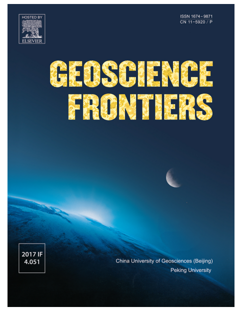
Journal Pre-proof



Big Data and Machine Learning in Geoscience and Geoengineering: Introduction

Wengang Zhang, Jianye Ching, Anthony Teck Chee Goh, Andy Y.F. Leung

PII: S1674-9871(20)30124-9

DOI: https://doi.org/10.1016/j.gsf.2020.05.006

Reference: GSF 1009

To appear in: *Geoscience Frontiers*

Editorial

Big Data and Machine Learning in Geoscience and Geoengineering: Introduction

Wengang Zhang[a,*], Jianye Ching[b], Anthony Teck Chee Goh[c], Andy Y.F. Leung[d]

[a] School of Civil Engineering, Chongqing University, Chongqing 400045, China.

[b] Department of Civil Engineering, National Taiwan University, Chinese Taipei

[c] School of Civil and Environmental Engineering, Nanyang Technological University, Singapore

[d] Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong, China

*Corresponding author: Wengang Zhang (zhangwg@cqu.edu.cn)

In recent years, we have entered the so-called Fourth Paradigm with the regular production of huge amount of observational data. Big data is often characterized by the three 'V's: Volume of data, Variety and Velocity. The concept of big data can potentially address some existing issues in areas of geoscience and geoengineering. Large-scale, comprehensive, multidirectional and multifield geotechnical monitoring is becoming a reality in the very near future. The in-depth analysis capabilities such as correction analysis, casual analysis, and decision support can become the core works for monitoring projects in the era of big data. Furthermore, site monitoring may be promoted to similar or even more important roles than the experimentation, theoretical analysis and numerical simulation. On the other hand, the nature of scientific geoscience and geoengineering data, and the processes used to retrieve and analyze them, may differ substantially from those in other fields. It is therefore timely that geoscience and geoengineering professionals pay more attention to big data research, create the environment to utilize data to add value to our fields and promote collaboration with data analysts of other disciplines.

Machine learning (ML) is the scientific study of algorithms and statistical models that allows computers to learn from existing data, in order to improve their performance on specific tasks without being explicitly programmed. In recent years, the application of ML in a wide range of industries has grown rapidly. ML can be very useful in solving problems where deterministic solutions are not available or are excessively expensive in terms of computational cost but for which there are ample observations and data available. Due to the nature of materials, geoscience and geo-engineering face more significant uncertainties than other fields of civil and mechanical engineering. Meanwhile, there is a lot of monitoring and site investigation data in geotechnical engineering which can be taken advantage of by using data analytic methods. Therefore, ML can be a suitable and effective alternative to solve geotechnical engineering problems. The combination of big data and machine learning may create unexpected solutions to the conventional geotechnical problems.

Accordingly, in this special issue of *Geoscience Frontiers*, we assemble 14 invited papers which provide insight into the latest developments and challenges in applying big data and machine learning to geoscience and geoengineering.

Predicting the performance of a tunneling boring machine is of vital importance to avoid any possible accidents during tunneling boring. Li et al. (2020) develop a long short-term memory neural network model to predict the TBM performances including the total thrust and the cutterhead torque in a real-time manner, based on the big data obtained from the 72.1 km long tunnel in the Yin-Song diversion project in China. The study also indicated that the missing of a key parameter can significantly reduce the accuracy of the model, while the supplement of a parameter that highly-correlated with the missing one can improve the prediction.

Successful application of classic geostatistical models requires prior characterization of spatial auto-correlation structures, which poses a great challenge for unexperienced engineers, particularly when only limited measurements are available. Shi and Wang (2020) propose an ensemble radial basis function network (RBFN) method not only to allows geotechnical anisotropy to be properly incorporated, but also quantifies uncertainty in spatial interpolation. The proposed method is illustrated using numerical examples of cone penetration test (CPT) data. Furthermore, a comparative study is performed to benchmark the proposed RBFN with two other non-parametric data-driven approaches. The results reveal that the proposed ensemble RBFN provides better estimation of spatial patterns and associated prediction uncertainty at unsampled locations when a reasonable amount of data is available as input.

Making use of a large volume of landslide data compiled in Hong Kong over the past few decades, Wang et al. (2020) introduce a novel machine-learning and deep-learning method to identify natural terrain landslides. Different types of landslide-related data were compiled, including topographic data, geological data and rainfall-related data. Three integrated geodatabases were also established, represented by Recent Landslide Database, Relict Landslide Database and Joint Landslide Database (JLD). Promising results were achieved by the machine learning and deep learning methods, in particular the convolutional neural networks (CNN) method, owing to its strengths in feature extraction and multi-layer two-dimensional data processing, which are important for landslide identification problems.

