Available online at www.eccomasproceedia.org
Eccomas Proceedia COMPDYN (2025) 554-562

ECCOMAS
Proceedia

COMPDYN 2025
10th ECCOMAS Thematic Conference on
Computational Methods in Structural Dynamics and Earthquake Engineering
M. Papadrakakis, M. Fragiadakis (eds.)
Rhodes Island, Greece, 15-18 June 2025

USE OF HPC IN THE DEVELOPMENT OF LARGE DATASETS FOR THE DEVELOPMENT OF PREDICTIVE MODELS THROUGH MACHINE LEARNING ALGORITHMS

Duan Calitz[1], Asma Yahiaoui[2], George Markou[3,1], Saida Dorbani

[1] Department of Civil Engineering, University of Pretoria, Private Bag X20, 002 Hatfield, Pretoria 0028, South Africa;
email: u19106743@tuks.co.za; george.markou@up.ac.za

[2]Department of Structures and Materials, Built Environmental Research Laboratory, Civil Engineering Faculty, University of Science and Technology Houari Boumediene (USTHB), BP 32 El Alia, Bab Ezzouar 16111, Algiers, Algeria
email: yahiaouiasma14@gmail.com; ayahiaoui1@usthb.dz

[3]Department of Civil Engineering and Geomatics, Cyprus University of Technology, Limassol 3036, Cyprus;
george.markou@cut.ac.cy

Abstract

In the modern world where the populations are increasing along with the built environment, the importance of seismic evaluation becomes more crucial in achieving safety and sustainability. The recent seismic activity in Turkey has brought the risk of earthquakes back into public consciousness, especially due to the extreme fatality rates. Typically, finite element analysis would be used to evaluate structures on a case-to-case basis. This procedure is usually time-consuming where the civil engineer is required to have a specific set of skills and know-how so as to obtain realistic and error-free results through a pushover analysis. One possible solution is to replace the tedious process of performing a pushover analysis through the use of a predictive machine learning-generated model. For the needs of this research work, 1248 reinforced concrete buildings were developed using such an approach. Each structure was unique in geometry, covering different footprint areas, and took into account the presence of different configurations of shear walls, material properties, and reinforcement ratios. Each base structure was analyzed around 20,000 times for different combinations of material properties and reinforcement detailing. The analyses were performed on LUMI supercomputer where the P-δ curves were recorded and stored in a relevant dataset that contains more than 21 million analysis results. This research work presents the method through which this large dataset was developed and discusses the future objectives of this research work.

Keywords: P-δ curves, Predictive machine learning-generated model, reinforced concrete buildings.

ISSN:2623-3347 © 2025 The Authors. Published by Eccomas Proceedia.
Peer-review under responsibility of the organizing committee of COMPDYN 2025.
doi: 10.7712/120125.12429.26143

1 INTRODUCTION

The development of P-δ curves through pushover analysis is a fundamental approach for assessing the seismic vulnerability of reinforced concrete (RC) buildings. Pushover analysis provides a computationally efficient method for evaluating structural performance under increasing lateral loads, allowing for a detailed understanding of nonlinear behavior and failure mechanisms. By incrementally applying horizontal forces to a numerical model of a structure, pushover analysis enables the derivation of P-δ curves, which illustrate the relationship between lateral displacement and base shear. These curves serve as a critical tool for identifying structural weaknesses and estimating the progression of damage under seismic excitation. According to Eurocode 8 [1], the evaluation of structural stability under seismic loading is crucial, as excessive lateral displacements can compromise the overall integrity of a building.

The accuracy of P-δ curves depends on several parameters, including material properties, geometric configurations, and boundary conditions. However, comprehensive studies on the influence of these factors in RC buildings remain limited due to the significant computational effort required to model various structural typologies [2-6]. The challenge lies in developing an efficient method capable of predicting reliable P-δ curves across a wide range of building configurations. Additionally, generating models for hundreds or thousands of structures presents difficulties in generalizing results for different RC building types. Addressing these challenges is essential for enhancing seismic risk assessment methodologies and improving the resilience of RC buildings against earthquakes, in line with the performance-based principles outlined in Eurocode 8.

The rapid expansion of machine learning (ML) applications in civil engineering has enabled researchers to solve complex problems and gain deeper insights into the significance of various parameters that were previously overlooked [7–13]. However, a fundamental limitation of most predictive ML models is their reliance on interpolation, meaning they can only generate accurate predictions within the range of the input parameters they were trained on. While extrapolation is theoretically possible, it often leads to significant inaccuracies due to the lack of representative training data.

