CSI (Cement Science Investigation): using machine learning to guess the OPC pastes composition from the elastic response

Tulio Honorio, Sofiane Ait Hamadouche, Amélie Fau

CSI (Cement Science Investigation): using machine learning to guess the OPC pastes composition from the elastic response

Tulio Honorio[1][0000-0003-0281-4502], Sofiane Ait-Hamadouche[1][.] and Amélie Fau[1][0000-0002-0902-969X]

[1] Université Paris-Saclay, ENS Paris-Saclay, CNRS, LMT - Laboratoire de Mécanique et Technologie, 91190, Gif-sur-Yvette, France.
tulio.honorio-faria-de-faria@ens-paris-saclay.fr

Abstract. Precise prediction of the elastic response is crucial to model cracking at early and late ages of cement-based materials and structures. Here, we use Machine Learning (ML) techniques to predict the elastic properties of Ordinary Portland Cement (OPC) pastes. A database with 365 observations is built on experimental studies from in the literature. We show that micromechanics-based estimations may provide missing data in databases to be interrogated by ML, increasing the accuracy of predictions. Finally, we explore the formulation space of OPC pastes using Monte Carlo computations, which enables guessing which are the compositions that can be associated with a given elastic response. Applications of our results include the development of tailored formulations for a target elastic response and also in the forensics of existing cement pastes.
Keywords: Elastic constants; Machine Learning; Data Science; Composition-Property Correlations.

1 Introduction

Modeling cracking in cement-based materials and structures requires a precise description prediction of the elastic response of the material. One of the main approaches to assess cracking risk in concrete is based on elastic analysis of the (e.g. crack risk index in JCI recommendation [1]), which of course require an accurate prediction of the elastic properties, especially at early-age when properties significantly change due to the progress of hydration process. The correct prediction of the evolution of mechanical properties is also a fundamental input in more sophisticated non-linear analysis of cracking of cement-based materials.

Establishing composition-structure-property correlations is the central paradigm in the optimization of the formulation and understanding the behavior of cement-based materials. Most of the concrete research community has focused on an increasingly costly and laborious trial-and-error experimental exploration of the formulation space, while much less significant attention has been devoted to establishing composition-structure-property correlations via fundamental component data and validated model-

ing approaches. The latter is a strategy that has been recognized with the potential to "revolutionize" cement and concrete research [1].

In this work, we first use Machine Learning (ML) techniques to predict the elastic properties of OPC pastes. Understanding cracking process in cement pastes is a first milestone to understand cracking at more complex system such as mortar and concrete in which the heterogeneities due to interfacial transition zone (ITZ) may play a role. Here, we show that micromechanics-based estimations can be useful to provide missing data in databases to be interrogated by ML. Finally, we explore the formulation space of OPC pastes using Monte Carlo computations. This strategy allows guessing what are the compositions that can be associated with a given elastic response, which can be useful for application requiring a target elastic response and also in the forensics of existing cement pastes.

2 Data science and machine learning methods

2.1 Database construction

A database with 365 observations is collected from experimental data from the literature [2–8]. The input and output, as well as statistical parameters associated, are displayed in Table 1. Both dynamic and quasi-static data are used to build the database. Dynamic and quasi-static methods are known to yield different values of elastic moduli (e.g. [9]). This fact adds to the uncertainties already present in experimental assessment of the elastic properties. To check the trustfulness of the data collected, we computed the theoretical bounds of Voigt-Reuss and Hashin-Shtrikman. All data collected regarding Young modulus respect the theoretical bounds.

Table 1. Statistical parameters of the training dataset on OPC cement pastes. Datasize of 365 observations. *DOH* refers to degree of hydration, w/c to the (mass) water-cement ratio, and the mass fraction of the four main clinker minerals (mC_3S, mC_2S, mC_3A and mC_4AF) and gypsum (*mGyp.*) are reported in %.

