Determination of sample size on surrogate model-based parameter inverse analysis of a super-high arch dam

Xi Liu[1[0000-0001-7303-367X]], Maria Pina Limongelli[2[0000-0002-9353-5439]], Fei Kang[1[0000-1111-2222-3333]]

[1] Dalian University of Technology, Dalian 116024, P. R. China
2 Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy
mariagiuseppina.limongelli@polimi.it

Abstract. This paper investigates the impact of the sample size on a surrogate model in the context of parameter inverse analysis for high arch dams. A deep learning-based surrogate model is developed and integrated with Jaya optimization algorithm to enhance the computational efficiency and accuracy of the inverse analysis. The input variables for the training set of the surrogate model are generated by Latin Hypercube Sampling (LHS). The output variables are obtained based on a high-precision finite element model calculation. By comparing the model accuracy and computation time across different sample sizes (ranges from 20 to 200 times the number of input variables), the optimal sample size is identified. The study was conducted for the case study of an actual high arch dam in China for wich measured data are available. The results indicate that a sample size of 100 times the number of input variables achieves a favorable balance between accuracy and computation time.

Keywords: Sample size, Surrogate Model, Inverse Analysis, Arch Dam, Parameter Identification.

1 Introduction

Materials parameters, such as the elastic modulus of concrete dams, hold significant importance in modelling their structural performance and evaluating their structural integrity through model updating procedures. The identification of the elastic modulus of a dam commonly relies on the utilization of measured displacements and inverse analysis techniques, which is essentially an optimization problem. For super high arch dams with multi-material zoning, the inversion process is inherently complex, involving a significant number of parameters, a large sample space, and multiple local extremes. To enhance the accuracy and efficiency of the inverse analysis, surrogate models are often employed.

The Kriging model [1], the response surface methodology[2], and neural network-based models are commonly employed techniques for surrogate modeling. In a recent study, the authors proposed a deep learning-based surrogate model for parameter inverse analysis [3]. In comparison to the commonly used shallow neural network-

based surrogate model, the proposed deep network exhibits superior high-dimensional data mapping capability and higher accuracy. In that first study, the sample size for establishing the surrogate model was set to 10~15 times the dimension of the input parameters based on the literature [4]. It is important to note that there is no universally applicable sample size for all practical applications. Moreover, the number of samples and their distribution in the parameter space directly affects the simulation efficiency and accuracy of the surrogate model. This paper serves as a continuation of the previous work, to further investigate the appropriate setting for the sample size.

In general, increasing the number of sample points tends to enhance the model fitting performance. However, this improvement comes at the cost of increased computational complexity and time consumption. In this paper, we aim to investigate the relationship between sample size and model prediction accuracy through the analysis of an engineering example. Based on the findings, we will provide a recommended sample size value, which can serve as a foundation for addressing the parameter inverse analysis problem of concrete arch dams with similar characteristics.

2 Methodology

2.1 Inverse analysis theory based on measured displacement

Assuming that the material of the dam and foundation is isotropic and homogeneous, the inverse analysis approach relies on the minimization of the distance between measured displacements and displacements calculated by a finite element model. The distance is quantified by an objective function which is minimized while considering prescribed constraints.

$$\begin{cases} \min \;\; J(\boldsymbol{E}) = \omega_i \sum_{i=1}^{N} \left(\left(u_{h,i}^{*}(\boldsymbol{E}) - u_{h,i} \right) / u_{h,i} \right)^2 \\ \text{s.t.} \;\; \mathbf{K} \cdot \boldsymbol{u} = \boldsymbol{R} \\ E_{j,\min} \le E_j \le E_{j,\max} \;\; (j = 1,\ 2,\ \cdots,\ N_E) \end{cases} \tag{1}$$

where J (.) is the objective function to be optimized, $\boldsymbol{E}$ is the vector of elastic/deformation moduli collecting the terms E_j each corresponding to the elastic modulus of a portion of the dam-foundation system. N_E is the number of portions of the dam: within each of them the elastic modulus is considered constant. The values $E_{j,\min}$ and $E_{j,\max}$ are the minimum and maximum value of E_j. ω_i is the weight assigned to the i-th displacement, $u_{h,i}^{*}$ is the hydrostatic component of the displacement, separated from the measured displacement, $u_{h,i}$ is the displacement calculated by FEM, and N is the number of locations where displacement is measured. $\mathbf{K}$ is the global stiffness matrix, which depends on the vector of elastic moduli $\boldsymbol{E}$, $\boldsymbol{u}$ is the nodal displacement vector collecting the displacement at the N locations, $\boldsymbol{R}$ is the equivalent load vector. This study employs a meta-heuristic intelligent optimization algorithm called Jaya [5] to solve the optimization problem. This algorithm has been selcted

Commentato [MGL1]: Where do you use this assumption?

