18TH WORLD CONFERENCE ON EARTHQUAKE ENGINEERING
WCEE2024
MILAN

MILAN, ITALY
30TH JUNE - 5TH JULY 2024

www.wcee2024.it

BUILDING SURROGATE MODELS FOR SEISMIC PERFORMANCE EVALUATIONS USING DIMENSIONALITY REDUCTION AND RF

M. Monsalve[1], E. Ferrario[2] & J.C de la Llera[3]

[1] Pontificia Universidad Catolica de Chile and Research Center for Integrated Disaster Risk Management (CIGIDEN) ANID/FONDAP/1522A0005, Santiago, Chile, mauricio.monsalve@cigiden.cl

[2] Ricerca sul Sistema Energetico - RSE S.p.A., Milano, Milano, Italy

[3] Pontificia Universidad Catolica de Chile and Research Center for Integrated Disaster Risk Management (CIGIDEN) ANID/FONDAP/1523A0009, Santiago, Chile

Abstract: *To assess the probabilistic seismic risk and resilience of an engineering system, the usual workflow considers the development of stochastically consistent seismic hazard scenarios, the estimation of earthquake demand on the structural and non-structural components, the evaluation of the operative performance of the affected system, and the final aggregation of outcome variables, say into exceedance probabilities or cost statistics. However, simulating large-scale, highly-detailed engineering systems, such as cities or large networks, may require considerable computational resources and time, especially as the number of evaluated earthquake scenarios increases in Monte Carlo simulations. This work proposes a new strategy to develop surrogate models for networked infrastructure systems using dimensionality reduction and random forests. The input to the surrogate model consists in the downtime of each element of the network, which is generally cheap to sample in case of earthquakes, for this requires evaluating the corresponding fragility curves and recovery times of the elements. The output consists in a loss measure (scalar or vector) that represents the behavior of the entire system. The surrogate model, then, reproduces the performance of the system to the affectation of its elements, which is typically the computationally most expensive task in simulating the system, because it often requires solving an optimization problem or reaching an equilibrium. Herein, the strategy proposed to generate the surrogate model roughly comprises two parts. First, it requires identifying structurally similar elements in the network, to later infer a reduced dimensionality input. And, second, it proposes projecting the input of downtimes to this lower dimensional representation and training a machine learning model to predict the loss measure. Results of this methodology are demonstrated using a practical application, by crafting a surrogate model of the electric power generation and transmission network in Chile, which has been thoroughly modeled using physics-based models and power flow equations. The surrogate model built following the proposed methodology was able to reduce the dimensionality of the problem from 1494 to merely 32 dimensions. The prediction errors were assessed and the predictions of the crafted surrogate model were unbiased.*

1 Introduction

To assess the probabilistic seismic risk and resilience of engineering systems, such as electric power network, public transportation and healthcare networks, the usual simulation workflow considers the following steps (Cardona et al, 2008; Marulanda et al, 2020; Liu et al, 2021; Silva-Lopez et al, 2022): (i) probabilistic sampling of consistent seismic hazard scenarios, (ii) the estimation of seismic loads on structural and non-structural components, (iii) the evaluation of the operative performance of the affected system, and (iv) the aggregation of outcome variables, such as exceedance probabilities, cost statistics, indicators and the identification of components-at-risk.

However, assessing the probabilistic seismic risk of a large-scale, highly-detailed engineering system, such as a city or a large network, may require prohibitive computational time and resources. This occurs because estimating the operational performance of the system (step (iii) above) consumes considerable time on each iteration. This situation worsens as the number of evaluated earthquake scenarios (step (i) above) increases in Monte Carlo simulations. In such case, reducing the computational time and resources needed to estimate the operational behavior of the system becomes desirable.

This work focuses in situation of large-scale networked infrastructure systems, these that comprise hundreds or more elements and span large areas. Such systems include urban drinking water networks, urban transportation networks, urban power distribution networks, urban telephony networks, regional highway networks, regional power transmission networks, etcetera. Of course, in the modeling of such systems, the number of elements considered depend on the level of abstraction suitable for the intended application, but this work focuses in systems comprising a considerable amount of elements.