Data	Variable	Min.	Max.	Mean	St. Dev	Ex. Kur.*
Input	*DOH* [-]	0.03	1.0	0.46	0.29	-0.68
	w/c [-]	0.25	0.80	0.44	0.10	-0.20
	mC_3S [%]	24.5	100	60.2	13.8	2.06
	mC_2S [%]	0	61.3	16.6	15.2	2.43
	mC_3A [%]	0	12.7	8.13	3.36	-0.70
	mC_4AF [%]	0	12.7	5.8	4.2	-1.53
	mGyp. [%]	0	6.8	2.9	2.9	-1.7
Output	*E* [GPa]	0.22	37.2	11.2	7.8	0.32

*normalized exceed kurtosis

2.2 Machine learning methods

We consider here two ML methods:

- **Artificial Neural Network (ANN)**: with this method a neural network is constituted of stacked layers. Each layer is related to simple computation. The information is then processed layer by layer starting at the input layer until the output layer. The neural network is trained to minimize a loss function on the training set. A gradient descent method is used to perform this minimization.
- **Gaussian Process (GP)**: with this method predictions are made using Bayesian inference on the Gaussian process conditioned to the training data (see, for example [10]). The assumption behind the method is that the prediction functions can be described by a Gaussian process, which is defined by its kernel. The training phase with GP consists of estimating the parameters of this kernel.

The algorithms are built in Mathematica®.

3 Results and discussion

3.1 Validation

To validate our ML approach against the training dataset, we deploy the K-fold method [11]. This method consists of dividing the training dataset in K folds in which elements are randomly sampled. ANN or GP predictor is then trained on K-1 folds and, next, it is used to predict the values in the remaining folds. This step is repeated K times so that all folds have been used for validation. Fig. 1 shows one of the stages of validation. The accuracy of the predictions is computed as the square root of the average variance obtained from the validation on all K-folds for the elastic constants. We obtain an average deviation of the Young modulus of 0.61 GPa for ANN and of 1.19 GPa for GP with K = 5 (which is a usual value adopted for validation).

3.2 Predictions on testing dataset

We define a testing dataset composed of experimental data from the literature, which are not included in the training dataset. Prediction using ANN and GP of the Young modulus of OPC pastes are gathered in Table 2 and compared against their experimental counterpart.

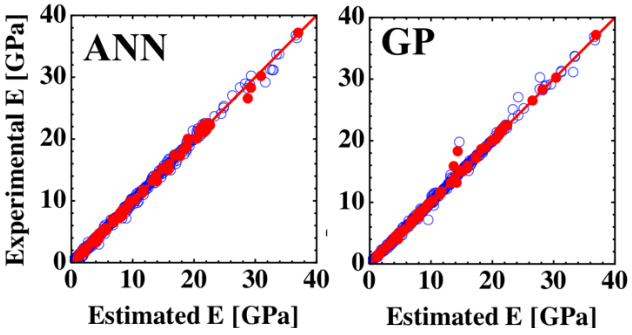

Fig. 1. An illustration of the K-fold (with K = 5) validation method: Experimental Young modulus plotted against the predicted Young modulus at one stage (out of 5) of validation for ANN and GP. The empty symbols are the values associated with the (K-1) folds (with 292 observations in total) used for training, and full symbols depict the 73 elements in the remaining fold used in the validation.

Table 2. Comparison of predictions with ANN and GP using the testing dataset (not comprised in the training dataset) against the experimental data provided in the literature by various authors. The mass fraction of clinker mineral and gypsum as provided by the authors as well as the w/c and degree of hydration were used as input in the predictions.