A well-known example is the dataset developed by Asteris [14], which examines the fundamental period of infilled reinforced concrete (RC) structures based on 4,026 cases, considering variations in five key input parameters: building height, opening percentage, span length, number of spans, and masonry wall stiffness. The main challenge with predictive models trained on this dataset [15, 16] is that their validity is restricted to structures with identical material properties and cross-sectional dimensions as those in the original dataset. Any deviation in these properties alters the mass or stiffness of the structure, directly affecting its fundamental period of vibration, which is highly sensitive to these parameters. This constraint highlights the challenge of developing ML models that can generalize across a broader range of structural configurations while maintaining predictive accuracy.

Therefore, developing a large dataset is beneficial in allowing the generalization of several building types and seismic demands. To this end, this research paper aims to develop a large dataset with more than 21 million analyses of the P-δ curves that cover different materials, mechanicals, and geometric characteristics of RC structures ranging from low-rise to high-rise buildings using the LUMI supercomputer. For this purpose, a pushover analysis is conducted first on the 1248 RC buildings modeled using Reconan FEA software [17].

Developing a large dataset is crucial for enabling the generalization of various building types and seismic demands. In this context, the objective of this research is to generate an extensive dataset comprising more than 21 million analyses of P-δ curves, encompassing a wide range of material properties, mechanical characteristics, and geometric configurations of RC structures.

The dataset covers structures from low-rise to high-rise buildings, utilizing the computational power of the LUMI supercomputer.

To achieve this, pushover analyses are first performed on 1,248 RC buildings modeled using Reconan FEA software [17]. These analyses serve as the foundation for constructing a comprehensive dataset that captures the nonlinear behavior of RC structures under seismic loading, ultimately enhancing the predictive capabilities of ML-based seismic vulnerability assessment models.

The motivation for integrating ML-based predictive modeling stems from the limitations of empirical formulas and physics-based simulations in handling large-scale structural datasets. While traditional FEA methods provide accurate results, they are computationally expensive and difficult to generalize across diverse structural configurations.

In this study, the dataset generated using HPC-driven simulations will be leveraged to train ML models capable of predicting P-δ curves based on input features such as material properties, geometry, and reinforcement ratios. The choice of ML algorithms will be guided by:

- Robustness in handling high-dimensional input features (e.g., Gradient Boosting, Deep Neural Networks)
- Computational efficiency for real-time predictions
- Ability to generalize across diverse RC structural types

Future work will compare multiple ML models to determine the most accurate and computationally efficient predictive framework.

2 NUMERICAL CAMPAIGN AND USE OF HPC

For the creation of the base dataset used to generate over 21 million analysis results, various structural types were modeled using Reconan FEA [17]. This software utilizes the Natural Beam-Column Flexibility-Based finite element [19], which reduces computational demands, and derives high numerical accuracy when modeling RC structures. A total of 1,248 RC buildings were modeled following the guidelines of Eurocode 2 and 8 [1,18].

Concrete compressive strength	60 MPa
Elastic modulus of concrete	44 GPa
Steel ultimate tensile strength	600 MPa
Elastic modulus of steel	210 GPa
Beams dimensions	0.25 m x 0.55 m
Columns dimensions	0.35 m x 0.35 m
Longitudinal reinforcement ratio of the beam, columns, and SW	4%
Shear reinforcement ratio of the beams, columns, and SW	2%
Poisson ratio of concrete	0.2
Poisson ratio of steel	0.3

Table 1: Maximum mechanical and geometrical characteristics of each of the base RC buildings

The dataset aims to be used for training predictive models that will be able to compute the P-δ curves of RC buildings by considering the variability across structural and material input parameters. These include average floor height, which is set at 2.5 m for buildings with 1, 6, and 12 stories, as well as building heights of 2.5, 15, and 30 m for the 1, 5, and 8-storey configurations. Additionally, building widths and lengths are modeled with values of 3, 6, 8, 24, and 48 m, with bay lengths of 3, 6, and 8 m, resulting in 1, 3, 4, 6, and 8 bays. This research work also examines four different shear wall (SW) percentages: 0%, 50%, 75%, and 100%. The 75% SW configuration is only considered for buildings with an average floor height of 3 m, with 1, 5, and 10 storeys, and building heights of 3, 15, and 30 m. The maximum mechanical and geometric characteristics of the 1,248 structures were outlined in Table 1 and were used to determine the maximum P-δ curve response from each base building model.

