Shimei Pan (Eds.), *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 617–624). Washington DC: IEEE Computer Society. doi:10.1109/ICTAI50040.2020.00100

Balakrishnan, S. (2020). (*Methods for risk and resilience evaluation in interdependent infrastructure networks*) (PhD thesis). University of Texas at Austin, Austin, TX, USA.

Balakrishnan, S., & Cassottana, B. (2022). Infrarisk: An open-source simulation platform for resilience analysis in interconnected power–water–transport networks. *Sustainable Cities and Society*, *83*, 103963. doi:10.1016/j.scs.2022.103963

Balakrishnan, S., Cassottana, B., & Verma, A. (2022). Application of clustering algorithms for dimensionality reduction in infrastructure resilience prediction models. arXiv preprint arXiv:2205.03316.

Balakrishnan, S., & Zhang, Z. (2020). Criticality and susceptibility indexes for resilience-based ranking and prioritization of components in interdependent infrastructure networks. *Journal of Management in Engineering*, *4*, 04020022.

Battegazzorre, E., Bottino, A., Domaneschi, M., & Cimellaro, G. P. (2021). Idealcity: A hybrid approach to seismic evacuation modeling. *Advances in Engineering Software*, *153*, 102956. doi:10.1016/j.advengsoft.2020.102956

Berk, R. A. (2008). *Statistical learning from a regression perspective* (Vol. 14). New York: Springer.

Bocchini, P., & Frangopol, D. M. (2012). Optimal resilience- and cost-based postdisaster intervention prioritization for bridges along a highway segment. *Journal of Bridge Engineering*, *17*(1), 117–129. doi:10.1061/(ASCE)BE.1943-5592.0000201

Boyles, S. D., Lownes, N. E., & Unnikrishnan, A. (2020). *Transportation network analysis* (Vol. 1, Version 0.85).

Brumbelow, K., Torres, J., Guikema, S., Bristow, E., & Kanta, L. (2007). Virtual cities for water distribution and infrastructure system research. In Karen C. Kabbes (Ed.), *Restoring Our Natural Habitat - Proceedings of the 2007 World Environmental and Water Resources Congress* (pp. 1–7). Reston, VA: American Society of Civil Engineers. doi:10.1061/40927(243)469

Bruneau, M., Chang, S. E., Eguchi, R. T., Lee, G. C., O'Rourke, T. D., Reinhorn, A. M., Shinozuka, M., Tierney, K., Wallace, W. A., & von Winterfeldt, D. (2003). A framework to quantitatively assess and enhance the seismic resilience of communities. *Earthquake Spectra*, *19*(4), 733–752. doi:10.1193/1.1623497

Cadini, F., Zio, E., & Petrescu, C. A. (2009). Using centrality measures to rank the importance of the components of a complex network infrastructure. In Roberto Setola and Stefan Geretshuber (Eds.), *CRITIS: International Workshop on Critical Information Infrastructures Security* (pp. 155–167). Berlin: Springer.

Cassottana, B., Biswas, P. P., Balakrishnan, S., Ng, B., Mashima, D., & Sansavini, G. (2022). Predicting resilience of interdependent urban infrastructure systems. *IEEE Access*, *10*, 116432–116442. doi:10.1109/ACCESS.2022.3217903

Chang, S. E., Seligson, H. A., & Eguchi, R. T. (1996). *Estimation of the economic impact of multiple lifeline disruption: Memphis light, gas and water division case study. Technical report.* Buffalo: National Center for Earthquake Engineering and Research, State University of New York at Buffalo.

Cimellaro, G. P., Mahin, S., & Domaneschi, M. (2019). Integrating a human behavior model within an agent-based approach for blasting evacuation. *Computer-Aided Civil and Infrastructure Engineering*, *34*(1), 3–20. doi:10.1111/mice.12364

Cimellaro, G. P., Tinebra, A., Renschler, C., & Fragiadakis, M. (2016). New resilience index for urban water distribution networks. *Journal of Structural Engineering*, *142*(8), C4015014. doi:10.1061/(ASCE)ST.1943-541X.0001433

Dehghani, N. L., Jeddi, A. B., & Shafieezadeh, A. (2021). Intelligent hurricane resilience enhancement of power distribution systems via deep reinforcement learning. *Applied Energy*, *285*, 116355. doi:10.1016/j.apenergy.2020.116355

Domaneschi, M., Cucuzza, R., Martinelli, L., Noori, M., & Marano, G. C. (2024). A probabilistic framework for the resilience assessment of transport infrastructure systems via structural health monitoring and control based on a cost function approach. *Structure and Infrastructure Engineering*, 1–13. doi:10.1080/15732479.2024.2318231

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. doi:10.1145/2347736.2347755

Dunn, S., Fu, G., Wilkinson, S., & Dawson, R. (2013). Network theory for infrastructure systems modelling. *Proceedings of the Institution of Civil Engineers: Engineering Sustainability*, *166*, 281–292.