Data	w/c	Exp.	**ANN**	**GP**
Constatinides et al. [12]	0.50	22.8	**20.7**	**14.0**
Tamtsia et al. [13]	0.35	20.70	**14.2**	**22.6**
	0.35	20.80	**17.6**	**22.6**
	0.35	21.50	**18.5**	**22.7**
	0.5	7.90	**4.3**	**10.1**
	0.5	10.40	**7.4**	**12.4**
	0.5	11.60	**7.7**	**12.5**
Savija et al. [14]	0.3	16.68	**27.7**	**16.4**
	0.4	12.79	**22.9**	**14.1**
	0.5	9.09	**18.0**	**14.9**
Haecker et al. [3]	0.6	1.7	**6.9**	**10.6**

3.3 Using Micromechanics to provide missing data

We use Micromechanics, i.e. another technique to establish composition-property correlations in cement-based materials [15], to provide data that may be missing in our training dataset. In the sequel, we verify if the extended database performs better than the original database in the predictions.

For the micromechanics-based estimations, we adopt the representation of the microstructure of cement paste combining self-consistent (SC) scheme to estimations at

the gel scale, and Mori-Tanaka scheme to estimations at the cement paste scale, as shown in Fig. 2A. Similar representations were successfully used in previous work to estimate mechanical [16, 17], thermal [18], and electromagnetic properties [19, 20] of cement pastes.

The accuracy of micromechanics-based estimations is illustrated in Fig. 2B, in which the data associated with late ages (DOH > 0.7) are identified. The average deviations in the Young modulus estimated by micromechanics are 6.9 GPa when all data is considered and 2.8 GPa when only late age data is taken into account.

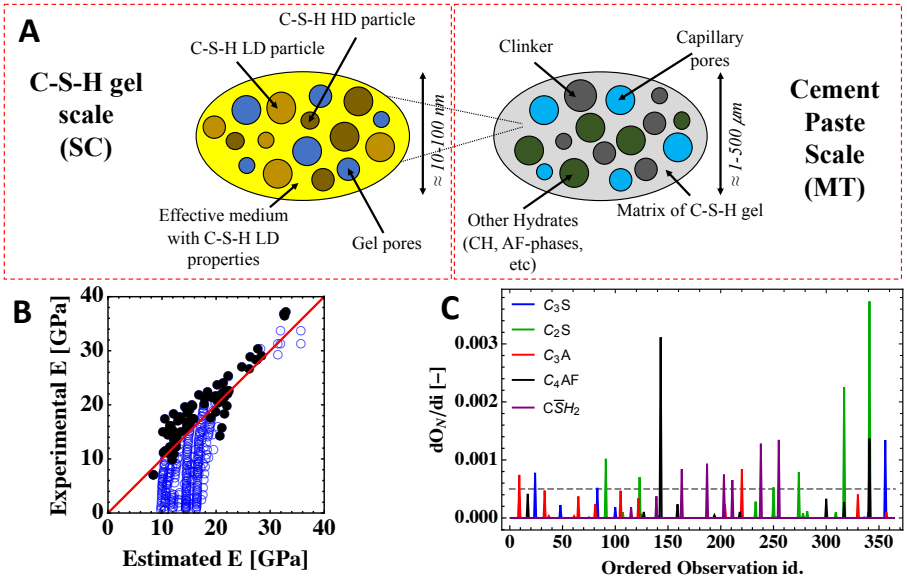

Fig. 2. A) Adopted representation of the microstructure of cement paste in the micromechanics estimations: combinations of self-consistent scheme estimate at the gel scale, and Mori-Tanaka estimate at the cement paste scale. B) Comparison of experimental and micromechanics predicted Young modulus taking into account the training database (black points corresponds to late ages (DOH > 0.7). C) Identification of the main domains with missing data regarding OPC composition using the derivative approach of the ordered input.