The determination of SW cross-sections was based on the specified SW percentage, except for the 100% configuration, where the dimensions of all SW sections were defined to be equal to 0.6 m x 3 m. Fig. 1 illustrates the geometry of the RC building models developed for the needs of this research work for four different SW percentages.

To predict the P-δ curves, it is necessary to determine the maximum lateral load that the structures can withstand before reaching their ultimate capacity or collapse, as well as the maximum roof displacement under these lateral loading conditions. These values are obtained through nonlinear static pushover analysis, conducted using Reconan FEA [17], for the cases where the buildings assume maximum material and reinforcement ratios.

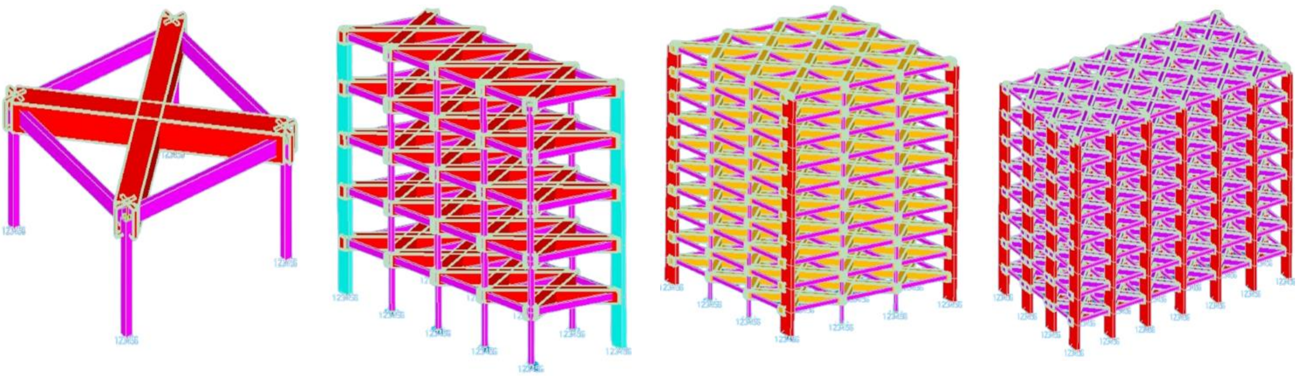

Figure 1: Different geometry of RC buildings with percentages of SW: 0%, 50%, 75%, and 100%

It is important to note here that all RC buildings were assumed to be fixed at their bases and that the plan view did not foresee any asymmetries. In addition to that, the seismic loads were distributed according to the seismic load distribution that is described by Eurocode 8 [1]. Furthermore, the slabs were modeled as diaphragms through the use of rigid elements.

2.1 The use of LUMI HPC in performing the analyses

The utilization of High-Performance Computing (HPC) resources was integral to the success of this research project, enabling the efficient execution of extensive Finite Element Analysis (FEA) simulations. The LUMI-C supercomputer was employed to handle the computational demands of analyzing 1,240 base structural models, each subjected to an average of 20,000 different configurations.

The workflow was designed to maximize computational efficiency while adhering to system constraints. For each of the 1,240 base models, a job was queued on the HPC system. Each job was responsible for the random sampling of the 20,000 configurations, which involved varying material properties and structural configurations to capture a wide range of potential seismic responses.

To manage this substantial computational task, each job was allocated 256 CPUs. These CPUs acted as individual workers, with each CPU dedicated to running the FEA software on one of the 20,000 configurations. This parallel processing approach significantly reduced the time required to complete the simulations, as multiple configurations could be analyzed simultaneously.

However, the number of jobs scheduled to run concurrently was limited to 70 base models in parallel. This limitation was imposed due to a file limit of 2 million files on the resource node at any given time. By restricting the number of simultaneous jobs, the process was able to run efficiently without exceeding this file limit that is set by default on the LUMI supercomputer. Additionally, this approach ensured that unnecessary files were cleared upon completion of each job, maintaining system performance and resource availability.

Once the FEA simulations were completed for a given configuration, the output data were retrieved and stored for further analysis. This data included detailed information on the P-δ curves, which were subsequently compiled into one of the largest datasets of its kind found currently in the international literature. In addition to the variation of the input features described in the previous section, this generated dataset includes the variation of 7 more input parameters, which are described in Table 2. The workflow per job as described herein is outlined in Figure 2.