Commentato [MGL2]: Here you have to shortly explain why only the hydrostatic components goes into the objective function

Commentato [MGL3]: W should go inside the summation since it depends on i

Commentato [MGL4]: Define s.t.

Commentato [MGL5]: I guess this is a vector?

Commentato [MGL6]: Are these physical constraints?

Commentato [MGL7]: What do you mean by 'separated'?

since the capability of Jaya in effectively exploring the search space and attaining the optimal solution for the given optimization problem has been empirically verified in various engineering optimization problem.

Commentato [MGL8]: add some references about this

2.2 CNN-based surrogate model

The utilization of a surrogate model, as an alternative to finite element calculations, is a commonly employed technique for improving the computational efficiency of inverse analysis. In this study, an advanced deep learning network is introduced to construct the surrogate model. The adopted architecture is built upon convolutional neural networks (CNN), and the network structure and hyperparameter settings are presented in Table 1. Performance criteria employed to evaluate the proposed CNN surrogate model include mean absolute percentage error (MAPE), mean absolute error (MAE), mean square root error (RMSE), which can be expressed as:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|(y_i - \hat{y}_i)/y_i\right| * 100 \tag{2}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, n is the number of output.

Table 1. Network structure and hyperparameter settings of CNN.

Parameter	Set
Convolutional layers	3 layers, filter size 3*64, 3*128, 3*256
Activation function	Rectified Linear Units (ReLU)
Fully connected layers	2 layers, size 256 and 10
Optimizer	ADAM
Maximum number of epochs	100
Size of the mini batch	32
Sequence length	50
Initial learn rate	1e-3
Learn rate drop period and factor	‘Piecewise’, 0.1

2.3 Sample datasets based on LHS and FEM

The sample dataset used for the surrogate model comprises two main variables: the elastic/deformation modulus and the dam displacement. The elastic/deformation modulus serves as the input variable, while the dam displacement is considered the output variable. To generate the initial points of the input variables within the parameter domain, Latin Hypercube Sampling (LHS) is employed. Subsequently, the output

variables are obtained from finite element calculations, considering a specific load and input variables for the limiting case.

The determination of the sample size plays a crucial role in achieving accurate predictions and modeling efficiency for the surrogate model. A small sample size may not yield reliable statistical results, while increasing the sample size can improve computational accuracy. However, it also introduces a higher computational burden, particularly for complex structures and systems, resulting in a significant increase in simulation computation time. To optimize the utilization of computational resources while ensuring satisfactory accuracy, the approach employed in this paper involves initially assessing the performance of the surrogate model using a small sample size. Subsequently, the sample size is incrementally increased until the desired accuracy requirements are achieved.

3 Results

3.1 Description of the concrete arch dam

The dam is a double-curved arch dam constructed in April 2007. The top elevation of the dam reaches 610 meters, while the bottom elevation is at 324.5 meters. The maximum height of the dam reaches 285.5 meters. The normal storage level is 600 meters. Notably, the main body of the dam is divided into 31 sections, and it is further classified into three distinct areas, namely A, B, and C, based on the varying design values assigned to the concrete material strength of the dam. The finite element model and material zoning of the dam-foundation system are shown in Fig. 1.

The data collection for horizontal radial displacement involved the selection of 10 monitoring locations within dam sections 10, 15, and 22. These monitoring locations were noted as PL10-2, PL10-3, PL10-5, PL15-1, PL15-2, PL15-4, PL15-5, PL22-1, PL22-2, and PL22-4.

Commentato [MGL9]: What does this mean?

Commentato [MGL10]: These are not visible in the figure. Why do we need to mention them?