Among the variety of possibility to guide the plan of experiments [20], the strategy deployed here to identify the domain in the 7-dimension space of input variables is to take a derivative of the ordered components of each input vector. The higher the peak in the derivative, the most significant is the domain with missing data. Fig. 2C shows the derivative for the mass fraction of OPC constituents. With this approach the points related to the most significant domain with missing data were identified as follows:

- DOH [-]: 0.70, 0.95
- w/c [-]: 0.28, 0.37, 0.72
- mC_3S [%]: 33.0, 52.4, 85.2
- mC_2S [%]: 5.40, 11.2, 16.7, 20.7, 31.7, 49.9

- mC_2S [%]: 1.10, 9.95
- mC_2S [%]: 4.40 11.2
- *mGyp.* [%]: 1.05, 2.00, 2.90, 3.65, 4.68, 6.09

where an arbitrary cut-off of 0.0001 for the w/c and degree of hydration, and 0.0005 for the clinker minerals and gypsum mass fractions was adopted.

In Fig. 3, we compare the average deviations of predictions in the testing dataset using the original (l_{orig}) and extended ($l_{Ex,6}$ and $l_{Ex,24}$) training datasets. The extended dataset is composed of a (i) $l_{Ex,6}$ with the minimum number of input points (6) comprising all the points above related to the most significative missing data domains; and (ii) $l_{Ex,24}$ with each point (24 in total) being associated with one of the points above related to the most significant missing data domains and the other input components are chosen also among these points related to the most significative missing data domains.

We observe that for both ANN and GP, using the extended dataset yields smaller average deviations than using the original dataset. This is an interesting result showing that micromechanics can be used to generate missing data in dataset for training ML methods.

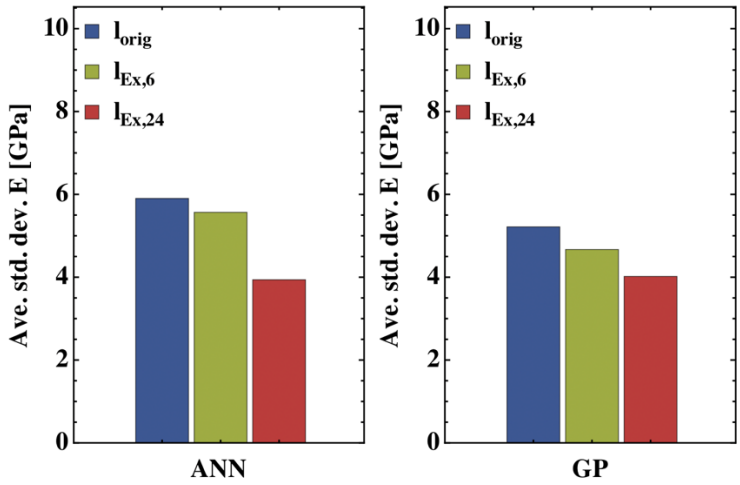

Fig. 3. Average deviations of predictions in the testing dataset using the original (l_{orig}) and extended ($l_{Ex,6}$ and $l_{Ex,24}$) training datasets.

3.4 Towards the estimation of the composition from a target elastic modulus

To estimate the composition from the elastic response, i.e. do the reverse path in the prediction done so far in this study, we adopt a Monte Carlo (MC) approach. We sample the given input following a uniform distribution with the limits defined by the minimum and maximum values in Table 1. We set the DOH at 1 to avoid the uncertainties related to hydration kinetics and early-age effects. Since we use as input the mass fraction of clinker minerals and gypsum, as a consistency condition, the sum of

all masses should be 100%, which means that not all mass fractions can vary independently in the MC trials, one of the phases mass fractions (we adopted C_3S) must be computed from the consistency condition. A random sample with 10000 MC trials is computed for each independent input variable.

The probability distribution functions of each input from MC trials yielding predictions that comply with the target 16.68±2 GPa (which corresponds to one of the experimental data from ref. [14]) are shown in Fig. 4. Yellow and blue histograms correspond to ANN and GP predictions, respectively. We also indicate the composition of the system according to the data provided in ref. [14]. The reverse predictions using GP yield distributions of input variables containing the experimental composition data. The distribution obtained with ANN failed in containing the experimental w/c. For both ML methods, the distribution obtained can be rather broad, which may indicate that the reverse problem might be ill-posed.