It must be stated at this point that during the LUMI analysis some models were computationally demanding while others were able to run significantly faster, leading to time constraints. This had as a result a limited number of analyses performed for the case of some large RC models.

Concrete compressive strength	20, 40, 60 MPa
Elastic modulus of concrete	20, 36, 44 GPa
Steel ultimate tensile strength	400, 500, 600 MPa
Elastic modulus of steel	190, 200, 210 GPa
Longitudinal reinforcement ratio of the beam (%)	0.13, 2, 4
Longitudinal reinforcement ratio of the columns and SW	0.2, 2, 4
Shear reinforcement ratio of the beam, columns, SW	0.2, 2, 4

Table 2: Characteristics of the developed dataset

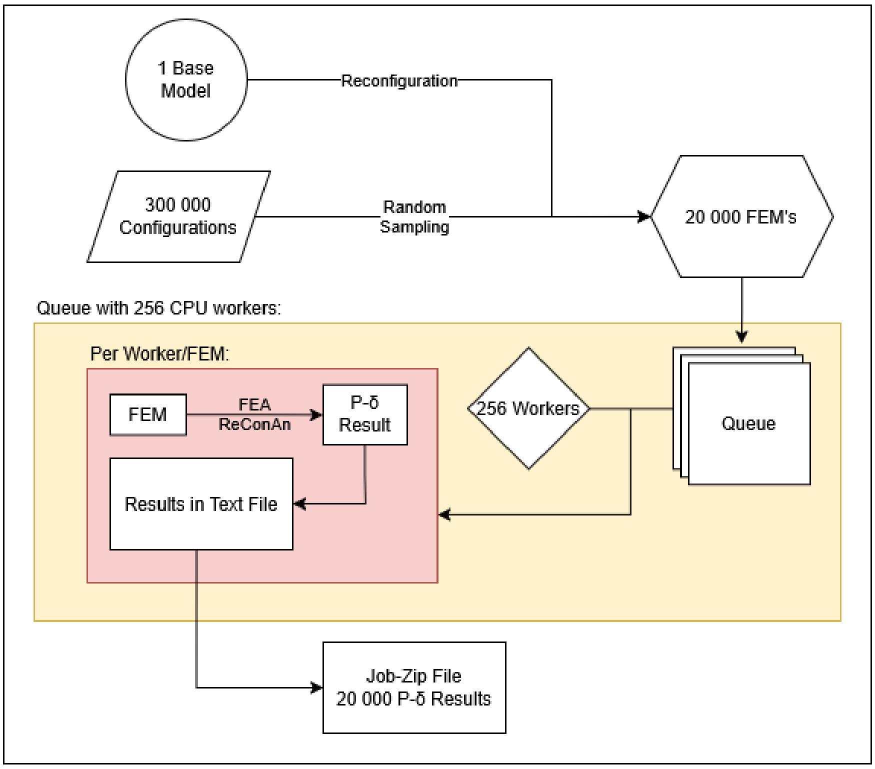

Figure 2: HPC workflow, per scheduled job

2.2 Programming and Results

The project's computational workflow was meticulously designed and implemented using a combination of scheduling batch scripts and Python processing scripts to ensure efficient execution of large-scale simulations. These scripts were responsible for orchestrating the submission, execution, and post-processing of numerous computational tasks in an HPC environment.

To effectively distribute and manage the computational workload, the SLURM job scheduler was employed. SLURM was used to schedule, queue, and allocate resources for each job, ensuring optimal utilization of the available computing infrastructure. Each job submission required the specification of CPU cores, memory allocation, wall-time limits, and dependency constraints to prevent resource conflicts and maximize throughput.

Each computational job involved the execution of processing scripts written in Python, which were responsible for running FEA simulations on RC models. These processing scripts were designed to dynamically load input parameters, execute simulations, and store results in structured output files.

Given the large scale of the study, it was essential to run multiple jobs simultaneously to reduce the overall computational time. The scheduling strategy was configured to execute 70 jobs in parallel, with each job focusing on analyzing a single RC model. This parallelization was crucial in handling the vast number of simulations efficiently. However, the choice of 70 concurrent jobs was not arbitrary; it was imposed due to a file limit restriction on the resource nodes. This limitation ensured that the system remained stable and prevented excessive file generation that could disrupt ongoing processes.