Commentato [MGL11]: Same as before. Either you inclde another figure with the notations or remove the notations in the text

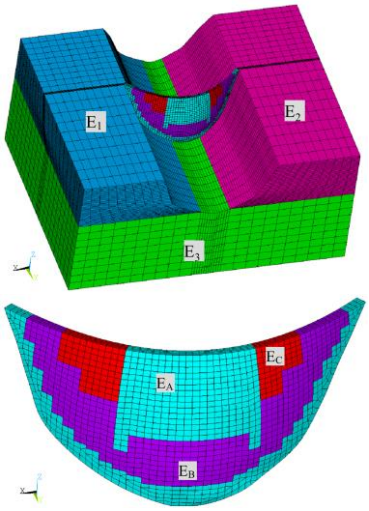


Fig. 1. FEM and material zoning of the dam-foundation system.

3.2 Determination of sample size

Table 2 displays the considered samples of different size. The training set consists of 20 to 200 times the number of input variables, whereas the test set is set at 25% of the training set size. The input variables were generated using LHS technique. Subsequently, these variables are input into the FEM to calculate the dam displacements, thereby obtaining the output dataset for the surrogate model. The computation time of the finite element analysis for varying sample sizes is illustrated in Fig. 2. It is evident that the finite element computation time exhibits a primarily linear increase with the growth of the sample size.

Table 2. Sample size setting.

Sample size	Training set	Test set	Total
20D	120	30	150
40D	240	60	300
60D	360	90	450
80D	480	120	600
100D	600	150	750
150D	900	225	1125
200D	1200	300	1500

Note: D is the number of input variables, in this case study D=6.

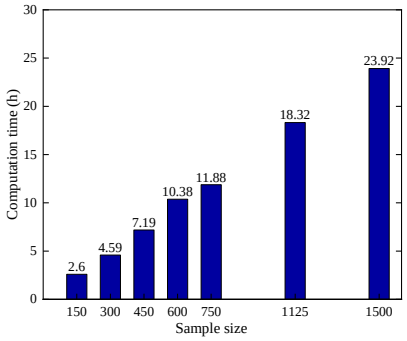


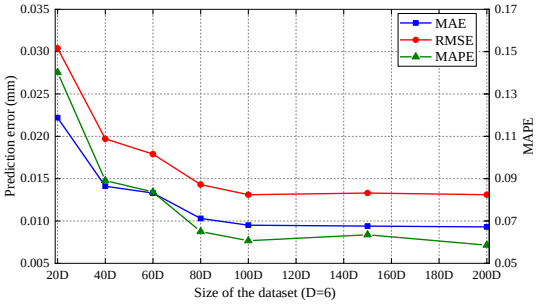
Fig. 2. Computation time for different sample size.

The prediction error of the surrogate models with different sample sizes is shown in Table 3 and Fig. 3. The observed trend reveals that the accuracy of the model improves as the sample size increases. Notably, as the sample size reaches 100D (D represents the number of variables), the growth rate of accuracy tends to approach zero. This suggests that the prediction performance of the model tends to stabilize at this point. Therefore, a sample size of 100D is selected for surrogate modeling.

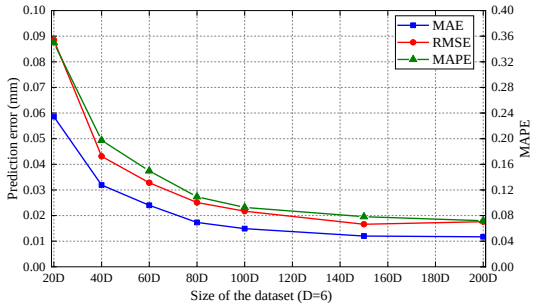
Table 3. Prediction error of the surrogate models for different sample sizes.

Dataset	Training			Test		
size	MAPE	MAE/mm	RMSE/mm	MAPE	MAE/mm	RMSE/mm
20D	0.1402	0.0222	0.0304	0.3497	0.0586	0.0885
40D	0.0891	0.0141	0.0197	0.1974	0.0319	0.0431
60D	0.0837	0.0133	0.0179	0.1497	0.024	0.0328
80D	0.0650	0.0103	0.0143	0.1094	0.0173	0.0251
100D	0.0607	0.0095	0.0131	0.0927	0.0149	0.0217
150D	0.0635	0.0099	0.0138	0.0782	0.0125	0.0173
200D	0.0586	0.0093	0.0131	0.0720	0.0117	0.0176

Note: MAPE is the mean absolute percentage error; MAE is the mean absolute error; RMSE is the root mean squared error.