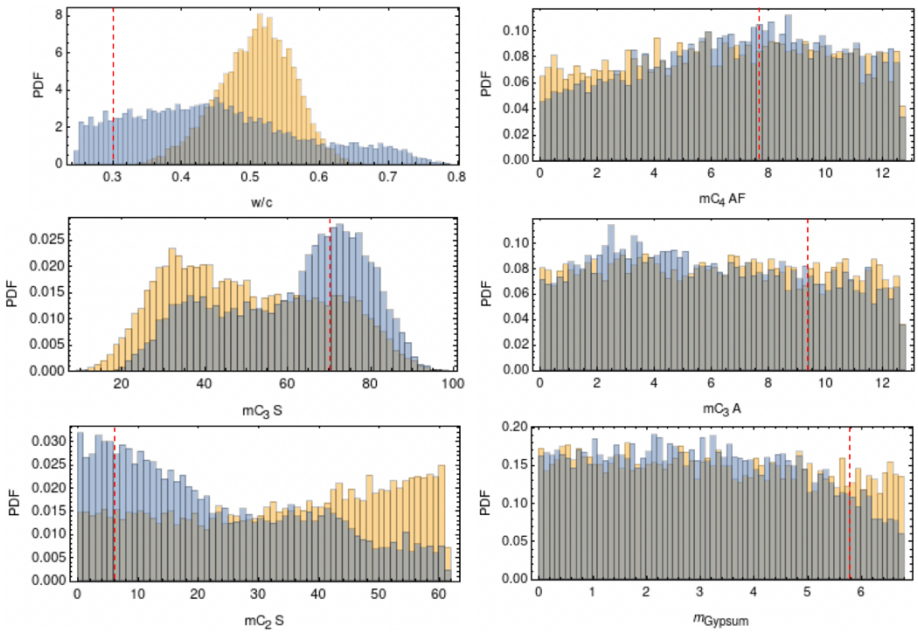

Fig. 4. Probability distribution function of each input from MC trials yielding predictions that comply with the target 16.68±2 GPa (which corresponds to one of the experimental data from ref. [14]). Yellow and blue histograms correspond to ANN and GP predictions, respectively. The red dashed line indicates the composition of the system according to data from ref. [14].

4 Conclusion

In this work, Machine Learning methods were deployed to establish correlations between the composition and the Young modulus of OPC pastes. Both ANN and GP

prediction performed here are not computer-intensive once the predictor function is trained, especially when one compares to the increasingly fastidious numerical homogenization approaches. This aspect indicates a clear advantage of ML methods, in particular in an application in which a larger exploration of the compositional and mixing design space is desired. ML prediction yields adequate precision with the training dataset proposed. Even in a relatively small dataset, ANN and GP have proven to be reliable and robust in the prediction of elastic constants of cement pastes from their composition. Micromechanics can be an ally of ML approaches by providing data missing in the database, increasing the accuracy of ML predictions. This result adds to the accumulating evidence showing that micromechanics-based approaches are a powerful tool to correlate composition and property in cement-based materials based on a few fundamental component data set and assumptions.

The strategy outlined in this work to explore the space of formulation design of cement pastes and correlating their composition to their effective Young modulus can be extended to other properties of cement-based materials.

References

1. JCI (2008) Japan Concrete Institute guidelines for control of cracking of mass concrete

2. Maruyama I, Igarashi G (2014) Cement Reaction and Resultant Physical Properties of Cement Paste. Journal of Advanced Concrete Technology 12:200–213. https://doi.org/10.3151/jact.12.200

3. Haecker C-J, Garboczi EJ, Bullard JW, et al (2005) Modeling the linear elastic properties of Portland cement paste. Cement and Concrete Research 35:1948–1960. https://doi.org/10.1016/j.cemconres.2005.05.001

4. Helmuth RA, Turk DH (1966) Elastic moduli of hardened Portland cement and tricalcium silicate pastes: effect of porosity. REt EIVED 135