Over the course of the study, 21 million individual analyses were completed, consuming a total of 380,000 CPU hours. These CPU hours were distributed across multiple HPC nodes, allowing for efficient execution of computations in parallel. It is important to note that not all allocated CPU hours were dedicated to actual result generation. A small portion was used for testing, debugging, and refining the workflow to ensure the accuracy and efficiency of the simulations.

Once the workflow was fully optimized and operational, the total real-time execution required approximately 60 days to complete. This two-month computational period reflects the extensive nature of the simulations and the vast dataset generated. Upon completion of the simulations, the total size of the output files amounted to 36 GB. The output files consisted of P-δ curve data, which were systematically structured to facilitate post-processing and ML model training. Given the large volume of data generated, efficient data management strategies were employed to ensure smooth retrieval and analysis of results. It is also important to note that the developed software was able to use Reconan FEA and solve on average 1800 – 4000 models per minute depending on the size of the model.

This computational framework provided a scalable, automated, and optimized approach to handling a large dataset of RC simulations, paving the way for further advancements in predictive modeling and structural engineering research.

3 CURRENT AND FUTURE RESEARCH WORK

After the successful performance of the nonlinear pushover analyses, the output data were gathered in a single folder where they are currently being processed. This research work foresees the extraction of all P-δ curves according to the numerically generated output files through the use of Reconan FEA and LUMI. The final dataset will be stored in an Excel sheet through which the authors will be able to train ML algorithms for the development of predictive models that will be able to compute the P-δ curves of RC buildings.

Therefore, the main objective of future research work is to use the generated dataset to develop ML-generated predictive models that are going to be able to provide the Civil Engineering community with the P-δ curve of RC buildings without the need to develop FEA models, which is a tedious and time-consuming process. The predictive models will be able to provide any engineer with the P-δ curve of RC buildings within milliseconds given that the engineer will provide the required input features of their RC building.

In addition to the above, this research work aims to numerically investigate the dataset and provide insights on the correlation between the overall mechanical response of RC buildings and the different input features that relate to the material properties of the frame, the geometry, and the reinforcement configurations that Civil Engineers use during the design of their RC buildings. Investigating the impact of the input features on the variation of the P-δ curves will be performed after the dataset is assembled.

Finally, it is important to note that the extensive use of HPC resources for dataset generation necessitates an analysis of computational cost versus accuracy trade-offs. A total of 380,000 CPU hours were allocated to perform the analyses, with an execution time of approximately 60 days. However, the computational effort required for ML model inference is expected to be significantly lower than the traditional FEA approach.

To assess scalability, we will investigate:

- The minimum dataset size required for accurate ML predictions
- Potential reduction in required simulations while maintaining dataset integrity
- The feasibility of deploying pre-trained models on standard engineering workstations

By quantifying these aspects, we aim to develop a computationally efficient yet highly accurate predictive model that can be integrated into real-world structural design processes.

4 CONCLUSION

This research underscores the significant role that HPC plays in the development of large datasets necessary for the creation of predictive models in Civil Engineering. By leveraging the computational power of the LUMI supercomputer, over 21 million FEA simulations were conducted on 1,248 RC buildings, covering a vast range of geometric configurations, material properties, and reinforcement ratios. These analyses generated detailed P-δ curves that are essential for seismic vulnerability assessment. The extensive dataset formed through this process paves the way for the development of ML models that can predict the seismic response of RC buildings in a fraction of the time it would take using traditional FEA methods.

The future goals of this research include developing ML models that can instantly provide P-δ curves for any given RC building based on its specific parameters. This would significantly streamline the design and evaluation process for engineers, improving both efficiency and safety. Moreover, the ongoing investigation into the correlations between material properties, geometry, and reinforcement configurations will provide invaluable insights into the optimization of structural designs, potentially enhancing the resilience of buildings to seismic activity. As the field advances, this approach represents a transformative shift toward more efficient and effective seismic risk assessment in Civil Engineering. By reducing the computational burden associated with seismic assessments, this research paves the way for scalable, AI-driven structural analysis, enhancing both safety and efficiency in earthquake engineering.

ACKNOWLEDGEMENT

This research work was performed for the needs of the LUMI project 465000832, Development of Large FEA Generated Datasets for the Training of AI and ML Algorithms in Civil Engineering (EHPC-DEV-2023D04-047) that was awarded in 2024.

REFERENCES

[1] Eurocode 8 : Design of Structures for Earthquake Resistance—Part 1: General Rules. Seismic Actions and Rules for Buildings. Brussels Eur. Comm. Stand. 2004.