Zhang et al. (2020c) formulate surrogate models for prediction of braced excavation response, via ensemble learning methods including the extreme Gradient Boost and Random Forest Regression methods. The surrogate model takes into account parameters such as excavation width, wall stiffness, wall penetration, and soil properties represented by ratio of passive to active strength and stiffness to strength ratios, etc. The surrogate model is trained by finite element models adopting the elastoplastic constitutive model NGI-ADP, leading to an efficient alternative tool for prediction of horizontal wall displacements. The proposed method also allows sensitivity analyses to be performed efficiently, where the soil properties show more significant influences on the excavation response than wall width, wall penetration depth and excavation width.

In the paper by Ray et al. (2020), the reliability analysis of shallow foundations on a clayey soil is assessed based on settlement criteria using three soft computing techniques: Minimax Probability Machine Regression (MPMR), Particle Swarm Optimisation-Artificial Neural Network (PSO-ANN) and Particle Swarm Optimisation-Adaptive Neuro-Fuzzy Inference System (PSO-ANFIS). The MPMR, PSO-ANN and PSO-ANFIS models were compared on the basis of various fitness parameters. All the three models were found to give good predictions of shallow foundation settlement on a clayey soil.

Precipitation-induced shallow landslides are common geological hazards in mountainous regions. In the paper by Liu et al. (2020), three machine learning algorithms, the Random Forest (RF), Boasted Regression Tree (BRT) and MultiLayer Perceptron neural network (MLP) are applied to model shallow landslides in Norway. A total of 8 landslide conditioning factors and 3 time-dependent triggering factors such as the slope angle, profile curvature and flow direction were selected as inputs. The results indicated that all three algorithms were capable of predicting the spatial distribution of landslides over a large area.

An important aspect in the geological risk assessment of tunnel face stability is the interpretation of the rock structure. The paper by Chen et al. (2020) present a framework for automated classification of rock mass structure based on the geological images of the tunnel face using CNN. Experimental results revealed that the proposed method is optimal and efficient for automated classification of rock structure using the geological images of the tunnel face.

The paper by Pan et al. (2020) propose a probabilistic analysis approach to assess seismic slope performance using a new metamodel that makes use of relevance vector machine and polynomial chaos expansion methodologies. In this work, a novel method was proposed to incorporate uncertainties associated with earthquake ground motions and soil shear strength.

Based on a large database of 4315 observations for 479 different anchors from 7 different projects, Shen et al. (2020) present a novel hybrid data-driven machine learning FastICA-MARS (Independent Component Analysis-multivariate adaptive regression splines) model for prediction of the load-displacement relationship of grouted anchors in weathered granite.

With sparse multivariate data obtained from geotechnical site investigation, it is impossible to identify outliers with certainty due to the distortion of statistics of geotechnical parameters caused by outliers and their associated statistical uncertainty resulted from data sparsity. Zheng et al. (2020) propose an approach quantifying the outlying probability of each data instance based on Mahalanobis distance and determining outliers as those data instances with outlying probabilities greater than 0.5. The proposed method tackled the distortion issue of statistics estimated from the dataset with outliers by a re-sampling technique and accounts, rationally, for the statistical uncertainty by Bayesian machine learning.

Compression index (C_c) is an essential parameter in geotechnical serviceability design. In the paper by Zhang et al. (2020a), a database with 311 clay cases with knowledge of C_c and three index properties was collected to train five commonly used machine learning (ML) models, including back-propagation neural network (BPNN), extreme learning machine (ELM), support vector machine (SVM), random forest (RF), and evolutionary polynomial regression

(EPR). The results indicated that ML models outperform empirical prediction formulations and that RF performs the best in terms of prediction accuracy for C_c, followed by BPNN, ELM, EPR and SVM.

Knowledge of pore-water pressure (PWP) variation is fundamental for slope stability. To explore the applicability and advantages of recurrent neural networks (RNNs) on PWP prediction, Wei et al. (2020) proposed three types of RNNs, i.e., standard RNN, long short-term memory (LSTM) and gated recurrent unit (GRU) to compare the predictive performances with a traditional static artificial neural network. The results indicated that the GRU and LSTM models produced the most precise and robust prediction among the four models.