The analysis of large-scale systems comes with the risk of falling under a *curse of dimensionality*, which is an umbrella term for a variety of problems arising when dealing with high dimensionality (Altman and Krzywinski, 2018; Chen et al, 2015; Kuo and Sloan, 2005; Wang, 2021). One type of curse of dimensionality occurs when systems comprise many elements and their computation, namely simulation, optimization or equilibrium finding, takes disproportionally (super-linearly, even exponentially) more time to solve (Bellman, 1961; Chen et al, 2015; Pereira and Pinto, 1991; Wang, 2021). This occurs because the description space (solution space or search space) of the problem grows exponentially with the number of dimensions (parameters, elements, details) considered.

Another type of curse of dimensionality affects learning (statistical and machine learning) models. Having too many dimensions (parameters) induces sparsity in the data (Altman and Krzywinski, 2018), models might fit the data too well and lack prediction power, a problem which is known as overfitting (Hawkins, 2004; Ying, 2019). Typical solutions to the problem of overfitting make use of dimensionality reduction, data augmentation and regularization (Ying, 2019). Another effective strategy includes using ensembles of models, where the consensus of possibly overfit models cancel each other estimation errors (Sollich and Krogh, 1995; Dietterich, 2002; Hastie et al, 2009).

Note that yet another type of curse of dimensionality stems from error or uncertainty propagation. The specification of large-scale systems requires specifying several parameters, all of which may be subjected to an estimation error, and which will be manipulated (say, through mathematical formulas) during system optimization or simulation, causing the propagation and enhancement of errors (Benke et al 2018). This work is not concerned with this problem, however.

This work is concerned with first two types of curse of dimensionality described. The objective is to bypass the first type, which is about computational complexity (runtime and computational resources). This is a typical use for surrogate models. However, the second type might affect the quality of the surrogate model crafted. In this sense, this work proposes using the network and spatial layout of the system components to inform a dimensionality reduction of the input space, using principal components (Jolliffe and Cadima, 2016), as well as the ensemble learning method of random forests (Breiman, 2001). Both techniques have been widely used in the development of surrogate models (Jun et al, 2020; Hou and Behdinan, 2022; Liu et al, 2021; Dasari et al, 2019; Hariri-Ardebili et al, 2021; Zheng et al, 2019).

The methodology introduced in this work is demonstrated through the development of a surrogate model for the electric power generation and transmission network in Chile, which supplies with electricity most of the territory. The network considered comprises 1494 nodes (substations and power plants), which were the

elements at risk, and 1195 edges (power lines), which represents its 2019 snapshot. The system model has been characterized in its entirety, both physically and operationally (Ferrario et al, 2019; Ferrario et al, 2020).

Following the proposed methodology, the surrogate model is built using the spatial and network distribution of the network elements (generating plants, substations and lines) and its simulated performance on 20000 seismic scenarios, sampled using importance sampling and following realistic recurrence relations for Chilean subduction seismicity (Poulos et al, 2019). The resulting surrogate model takes a vector having the downtime in hours of each node of the network, which has 1494 dimensions, and reduces it to 32 dimensions, to then estimate the loss measure, which was chosen to be the energy not supplied to the Greater Valparaíso conurbation. The quality of the surrogate model is measured and validated through its cross-validation prediction error.

2 Proposed methodology

The general objective of this methodology is to guide the crafting of surrogate models using information about the layout of the engineering system under consideration, namely, the location and interconnections of its components, and a performance dataset for the system, in which each sample datum consists in the damage or disruption level of each component of interest and the loss outcome of the system.

2.1 Proposed guide

In particular, the methodology consists in the following steps:

1. Devise a similarity matrix for the components of the system. The idea here is that this similarity matrix may help identify functionally equivalent components. The components that must be considered in the matrix are the elements-at-risk only. The similarity between two components may be defined according to the available data. For specific cases:
    1. If the components have known spatial locations, the similarity measure may consider that closer elements are more similar than distant ones. For instance, if $d(u,v)$ is the Euclidean distance between components *u* and *v*, a similarity measure such as $\exp(-\alpha d^2(u,v))$ may be used.
    2. If the components belong in a network, the similarity measure may evaluate their structural equivalence or similarity (Lorrain et al, 1971; Audenaert et al, 2018). If the set of neighbors of node *n* is denoted by $\Gamma(n)$, then Jaccard's score, $J(u,v)=(\Gamma(u)\cap\Gamma(v))/(\Gamma(u)\cup\Gamma(v))$, may be used as similarity measure. This score may be improved if a less strict node similarity measure is used.
    3. If data about the behaviour or performance of components *u*, *v* are available, then the correlation may be used as similarity measure.
    4. If similarity measures measuring different aspects are available, e.g., spatial proximity and node similarity, then the measures may be combined.
2. Compute the top eigenvectors of the matrix. This step is essentially principal components analysis (Fan et al, 2014). The eigenvectors may be ranked by their eigenvalues (the greater, the better) or by their relation to an area or element of interest (this requires looking at specific components within the vectors obtained).
3. Simulate a number of scenarios using the original, detailed system model, and estimate the loss for each scenario. The number of scenarios to consider may depend on the amount of data needed for statistical fitting, the probability of generating damaging events, and the resources available. The loss measures (may just be one measure, i.e., a scalar) should depend on the intended use for the model. In the generated dataset, each entry should consist in the damage or disruption level of each component of interest of the system (e.g., the downtime of each element) and the loss measures obtained.
4. Generate a low dimensionality performance dataset. This must be done by projecting each vector of damage or disruption levels on the eigenvectors obtained (a simple dot product), plus the summation of all the components of the vector. Each entry should retain, however, the loss measures.

5. Train a random forest model on the low dimensionality dataset previously generated. This random forest model will estimate the *expected* loss measures given a vector of damage or disruption levels.
6. Estimate the error model of the model by performing cross-validation. Cross-validation is a technique for assessing the prediction error of a model by repeatedly splitting the dataset into training and testing sets, and each time fitting the model in the training set to test it on the testing set (Browne, 2000; Hawkins et al, 2010). For simplicity, 10-fold cross-validation may be used. Then, by pooling the prediction errors of the random forest models, it is possible to assess, for each loss measure, whether its error is additive, multiplicative or affine, and what is the magnitude of this error.

2.2 How to use the crafted surrogate model

To evaluate the system performance using the surrogate model, the following steps may be followed:

1. From the seismic event, estimate the seismic loads (intensity measure) on each component-at-risk of the system.
2. From the seismic loads, estimate the damage or disruption level of the components-at-risk of the system. This has to be in the same units used to train the surrogate model. For example, it could be downtime in hours or an index for damage level (0 for no damage, then 1, 2, 3, etc., for additional severity levels).
3. Generate a low dimensional representation of the previous step. This is done by applying the dot product between the damage or disruption level vector and the eigenvectors, plus the additional dimension that results from summing all the components in the vector.
4. Estimate the expected loss measures using the random forest model trained on the entire original low dimensional dataset.
5. Use the error model defined previously to generate a random loss vector using the expected loss measures and the error model, or provide confidence bounds for the expected loss vector.

3 Electric power network

As a practical application of the methodology previously introduced, a surrogate model of the electric power network of Chile is developed following the methodology. This section describes the system model used and the dataset built using its performance on a sample of seismic scenarios.

3.1 Description of the model

The electric power network of Chile, called *Sistema Eléctrico Nacional* (National Electric System) and abbreviated *SEN*, supplies with electricity most of the country and even sells electricity to neighboring country of Argentina. The SEN spans about 3100 km in length from North to South, serves about 98% of the population of Chile and was formed by merging the previous Greater North and Central-Southern electric systems (Coordinador Eléctrico Nacional, 2021).

The model of the SEN considered in this work consists in its 2019 snapshot (Ferrario et al, 2019; Ferrario et al, 2020). This is a network model comprising 1494 nodes and 1195 links. The 1494 nodes are the elements at seismic risk, and consists in 994 substations and 500 power generation plants. The power plants are of various kinds, ranging from hydroelectric to wind generation. Fig. 1 illustrates the spatial distribution and network layout of the 2019 SEN.

Figure 1. The electric power network of Chile (SEN).

Note that Fig. 1 also illustrates the administrative division of the covered area of the country in 2019. The map leaves out the southernmost part of Chile, because it is not supplied by the SEN. The line crossing the eastern border, which is on the Andes mountain range, supplies a northern location in Argentina with electricity.