[2] Mourlas, Ch., Khabele, N., Bark, H.A., Karamitros, D., Taddei, F., Markou, G. and Papadrakakis, M. (2020), The Effect of Soil-Structure Interaction on the Nonlinear Dynamic Response of Reinforced Concrete Structures, International Journal of Structural Stability and Dynamics, **20** (13), 2041013, 2020, https://doi.org/10.1142/S0219455420410138.

[3] Markou, G., Sabouni, R., Suleiman, F. and El-Chouli, R. (2015), Full-Scale Modeling of the Soil-Structure Interaction Problem Through the use of Hybrid Models (HYMOD), International Journal of Current Engineering and Technology, **5** (2), pp. 885-892.

[4] Zavala DJ, Torres CJ, Moreno JD. Influence of the P-delta Effect and Stiffness Irregularity on the Structural Behavior of Reinforced Concrete Buildings. J Phys Conf Ser. 2022;2287:0–9.

[5] Markou, G. and Genco, F. (2019), Seismic Assessment of Small Modular Reactors: NuScale Case Study for the 8.8 Mw Earthquake in Chile, Nuclear Engineering and Design, **342**(2019), 176-204, https://doi.org/10.1016/j.nucengdes.2018.12.002.

[6] Pettinga D, Priestley N. Accounting for P-Delta Effects in Structures When Using direct displacement-based design. 14 World Conf Earthq Eng. 2008. p. 1–8.

[7] Yahiaoui A, Dorbani S, Yahiaoui L. Machine learning techniques to predict the fundamental period of infilled reinforced concrete frame buildings. Structures [Internet]. 2023;54:918–27. Available from: https://doi.org/10.1016/j.istruc.2023.05.052

[8] Yahiaoui A, Matos J., Dorbani S. Probabilistic forecast of concrete compressive strength using ML. 20th Int Probabilistic Work IPW 2024. Cham : Spr. Cham : Springer Nature Switzerland; 2024. p. 281–6.

[9] AlHamaydeh M, Markou G, Bakas N, Papadrakakis M. AI-based shear capacity of FRP-reinforced concrete deep beams without stirrups. Eng Struct. 2022;264:114441.

[10] Spijkerman Z, Bakas N, Markou G, Papadrakakis M. Predicting the shear capacity of reinforced concrete slender beams without stirrups by applying artificial intelligence algorithms. Proc Compdyn. 2021.

[11] Westhuizen AM van der, Bakas N, Markou G. Developing an artificial neural network model that predicts the fundamental period of steel structures using a large dataset. Proc Compdyn. 2023.

[12] Markou G, Bakas NP, Chatzichristofis SA, Papadrakakis M. A General Framework of High-Performance Machine Learning Algorithms: Application in Structural Mechanics. Comput Mech Press. 2023;73:705–29.

[13] Gravett DZ, Mourlas C, Taljaard VL, Bakas N, Markou G, Papadrakakis M. New fundamental period formulae for soil-reinforced concrete structures interaction using machine learning algorithms and ANNs. Soil Dyn Earthq Eng. 2021;144:106656.

[14] Asteris PG. The FP4026 Research Database on the fundamental period of RC infilled frame structures. Data Br [Internet]. 2016;9:704–9. Available from: http://dx.doi.org/10.1016/j.dib.2016.10.002

[15] Charalampakis AE, Tsiatas GC, Kotsiantis SB. Machine learning and nonlinear models for the estimation of fundamental period of vibration of masonry infilled RC frame structures. Eng Struct [Internet]. 2020;216. Available from: https://doi.org/10.1016/j.engstruct.2020.110765

[16] Mirrashid M, Naderpour H. Computational intelligence-based models for estimating the fundamental period of infilled reinforced concrete frames. J Build Eng [Internet]. 2022;46:103456. Available from: https://doi.org/10.1016/j.jobe.2021.103456

[17] ReConAn v2 . 00 Finite Element Analysis Software User's Manual [Internet]. 2020. Available from: https://www.researchgate.net/publication/342361609_ReConAn_v200_Finite_Element_Analysis_Software_User's_Manual

[18] Eurocode 2: design of concrete structures-part 1–1: General rules and rules for buildings. London: British Standard Institution; 2005.

[19] Markou, G. and Papadrakakis, M., Accurate and Computationally Efficient 3D Finite Element Modeling of RC Structures, Computers & Concrete, **12** (4), 443-498, 2013. https://doi.org/10.12989/cac.2013.12.4.443