Zhang et al. (2020b) address the estimation of the undrained shear strength (USS) for soft sensitive clays. Based on the soil data sets from TC304 database, this study constructed multivariate regression models to predict USS based on preconsolidation stress (PS), vertical effective stress (VES), liquid limit (LL), plastic limit (PL), and natural water content (W). Two relatively recent machine learning (ML) methods, XGBoost & Random Forest (RF), were investigated. These two methods were special in the way that they did not make predictions based on a single model but based on multiple models to combine their prediction powers. It was found that these two methods outperform some other ML approaches.

Compared with RF, XGBoost further provided feature importance ranks, which can enhance the interpretability of model.

Traditional approaches to develop 3D geological models employ a mix of quantitative and qualitative scientific techniques, which do not fully provide quantification of uncertainty in the constructed models and fail to optimally weight geological field observations against constraints from geophysical data. Olierook et al. (2020) develop a methodology to fuse lithostratigraphic field observations with aeromagnetic and gravity data to build a 3D model in a small region of the Gascoyne Province, Western Australia. The results revealed that surface geological observations fused with geophysical survey data can yield reasonable 3D geological models with narrow uncertainty regions at the surface and shallow subsurface.

We are privileged to be invited by Prof. Xuanxue Mo, Editor-in-Chief, and Prof. M. Santosh, Editorial Advisor of Geoscience Frontiers to edit this special issue. We are grateful to the authors for their generous contributions and patience during the review process. Our heartfelt thanks also go to the dedicated reviewers for their useful comments. We would also like to acknowledge the immense support from Dr. Lily Wang, Editorial Assistant of Geoscience Frontiers for her dedicated assistance during the review and processing of the manuscripts for this issue.

References

Li J, Li P, Guo D, Li X, Chen ZY. 2020. Advanced prediction of tunnel boring machine performance based on big data. Geoscience Frontiers. https://doi.org/10.1016/j.gsf.2020.02.011

Shi C, Wang Y. 2020. Non-parametric machine learning methods for interpolation of spatially varying non-stationary and non-Gaussian geotechnical properties. Geoscience Frontiers. https://doi.org/10.1016/j.gsf.2020.01.011

Ray R, Kumar D, Samui P, Roy LB, Goh ATC, Zhang W. 2020. Application of soft computing techniques for shallow foundation reliability in geotechnical engineering. Geoscience Frontiers.

Liu Z, Gilbert G, Cepeda JM, Lysdahl AOK, Piciullo L, Hefre H, Lacasse S. 2020. Modelling of shallow landslides with machine learning algorithms. Geoscience Frontiers. https://doi.org/10.1016/j.gsf.2020.04.014

Chen J, Yang T, Zhang D, Huang H, Tian Y. 2020. Deep learning based classification of rock structure of tunnel face. Geoscience Frontiers.

https://doi.org/10.1016/j.gsf.2020.04.003

Pan Q, Leung YF, Hsu S. 2020. Stochastic seismic slope stability assessment using polynomial chaos expansions combined with relevance vector machine. Geoscience Frontiers. https://doi.org/10.1016/j.gsf.2020.03.016

Zhang P, Yin ZY, Jin YF, Chan THT, Gao FP. (2020a). Intelligent modelling of clay compressibility using hybrid meta-heuristic and machine learning algorithms. Geoscience Frontiers. https://doi.org/10.1016/j.gsf.2020.02.014

Zhang WG, Wu CZ, Zhong HY, Li YQ, Wang L. (2020b). Prediction of undrained shear strength using Extreme Gradient Boosting and Random Forest based on Bayesian optimization. Geoscience Frontiers. https://doi.org/10.1016/j.gsf.2020.03.007

Zhang RH, Wu CZ, Goh ATC, Thomas Böhlke, Zhang WG. (2020c). Estimation of Diaphragm Wall Deflections for Deep Braced Excavation in Anisotropic Clays Using Ensemble Learning. Geoscience Frontiers. DOI: 10.1016/j.gsf.2020.03.003

Hugo K. H. Olierook, Richard Scalzo, David Kohn, Rohitash Chandra, Ehsan Farahbakhsh, Chris Clark, Steven M. Reddy, R. Dietmar Müller. Bayesian geological and geophysical data fusion for the construction and uncertainty quantification of 3D geological models. 2020. Geoscience Frontiers.