The system model is implemented as a mathematical optimization model, by which the system tries to satisfy the demand using the least cost to do so. This is done by solving a direct current, optimal power flow (DC-OPF) model (Ferrario et al, 2019; Ferrario et al, 2020). The objective function, to minimize, considers the generation cost and load shedding. The constraints ensure the flow conservation at each node and model the energy loss due to line electrical resistance. Additional constraints involve capacity, such as line capacity, generation capacity and client consumption capacity, which must not be exceeded.

In addition to the above, the hourly demand during year 2019 has also been collected.

3.2 Seismic scenarios considered

The system model was evaluated on a set of 20000 seismic scenarios which were sampled following the earthquake recurrence relations of Poulos et al (2019). These recurrence relations indicate the probability of mainshocks originating from the subduction mechanism between the oceanic and continental plates. The relations indicate event frequencies for 7 zones (3 coastal and 4 inland) and according to moment magnitudes (M_w) above 5.0, following the Gutenberg-Richter relation (Gutenberg and Richter, 1944).

The 20000 scenarios were sampled following the aforementioned recurrence relations. However, earthquake magnitudes were sampled uniformly between magnitudes M_w 5.0 and M_w 9.6, in other words, following an importance sampling scheme. This was done because it was necessary to sample higher magnitude mainshocks, which in practice are much less frequent than lower magnitude ones, since the performance of the electric power network is unaffected by events except for these of higher magnitude ($M_w \geq 8$), in which the disruption and performance degradation may be massive. This configures what is often referred to as an extreme, rare events situation (Broska et al, 2020; Glette-Iversen and Aven, 2021).

For each scenario, the seismic load on each node was evaluated (following the use of a GMPE). From the application of the fragility curves of each node, its damage state and downtime were estimated. Then, by assigning a random hour of occurrence in 2019, the power flow optimization problem was solved to check if the system could meet the demand. The chosen loss measure was the energy not supplied (ENS) to the Greater Valparaíso conurbation.

Fig. 2 shows the ENS to the Greater Valparaíso conurbation for the 20000 seismic scenarios considered. As the plot shows, most of the losses are concentrated in earthquakes with magnitudes $M_w \geq 8.0$. Overall, only 3307 (16.5%) of the scenarios had any losses associated. Also, note that for magnitudes $M_w \geq 9.3$ there is a reduction of damaging scenarios, which is caused by the spatial distribution of the most damaging scenarios (which are dominated by events far from the Valparaíso region).

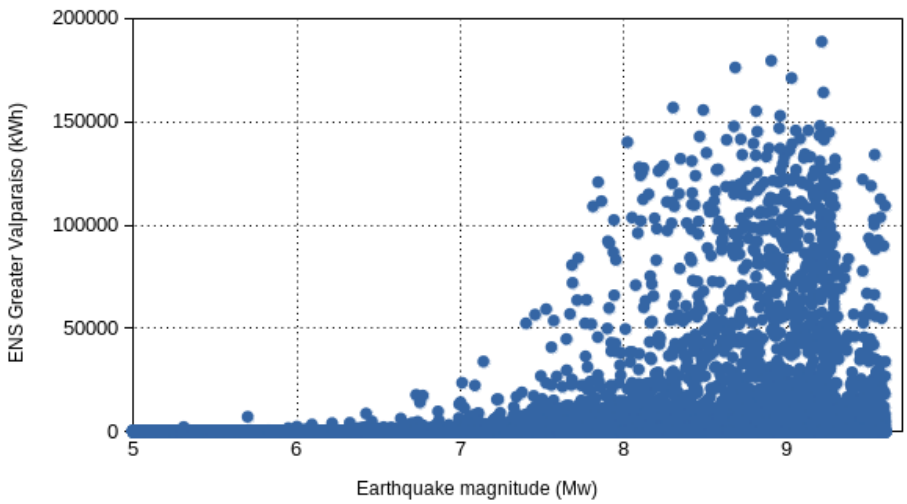

Figure 2. ENS to the Greater Valparaíso conurbation by scenario.

4 Application and results

4.1 Development of the surrogate model

The first step of the methodology consists in defining a similarity matrix for the elements-at-risk, which are the 1494 nodes in this case. Since each node has a spatial location as well as a position in the network, the similarity measure used combined both.