Farnaaz, N., & Jabbar, M. A. (2016). Random forest modeling for network intrusion detection system. *Procedia Computer Science*, *89*, 213–217. doi:10.1016/j.procs.2016.06.047

Francis, R., & Bekera, B. (2014). A metric and frameworks for resilience analysis of engineered and infrastructure systems. *Reliability Engineering & System Safety*, *121*, 90–103. doi:10.1016/j.ress.2013.07.004

Gay, L. F., & Sinha, S. K. (2013). Resilience of civil infrastructure systems: Literature review for improved asset management. *International Journal of Critical Infrastructures*, *9*(4), 330–350. doi:10.1504/IJCIS.2013.058172

Goforth, E., Yosri, A., El-Dakhakhni, W., & Wiebe, L. (2022). Rapidity prediction of power infrastructure forced outages: Data-driven approach for resilience planning. *Journal of Energy Engineering*, *148*(3), 04022016. doi:10.1061/(ASCE)EY.1943-7897.0000836

Haggag, M., Yorsi, A., El-Dakhakhni, W., & Hassini, E. (2021). Infrastructure performance prediction under climate-induced disasters using data analytics. *International Journal of Disaster Risk Reduction*, *56*, 102121. doi:10.1016/j.ijdrr.2021.102121

Haimes, Y. Y., Horowitz, B. M., Lambert, J. H., Santos, J., Lian, C., & Crowther, K. (2005). Inoperability input-output model for interdependent infrastructure sectors. I: Theory and methodology. *Journal of Infrastructure Systems*, *11*(2), 67–79. doi:10.1061/(ASCE)1076-0342(2005)11:2(67)

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, *28*(1), 100. doi:10.2307/2346830

Holden, R., Val, D. V., Burkhard, R., & Nodwell, S. (2013). A network flow model for interdependent infrastructures at the local scale. *Safety Science*, *53*, 51–60. doi:10.1016/j.ssci.2012.08.013

Iuliis, M. D., Battegazzorre, E., Domaneschi, M., Cimellaro, G. P., & Bottino, A. G. (2023). Large scale simulation of pedestrian seismic evacuation including panic behavior. *Sustainable Cities and Society*, *94*, 104527. doi:10.1016/j.scs.2023.104527

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: Springer.

Kajitani, Y., & Tatano, H. (2009). Estimation of lifeline resilience factors based on surveys of Japanese industries. *Earthquake Spectra*, *25*(4), 755–776. doi:10.1193/1.3240354

Klise, K., Hart, D., Bynum, M., Hogge, J., Haxton, T., Murray, R., & Burkhardt, J. (2020). *Water network tool for resilience (WNTR) user manual. Technical report*. Albuquerque, NM: Sandia National Lab.(SNL-NM).

Labaka, L., Hernantes, J., & Sarriegi, J. M. (2016). A holistic framework for building critical infrastructure resilience. *Technological Forecasting and Social Change*, *103*, 21–33. doi:10.1016/j.techfore.2015.11.005

Liu, X., Ferrario, E., & Zio, E. (2019). Identifying resilient-important elements in interdependent critical infrastructures by sensitivity analysis. *Reliability Engineering & System Safety*, *189*, 423–434. doi:10.1016/j.ress.2019.04.017

Luiijf, E., Nieuwenhuijs, A., Klaver, M., van Eeten, M., & Cruz, E. (2009). Empirical findings on critical infrastructure dependencies in Europe. In Roberto Setola and Stefan Geretshuber (Eds.), *Critical information infrastructure security* (pp. 302–310). Berlin: Springer.

Manyena, B., Machingura, F., & O'Keefe, P. (2019). Disaster resilience integrated framework for transformation (drift): A new approach to theorising and operationalising resilience. *World Development*, *123*, 104587. doi:10.1016/j.worlddev.2019.06.011

Marasco, S., Cardoni, A., Zamani Noori, A., Kammouh, O., Domaneschi, M., & Cimellaro, G. P. (2021). Integrated platform to assess seismic resilience at the community level. *Sustainable Cities and Society*, *64*, 102506. doi:10.1016/j.scs.2020.102506

McDaniels, T., Chang, S., Peterson, K., Mikawoz, J., & Reed, D. (2007). Empirical framework for characterizing infrastructure failure interdependencies. *Journal of Infrastructure Systems*, *13*(3), 175–184. doi:10.1061/(ASCE)1076-0342(2007)13:3(175)

