Cement and Concrete Research 115 (2019) 379–388

Contents lists available at ScienceDirect

Cement and Concrete Research

journal homepage: www.elsevier.com/locate/cemconres

Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?: New insights from statistical analysis and machine learning methods

Benjamin A. Young[a], Alex Hall[b], Laurent Pilon[a], Puneet Gupta[c], Gaurav Sant[d,e,f,*]

[a] Department of Mechanical and Aerospace Engineering, Henry Samueli School of Engineering and Applied Science, University of California, Los Angeles, CA 90095, United States
[b] Suffolk Construction, Boston, MA 02119, United States
[c] Department of Electrical and Computer Engineering, Henry Samueli School of Engineering and Applied Science, University of California, Los Angeles, CA 90095, United States
[d] Laboratory for the Chemistry of Construction Materials (LC^2), Department of Civil and Environmental Engineering, Henry Samueli School of Engineering and Applied Science, University of California, Los Angeles, CA 90095, United States
[e] California Nanosystems Institute, University of California, Los Angeles, CA 90095, United States
[f] Department of Materials Science and Engineering, Henry Samueli School of Engineering and Applied Science, University of California, Los Angeles, CA 90095, United States

ARTICLE INFO

Keywords:
Strength
Mixture design
Predictive modeling
Machine learning
Regression analysis

ABSTRACT

The use of statistical and machine learning approaches to predict the compressive strength of concrete based on mixture proportions, on account of its industrial importance, has received significant attention. However, previous studies have been limited to small, laboratory-produced data sets. This study presents the first analysis of a large data set (> 10,000 observations) of measured compressive strengths from actual (job-site) mixtures and their corresponding actual mixture proportions. Predictive models are applied to examine relationships between the mixture design variables and strength, and to thereby develop an estimate of the (28-day) strength. These models are also applied to a laboratory-based data set of strength measurements published by Yeh et al. (1998) and the performance of the models across both data sets is compared. Furthermore, to illustrate the value of such models beyond simply strength prediction, they are used to design optimal concrete mixtures that minimize cost and embodied CO_2 impact while satisfying imposed target strengths.

1. Introduction

A concrete's compressive strength after 28 days of aging is the most commonly used metric of its engineering properties and performance and forms a critical input in structural design [1,2]. Indeed, structural concrete is most often specified on the basis of its compressive strength after 28 days of aging, and the compressive strength is known to be proportional to other mechanical properties such as the flexural and tensile strength [1]. Furthermore, it is well-known that concrete strength is chiefly influenced by w/c (water-to-cement ratio, mass basis). But, even for a given w/c, substantial variations in concrete strength may be observed based on the characteristics of constituent materials; e.g., cement type, the type of aggregate used, the paste content, mineral and chemical admixtures, etc. [1,2]. Therefore, a robust, predictive model that could estimate compressive strength as a function of the mixture proportions would be useful in enabling high-throughput mixture design, and reducing the empirical, labor intensive nature of “trial batching” approaches that are the basis of industrial practice today.

Physical models that are capable of strength prediction (i.e., without empirical calibration) are difficult to construct, due to: (i) our inability to rigorously model cement hydration, and microstructure development, (ii) the unavailability of constituent material properties – for example, while the mechanical properties of the constituent materials, and cement hydrates are better known, chemical data especially as needed to model reaction kinetics is much less available [3–5], (iii) nonlinear elastic behavior of the cement paste which evolves with time [6–9], and (iv) the unpredictable effects of mineral and chemical

* Corresponding author at: Laboratory for the Chemistry of Construction Materials (LC^2), Department of Civil and Environmental Engineering, Henry Samueli School of Engineering and Applied Science, University of California, Los Angeles, CA 90095, United States.
E-mail address: gsant@ucla.edu (G. Sant).

https://doi.org/10.1016/j.cemconres.2018.09.006
Received 16 December 2017; Received in revised form 30 July 2018; Accepted 21 September 2018
Available online 27 September 2018
0008-8846/

admixture interactions, such as water-reducing admixtures (WRAs) or air-entraining admixtures (AEAs), on cement hydration and strength development. Therefore, there is great interest in applying statistical and machine learning (ML) methods (e.g. multiple linear regression, artificial neural networks) to model compressive strength evolution as a function of the concrete's mixture proportions.

However, the vast majority of prior studies based on statistical or ML approaches have been limited to smaller data sets consisting of around 1000 compressive strength measurements or less [10–19]. Furthermore, these data sets typically encompass laboratory specimens produced under controlled conditions. As such, it is unclear as to how well these predictive models may perform when applied to data collected from industrial concrete production ("ready mix") operations. This is because production data is likely to contain more unexplainable variance whose effects on the accuracy of strength prediction models remains unknown. Therefore, the present study aims to evaluate the performance of predictive models for estimating a concrete's compressive strength using a large data set of job-site based concrete strength measurements.

2. Background

2.1. Machine learning/data mining algorithms

Machine learning and data mining algorithms that have been developed over the past few decades provide a means of developing predictive models from empirical data, without a need for detailed knowledge of the underlying physical mechanisms [20,21]. Therefore, such models may be well-suited for predicting a concrete's compressive strength – a material property that can be influenced by compositional, processing and testing variables. This section provides a brief overview of the modeling techniques used in this study while further details regarding their formulation, and implementation can be found elsewhere [21].