For the spatial location, the similarity measure chosen was

$$S_{\text{dist}}(u,v)=\exp\left(-\frac{d(u,v)^2}{471^2}\right), \tag{1}$$

where $d(u,v)$ is the Euclidean distance in kilometers between elements u and v.

For the network position, the similarity measure chosen was

$$S_{\text{netw}}(u,v)=\frac{\Gamma(u)\cap\Gamma(v)}{\Gamma(u)\cup\Gamma(v)}, \tag{2}$$

in other words, by Jaccard's similarity.

Finally, the combined similarity was defined as

$$S(u,v)=1-(1-S_{\text{dist}}(u,v))(1-S_{\text{netw}}(u,v)). \tag{3}$$

The reasoning behind Eq. (3) is that if two nodes are extremely close or structurally equivalent in terms of network position, then the nodes should be consider to fulfil practically the exact same role and, hence, should be considered perfectly similar.

Using the similarity measure S, a similarity matrix of dimensions 1494×1494 was defined and its eigen-decomposition was calculated. It turned out that only 40 eigenvectors with non zero eigenvalues were found. Moreover, the first 10 explained 98.5% of the 1494 dimensions. Still, the first 31 eigenvectors were used, which where the only associated with eigenvalues of 0.1 or more (this was the cut criterion used).

Following the 31 eigenvectors chosen and an additional vector of 1s (to sum all the damage or disruption levels in the affectation vector), the dataset with the system performance on the 20000 was reduced from having 1494+1 columns (1494 elements-at-risk plus the loss measure ENS) to a dataset of 32+1 columns. The random forest model was trained on this reduced dataset.

4.2 Prediction error

The next step in the proposed methodology consists in assessing the prediction error and defining an error model for the random forest predictions. Thus, a 10-fold cross-validation procedure was performed. In each iteration, a random forest model was fitted on the training set and its predictions were tested on the testing set. The pooled predicted and actual ENS in the 10 testing sets are shown in Fig. 3.

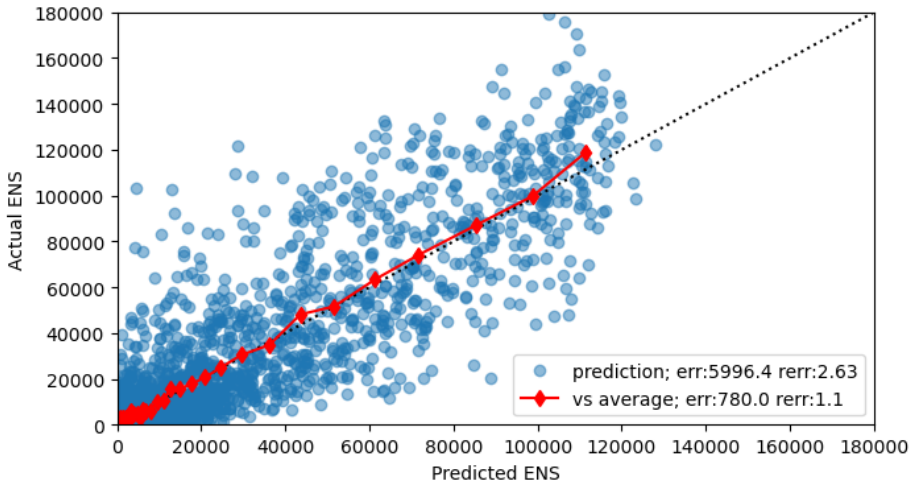

Figure 3. Prediction results of 10-fold cross validation.

The results suggest that the error model should be multiplicative instead of additive or affine. In fact, it is suggested that the predictions of the random forest model should be multiplied by a random number generated following a normal distribution with mean 1 and standard deviation of 2.91.

The plot shown in Fig. 3 also suggests that the predictions by the random forest model are unbiased. To assess this, all the surrogate model predictions were sorted by their predicted ENS, they were partitioned in 250 contiguous bins, and then, for each bin, the actual (true) ENS were averaged. The result was the red line in Fig. 3, which, as shown, closely follows the diagonal line. In other words, the model predictions are unbiased.

Note that the random forest models were configured to consist of 150 random regression trees inside (which are like decision trees, but whose leaves have linear regression functions) with depth of 5 levels. These settings were hand-chosen to yield a decent performance, but were not chosen automatically, so they are unlikely to be optimal.