Mitsova, D. (2021). Integrative interdisciplinary approaches to critical infrastructure interdependency analysis. *Risk Analysis*, *41*(7), 1111–1117. doi:10.1111/risa.13129

Nan, C., & Sansavini, G. (2017). A quantitative method for assessing resilience of interdependent infrastructures. *Reliability Engineering & System Safety*, *157*, 35–53. doi:10.1016/j.ress.2016.08.013

Nicholson, C. D., Barker, K., & Ramirez-Marquez, J. E. (2016). Flow-based vulnerability measures for network component importance: Experimentation with preparedness planning. *Reliability Engineering & System Safety*, *145*, 62–73. doi:10.1016/j.ress.2015.08.014

Powell, D. R., DeLand, S. M., & Samsa, M. E. (2008). Critical infrastructure protection decision making. In John G. Voeller (Ed.), *Wiley handbook of science and technology for homeland security* (pp. 1–15). Hoboken, NJ: John Wiley & Sons.

Praks, P., Kopustinskas, V., & Masera, M. (2017). Monte-Carlo-based reliability and vulnerability assessment of a natural gas transmission system due to random network component failures. *Sustainable and Resilient Infrastructure*, *2*(3), 97–107. doi:10.1080/23789689.2017.1294881

Rahimi-Golkhandan, A., Aslani, B., & Mohebbi, S. (2021). Predictive resilience of interdependent water and transportation infrastructures: A sociotechnical approach. *Socio-Economic Planning Sciences*, *80*, 101166. doi:10.1016/j.seps.2021.101166

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLoS One*, *14*(1), e0210236. doi:10.1371/journal.pone.0210236

Santos, J. R. (2006). Inoperability input-output modeling of disruptions to interdependent economic systems. *Systems Engineering*, *9*(1), 20–34. doi:10.1002/sys.20040

Satopää, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *Proceedings - International Conference on Distributed Computing Systems* (pp. 166–171).

Sun, J., & Zhang, Z. (2020). A post-disaster resource allocation framework for improving resilience of interdependent infrastructure networks. *Transportation Research Part D: Transport and Environment*, *85*, 102455. doi:10.1016/j.trd.2020.102455

Svendsen, N., & Wolthusen, S. (2008). Multigraph dependency models for heterogeneous infrastructures. In E. Goetz & S. Shenoi (Eds.), *Critical infrastructure protection* (pp. 337–350). Boston, MA: Springer US.

Thompson, J. R., Frezza, D., Necioglu, B., Cohen, M. L., Hoffman, K., & Rosfjord, K. (2019). Interdependent critical infrastructure model (ICIM): An agent-based model of power and water infrastructure. *International Journal of Critical Infrastructure Protection*, *24*, 144–165. doi:10.1016/j.ijcip.2018.12.002

Thurner, L., Scheidler, A., Schafer, F., Menke, J. H., Dollichon, J., Meier, F., Meinecke, S., & Braun, M. (2018). Pandapower - An open-source Python tool for convenient modeling, analysis, and optimization of electric power systems. *IEEE Transactions on Power Systems*, *33*(6), 6510–6521. doi:10.1109/TPWRS.2018.2829021

Trivedi, S., Pardos, Z. A., & Heffernan, N. T. (2015). The utility of clustering in prediction tasks. *arXiv Preprint arXiv:1509.06163*.

Turati, P., Pedroni, N., & Zio, E. (2016). Dimensionality reduction of the resilience model of a critical infrastructure network by means of elementary effects sensitivity analysis. In Lesley Walls, Matthew Revie and Tim Bedford (Eds.), *European Safety and RELiability Conference 2016* (pp. 2797–2804). Boca Raton, FL: CRC Press, Taylor & Francis Group.

Wang, F., Magoua, J. J., & Li, N. (2022). Modeling cascading failure of interdependent critical infrastructure systems using HLA-based co-simulation. *Automation in Construction*, *133*, 104008. doi:10.1016/j.autcon.2021.104008

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165–193. doi:10.1007/s40745-015-0040-1

Yuan, C., & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. *J*, *2*(2), 226–235. doi:10.3390/j2020016

Zou, Q., & Chen, S. (2021). Resilience-based recovery scheduling of transportation network in mixed traffic environment: A deep-ensemble-assisted active learning approach. *Reliability Engineering & System Safety*, *215*, 107800. doi:10.1016/j.ress.2021.107800

Abbreviations

DT	decision tree
HLA	high-level architecture
IO	input-output
ML	machine learning
MLR	multiple linear regression
NN	neural network
PCS	prioritized consumer serviceability
RF	random forest
SVR	support vector regression