Xin Wei, Lulu Zhang, Hao-Qing Yang, Limin Zhang, Yang-Ping Yao. 2020. Machine learning for pore-water pressure time-series prediction: Application of recurrent neural networks. Geoscience Frontiers. https://doi.org/10.1016/j.gsf.2020.04.011

Shuo Zheng, Yu-Xin Zhu, Dian-Qing Li, Zi-Jun Cao, Qin-Xuan Deng, Kok-Kwang Phoon. 2020. Probabilistic outlier detection for sparse multivariate geotechnical site investigation data using Bayesian learning. Geoscience Frontiers. https://doi.org/10.1016/j.gsf.2020.03.017

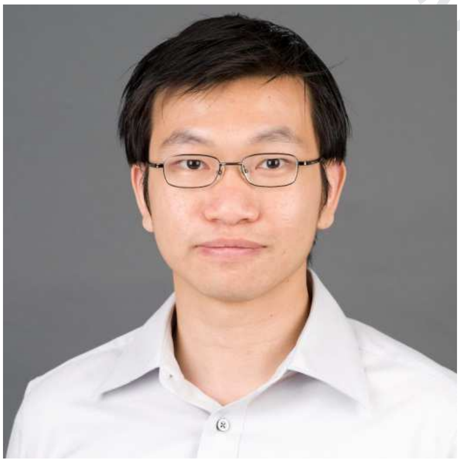
Haojie Wang, Limin Zhang, Te Xiao, Lulu Zhang, Jinhui Li. 2020. Landslide Identification Using Machine Learning. Geoscience Frontiers. https://doi.org/10.1016/j.gsf.2020.02.012

Hao Shen, Jinhui Li, Sixin Wang, Zewei Xie. 2020. Prediction of load-displacement performance of grouted anchors in weathered granites using FastICA-MARS as a novel model. Geoscience Frontiers.

Bio+Photo

 **Dr. Wengang ZHANG**	Dr. Wengang ZHANG is currently full professor in School of Civil Engineering, Chongqing University, China. He obtained his BEng and MEng degrees in Hohai University, China, as well as PhD in Nanyang Technological University, Singapore. His research interests focus on assessment of influences on the built environment induced by underground construction, in circumstances of complicated geological and geophysical conditions, as well as the big data and machine learning in geotechnics and geoengineering. He is now the members of the ISSMGE TC304 (Reliability), TC309 (Machine Learning), and TC219 (System Performance of Geotechnical Structures). His paper "Multivariate adaptive regression splines and neural network models for prediction of pile drivability " is highly cited paper and "Multivariate adaptive regression splines for analysis of geotechnical engineering systems" won Sloan outstanding paper

	award in 2019.
 Prof Jianye Ching	Dr. Ching is Distinguished Professor at the Dept. of Civil Engineering at National Taiwan University, China. His main research interests are geotechnical reliability analysis & reliability-based design, basic uncertainties in soil properties, random fields & spatial variability, reliability-based geotechnical design codes, and probabilistic site characterization. He is the author or co-author of more than 100 publications in international journals. Dr. Ching is now Editor-in-Chief of Journal of GeoEngineering, Managing Editor of Georisk, Associate Editor of Canadian Geotechnical Journal & Editorial Board Member of Structural Safety. He has received the following awards and recognitions: Georisk Best Paper Award (2014), Editor's Choice Paper from Canadian Geotechnical Journal (2014, 2017), Highly Cited Research from Structural Safety (2016), Outstanding Reviewer from Canadian Geotechnical Journal (2015, 2016), and Outstanding Reviewer from ASCE-ASME Journal of Rick and Uncertainty in Engineering System (2016).

 Prof. ATC Goh	Prof. Goh ATC is Associate professor in the School of Civil and Environmental Engineering at Nanyang Technological University, Singapore. He received his PhD and BEng in Monash University, Australia. He is a registered Professional Engineer in Singapore and Australia. His teaching, research and professional practice have covered many aspects of geotechnical engineering including soft computing, finite element analysis, earth retaining structures, pile foundations and slope stability.
 Dr. Andy Y.F. Leung	Dr. Leung is currently Associate Professor at The Hong Kong Polytechnic University (PolyU). He graduated from The University of Hong Kong (BEng) and University of California, Berkeley (MS), before he obtained PhD degree at the University of Cambridge, UK. Before joining PolyU, Dr Leung has practiced geotechnical engineering both in Hong Kong and the United States, and had been involved in a number of large-scale civil engineering projects in Hong Kong, the United States, United Kingdom, India and the Middle East. He is a registered professional engineer

	in the State of California, US. His research interests include soil-structure interaction, reliability of geotechnical and structural systems, probabilistic analysis and machine learning approaches and applications. He has received awards including the HKIE Fugro Prize, Departmental Teaching Excellence Award, Dean's Award for Outstanding Achievement in Research Funding, etc. He is a member of several technical committees of the International Society for Soil Mechanics and Geotechnical Engineering (ISSMGE), including TC304 Engineering Practice of Risk Assessment and Management, and TC309 Machine Learning and Big Data. Currently, he also serves as the Secretary-General of Hong Kong Geotechnical Society.