5 Conclusions

Assessing the probabilistic seismic risk of a large-scale, highly-detailed engineering system, such as city or region-wide infrastructure network, may require prohibitive computational time and resources. This cost comes from estimating the operational performance of the system under each seismic scenario. Therefore, to reduce the overall computational time needed to carry the probabilistic seismic risk assessment of the system, a surrogate model of the system may be used instead.

This work introduced a methodology to develop surrogate models that take as input the downtimes of the elements of the system and predict the overall operative loss in the performance of the system, according to a pre-specified loss measure.

The methodology proposes crafting surrogate models using dimensionality reduction and machine learning. The dimensionality reduction step takes into account the spatial distribution and network topology of the system, to identify which elements of the input can be combined using principal components analysis, to yield a low dimensionality projection of the downtimes. The machine learning step consists in training a random forest model to learn the association between the low dimensional input and the loss measure, as well as to learn the prediction error. The resulting surrogate models are expected to be robust against the curse of dimensionality.

As a practical application, a surrogate model for the electric power network of Chile was developed following the proposed methodology. The system model comprises 1494 nodes (substations and power plants), which were the elements at risk, and 1195 edges (power lines). Through the application of the methodology, the surrogate model takes a vector having the downtime in hours of each node of the network, totalling 1494 components, and reduces it to 32 dimensions, to then estimate the loss measure, which was chosen to be the energy not supplied to the Greater Valparaíso conurbation. Cross-validation error estimation validated the quality of the surrogate model developed following the proposed methodology.

6 Acknowledgements

This work has been sponsored by the Chilean government through the Research Center for Integrated Disaster Risk Management (CIGIDEN), ANID/FONDAP/1523A0009, and the research project Multiscale earthquake risk mitigation of healthcare networks using seismic isolation, ANID/FONDECYT/1220292.

7 References

Altman N., Krzywinski M. (2018). The curse(s) of dimensionality. *Nature Methods*, 15(6):399-400.

Audenaert P., Colle D., Pickavet M. (2018). Regular equivalence for social networks. *Applied Sciences,* 9(1):117.

Bellman R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton.

Benke K.K., Norng S., Robinson N.J., Benke L.R., Peterson T.J. (2018). Error propagation in computer models: analytic approaches, advantages, disadvantages and constraints. *Stochastic Environmental Research and Risk Assessment*, 32:2971-85.

Breiman L. (2001). Random forests. *Machine learning*, 45:5-32.

Broska L.H., Poganietz W.R., Vögele S. (2020). Extreme events defined—A conceptual discussion applying a complex systems approach. *Futures*, 115:102490.

Browne M.W. (2000). Cross-validation methods. *Journal of mathematical psychology* 44(1):108-32.

Cardona O.D., Ordaz M.G., Yamín L., Arámbula S., Marulanda M.C., Barbat A. (2008). Probabilistic seismic risk assessment for comprehensive risk management: modeling for innovative risk transfer and loss financing mechanisms. *Proceedings ot The 14th World Conference on Earthquake Engineering*.

Chen S., Montgomery J., Bolufé-Röhler A. (2015). Measuring the curse of dimensionality and its effects on particle swarm optimization and differential evolution. *Applied Intelligence*, 42:514-26.

Coordinador Eléctrico Nacional (2021). Sistema Eléctrico Nacional (SEN). Website of the Coordinador Eléctrico Nacional, https://www.coordinador.cl/sistema-electrico/ (retrieved October 31, 2023).

Dasari S.K., Cheddad A., Andersson, P. (2019). Random forest surrogate models to support design space exploration in aerospace use-case. *Proceeding of the 15th IFIP WG 12.5 International Conference, Artificial Intelligence Applications and Innovations 2019*, Hersonissos, Crete, Greece. Springer International Publishing.

Dietterich T.G. (2002). Ensemble learning. *The handbook of brain theory and neural networks,* 2(1):110-25.

Fan Z., Xu Y., Zuo W., Yang J., Tang J., Lai Z., Zhang D. (2014). Modified principal component analysis: An integration of multiple similarity subspace models. *IEEE transactions on neural networks and learning systems,* 25(8):1538-52.