5. Boumiz A, Vernet C, Tenoudji FC (1996) Mechanical properties of cement pastes and mortars at early ages: Evolution with time and degree of hydration. Advanced Cement Based Materials 3:94–106. https://doi.org/10.1016/S1065-7355(96)90042-5

6. Chamrova R (2010) Modelling and measurement of elastic properties of hydrating cement paste. PhD Thesis, EPFL

7. Wang X, Subramaniam KV (2011) Ultrasonic monitoring of capillary porosity and elastic properties in hydrating cement paste. Cement and Concrete Composites 33:389–401. https://doi.org/10.1016/j.cemconcomp.2010.11.001

8. Sun Z, Garboczi EJ, Shah SP (2007) Modeling the elastic properties of concrete composites: Experiment, differential effective medium theory, and numerical simulation. Cement and Concrete Composites 29:22–38. https://doi.org/10.1016/j.cemconcomp.2006.07.020

9. Benboudjema F, Carette J, Delsaute B, et al (2019) Mechanical Properties. In: Fairbairn EMR, Azenha M (eds) Thermal Cracking of Massive Concrete Structures: State of the Art Report of the RILEM Technical Committee 254-CMS. Springer International Publishing, Cham, pp 69–114

10. Williams CKI, Rasmussen CE (1996) Gaussian Processes for Regression. 7

11. Bengio Y, Grandvalet Y (2004) No Unbiased Estimator of the Variance of K-Fold

Cross-Validation. Journal of Machine Learning Research 5:1089–1105

12. Constantinides G, Ulm F-J (2004) The effect of two types of C-S-H on the elasticity of cement-based materials: Results from nanoindentation and micromechanical modeling. Cement and Concrete Research 34:67–80. https://doi.org/10.1016/S0008-8846(03)00230-8

13. Tamtsia BT, Beaudoin JJ, Marchand J (2004) The early age short-term creep of hardening cement paste: load-induced hydration effects. Cement and Concrete Composites 26:481–489. https://doi.org/10.1016/S0958-9465(03)00079-9

14. Šavija B, Zhang H, Schlangen E (2020) Micromechanical testing and modelling of blast furnace slag cement pastes. Construction and Building Materials 239:117841. https://doi.org/10.1016/j.conbuildmat.2019.117841

15. Wyrzykowski M, Sanahuja J, Charpin L, et al (2017) Numerical benchmark campaign of COST Action TU1404 – microstructural modelling. RILEM Technical Letters 2:99–107. https://doi.org/10.21809/rilemtechlett.2017.44

16. Honorio T, Brochard L, Bary B (2018) Statistical variability of mechanical fields in thermo-poro-elasticity: Multiscale analytical estimations applied to cement-based materials at early-age. Cement and Concrete Research 110:24–41. https://doi.org/10.1016/j.cemconres.2018.05.010

17. Honorio T, Bary B, Benboudjema F (2016) Multiscale estimation of ageing viscoelastic properties of cement-based materials: A combined analytical and numerical approach to estimate the behaviour at early age. Cement and Concrete Research 85:137–155. https://doi.org/10.1016/j.cemconres.2016.03.010

18. Honorio T, Bary B, Benboudjema F (2018) Thermal properties of cement-based materials: Multiscale estimations at early-age. Cement and Concrete Composites 87:205–219. https://doi.org/10.1016/j.cemconcomp.2018.01.003

19. Honorio T, Carasek H, Cascudo O (2020) Electrical properties of cement-based materials: Multiscale modeling and quantification of the variability. Construction and Building Materials 245:118461. https://doi.org/10.1016/j.conbuildmat.2020.118461

20. Honorio T, Bore T, Benboudjema F, et al (2020) Dielectric properties of the pore solution in cement-based materials. Journal of Molecular Liquids 302:112548. https://doi.org/10.1016/j.molliq.2020.112548