(a) Training set

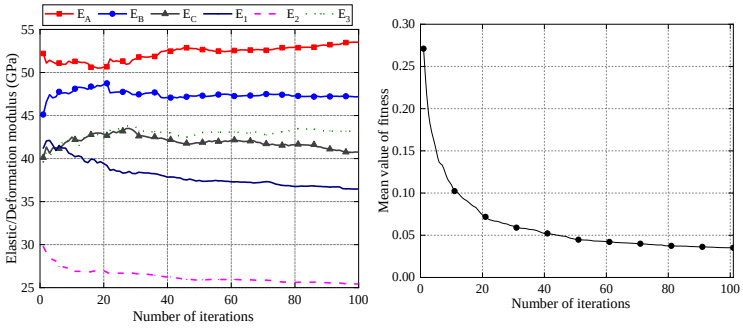


(b) Test set

Fig. 3. Prediction error of the surrogate models for different sample sizes.

3.3 Results of parametric inverse analysis

The results of parametric inverse analysis are E_A=53.541GPa, E_B=47.184GPa, E_C=40.699GPa, E_1=36.453GPa, E_2=25.414GPa, and E_3=43.223GPa. Figure 4 shows the iterative process of parameters and mean fitness. The minimal variations in the parameters coupled with the small fitness values indicate that the inverse results are stable and accurate. Subsequently, the outcomes are utilized as input for the forward calculation within the finite element model, yielding the displacement values. The comparison between the calculated and the measured displacement is presented in Table 4. The results indicate that the maximum relative error is below 1%, thus providing substantial evidence for the high accuracy of the inverse analysis results obtained through the proposed approach.



(a) (b)

Fig. 4. Process of parametric inverse analysis. (a) Parametric evolution process; (b) Fitness evaluation process.

Table 4. Performance of the Jaya-CNN for forward analysis.

Monitoring locations	Measured (mm)	Calculated (mm)	Absolute error (mm)	Relative error (%)
PL10-2	21.6218	21.5999	0.0219	0.1014
PL10-3	18.4370	18.3823	0.0547	0.2965
PL10-5	3.3131	3.2985	0.0146	0.4409
PL15-1	27.3852	27.2674	0.1178	0.4300
PL15-2	24.4985	24.3801	0.1184	0.4833
PL15-4	15.2085	15.0832	0.1253	0.8242
PL15-5	7.1204	7.1100	0.0104	0.1458
PL22-1	23.2238	23.0915	0.1323	0.5698
PL22-2	19.9128	19.8625	0.0503	0.2525
PL22-4	10.1150	10.0660	0.0490	0.4840

4 Conclusion

This paper explores the influence of the sample size on a surrogate model for parametric inverse analysis, utilizing a case study of a high arch dam example project. The sample size is varied from 20 to 200 times the number of input variables. The results reveal that a favorable trade-off between accuracy and computation time is achieved when the sample size is 100 times the number of input variables. Based on this sample size, a deep learning-based surrogate model is proposed for the parametric inverse analysis. The efficacy of this proposed approach is validated using measured displacement data, thereby establishing a reliable framework for addressing similar inverse problems in super high arch dams.

References

1. Kleijnen, J.P.C.: Kriging metamodeling in simulation: A review. European Journal of Operational Research. 192, 707–716 (2009).
2. Simpson, T.W., Poplinski, J.D., Koch, P.N., Allen, J.K.: Metamodels for Computer-based Engineering Design: Survey and recommendations. EWC. 17, 129–150 (2001).
3. Liu, X., Kang, F., Limongelli, M.P.: Multi-zone parametric inverse analysis of super high arch dams using deep learning networks based on measured displacements. Advanced Engineering Informatics. 56, 102002 (2023).
4. Kang, F., Liu, X., Li, J., Li, H.: Multi-parameter inverse analysis of concrete dams using kernel extreme learning machines-based response surface model. ENGINEERING STRUCTURES. 256, (2022).
5. Venkata Rao, R.: Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems. INT J IND ENG COMP. 19–34 (2016).