Ferrario E., Poulos A., de la Llera J.C., Lorca A., Oneto A., Magnere C. (2019). Representation and modeling of the chilean electric power network for seismic resilience analysis. *Proceedings of the 29th European Safety and Reliability Conference (ESREL)*. Research Publishing Services.

Ferrario E., Monsalve M., Poulos A., de la Llera J.C., Sansavini G. (2020). Estimating the impact of earthquake-induced power outages on different economic sectors in Chile. *Proceedings of the 30th European Safety and Reliability Conference (ESREL) and the 15th Probabilistic Safety Assessment and Management Conference (PSAM) 2020*. Research Publishing Services.

Glette-Iversen I., Aven T. (2021). On the meaning of and relationship between dragon-kings, black swans and related concepts. *Reliability Engineering & System Safety,* 211:107625.

Gutenberg B., Richter C.F. (1944). Frequency of earthquakes in California. *Bulletin of the Seismological society of America*, 34(4):185-8.

Hariri-Ardebili M.A., Mahdavi G., Abdollahi A., Amini A. (2021). An RF-PCE hybrid surrogate model for sensitivity analysis of dams. *Water*, 13(3):302.

Hastie T., Tibshirani R., Friedman J. (2009). Ensemble learning. *The elements of statistical learning: data mining, inference, and prediction*, 605-24.

Jolliffe I.T., Cadima J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences,* 374(2065):20150202.

Hawkins D.M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1-2.

Hawkins D.M., Basak S.C., Mills D. (2003). Assessing model fit by cross-validation. *Journal of chemical information and computer sciences,* 43(2):579-86.

Hou C.K., Behdinan K. (2022). Dimensionality Reduction in Surrogate Modeling: A Review of Combined Methods. *Data Science and Engineering*, 7(4):402-27.

Jun T.A., Gang S.U., Liqiang G.U., Xinyu W.A. (2020). Application of a PCA-DBN-based surrogate model to robust aerodynamic design optimization. *Chinese Journal of Aeronautics,* 33(6):1573-88.

Kuo F.Y., Sloan I.H. (2005). Lifting the curse of dimensionality. *Notices of the AMS*, 52(11):1320-8.

Liu Y., Li L., Zhao S., Song S. (2021). A global surrogate model technique based on principal component analysis and Kriging for uncertainty propagation of dynamic systems. *Reliability Engineering & System Safety,* 207:107365.

Liu Y, Wotherspoon L, Nair NK, Blake D. (2021). Quantifying the seismic risk for electric power distribution systems. *Structure and Infrastructure Engineering*, 17(2):217-32.

Lorrain F, White HC. (1971). Structural equivalence of individuals in social networks. *The Journal of mathematical sociology,* 1(1):49-80.

Marulanda M.C., de la Llera J.C., Bernal G.A., Cardona O.D. (2021). Uncertainty Range in Probabilistic Seismic Risk Metrics Resulting from Multiple Hazard Models. *Natural Hazards,* 2021.

Pereira M.V., Pinto L.M. (1991). Multi-stage stochastic optimization applied to energy planning. Mathematical programming, 52:359-75.

Poulos A., Monsalve M., Zamora N., de la Llera J.C. (2019). An updated recurrence model for Chilean subduction seismicity and statistical validation of its Poisson nature. *Bulletin of the Seismological Society of America,* 109(1):66-74.

Silva-Lopez R., Bhattacharjee G., Poulos A., Baker J.W. (2022). Commuter welfare-based probabilistic seismic risk assessment of regional road networks. *Reliability Engineering & System Safety,* 227:108730.

Sollich P., Krogh A. (1995). Learning with ensembles: How overfitting can be useful. *Advances in neural information processing systems*, 8.

Wang Q. (2021). Knowledge-based approach for dimensionality reduction solving repetitive combinatorial optimization problems. *Expert Systems with Applications*, 184:115502.

Ying X. (2019). An overview of overfitting and its solutions. In *Journal of physics: Conference series*, 1168: 022022. IOP Publishing.

Zheng Y., Fu X., Xuan Y. (2019). Data-driven optimization based on random forest surrogate. Proceeding of the *2019 6th international conference on systems and informatics (ICSAI),* pp. 487-491. IEEE.