2.1.1. Artificial neural networks (ANNs)

Artificial neural networks (ANNs) are statistical models that seek to determine input-output relationships via a series of connected data structures or "neurons" [21,22]. The neurons are organized into layers, with each neuron being functionally related to all neurons in the previous layer. Fig. 1 shows a schematic of a typical neural network model with $p = 4$ input variables, a hidden layer with $k = 3$ hidden neurons, and a single output.

Mathematically, the relationship between each hidden neuron and the input variables can be expressed as [21],

$$h_i = \sigma(\mathbf{w_i}^T\mathbf{x}) \tag{1}$$

where, $\mathbf{x} = [x_1, \ldots, x_p]^T$ is the vector of p input variables, $\mathbf{w_i} = [w_{i,\ 1}, \ldots, w_{i,\ p}]^T$ are the "weights" corresponding to each input variable, and σ is a nonlinear "activation function." For regression problems, a sigmoidal function is most commonly used as the activation function, i.e., [21],

$$\sigma(\mathbf{w_i}^T\mathbf{x}) = \frac{1}{1 + e^{\mathbf{w}_i^T\mathbf{x}}} \tag{2}$$

Similarly, the predicted output $\widehat{Y}$ is related to the hidden neurons h according to [21],

$$\widehat{Y} = g(\boldsymbol{\beta}^T\mathbf{h}) \tag{3}$$

where, $h = [h_1, \ldots, h_k]^T$ is the vector of hidden neuron values and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_k]^T$ is another set of weight factors. Here again, g can be a nonlinear function, however, it is typically taken as the identity matrix such that $\widehat{Y} = \boldsymbol{\beta}^T\mathbf{h}$.

Neural networks are "trained" by identifying the appropriate values of the weight factors $\mathbf{w}$ and $\boldsymbol{\beta}$ which minimize a measure of prediction error for a given training data set. Often, the root-mean-squared error (RMSE) provides an adequate measure of error, and can be expressed as,

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{(Y_i - \widehat{Y}_i)^2}{N}} \tag{4}$$

where, Y_i is the observed or measured output corresponding to observation i, $\widehat{Y}_i$ is the predicted output, and N is the number of observations or data points in the training set. Common numerical methods such as the Levenberg–Marquardt algorithm [22], can be used to identify the appropriate weight factors. The large number of fitting parameters in neural network models (i.e., the weight factors $\mathbf{w}$ and $\boldsymbol{\beta}$) allows them to easily identify nonlinear interactions between input variables, resulting in a powerful predictive tool. But, if used indiscriminately, neural networks that have a large number of hidden layers and/or hidden neurons in each layer are prone to over-fitting data, leading to generalization errors [20,21]. Therefore, the size of a neural network should be chosen carefully based on cross-validation, as described below.

2.1.2. Decision trees

Decision trees are a family of machine learning methods that can be used for both data classification and regression problems [20,21,23]. In contrast to neural networks, decision trees are "rule-based" models, i.e., they aim to identify logical splits in the data rather than fit a set of parameters in a mathematical formula. In other words, the tree splits the input space into a series of partitions or "leaf nodes," and then uses a simple model (i.e., often simply a constant value) to predict the output in each partition [21]. The splits are selected to minimize some metric of error, typically the sum-of-squared errors between predicted and observed outputs (Eq. (4)). A popular method of determining the splits is the classification and regression tree (CART) algorithm [23]. Clearly, the size of a tree must be limited in some way to prevent the tree from becoming too large and over-fitting the data; this is typically done by limiting the number of splits, or by ensuring that each leaf node contains a minimum number of data points [21,23].

The performance of decision tree regression models can be improved by building ensembles, or large collections of individual decision trees, and aggregating their predictions. Such ensemble methods can both increase prediction accuracy and reduce over-fitting and generalization errors [21]. One popular ensemble model is the so-called random forest model, which trains a large number of trees individually using only a random subset of the input variables [20,21]. In addition, each tree does not use the entire set of training data, but rather a bootstrap sample of the training data [21]. This procedure is known as bootstrap aggregation or "bagging." The predictions of each individual tree are then averaged to obtain the prediction of the random forest ensemble. Gradient boosting [21] is another established tree ensemble method. In this case, an initial tree is trained using the entire set of input data and all input variables. Then, a second tree is trained to fit the residuals of the first tree (i.e., the differences between the predicted and observed values) to the input data. This procedure is repeated for a specified number of iterations (typically several hundred). Then, the predictions of each tree are added to obtain the predictions of the gradient-boosted ensemble [20,21]. In order to avoid over-fitting, the contribution of each tree beyond the first to the sum is reduced by a factor known as the learning rate, which typically varies between 0.01 and 0.1 [20] and can be tuned by cross-validation.

2.1.3. Support vector machines (SVMs)

Support vector machines (SVMs) are a family of regression models that are effective for nonlinear problems [20,21,24,25]. Rather than using the sum-of-squared errors (Eq. (4)) as an objective function when fitting the model, SVMs use another measure of error known as hinge loss or ε-insensitive loss. In this case, errors smaller than a set threshold ε do not contribute to the overall error measure. The hinge loss function

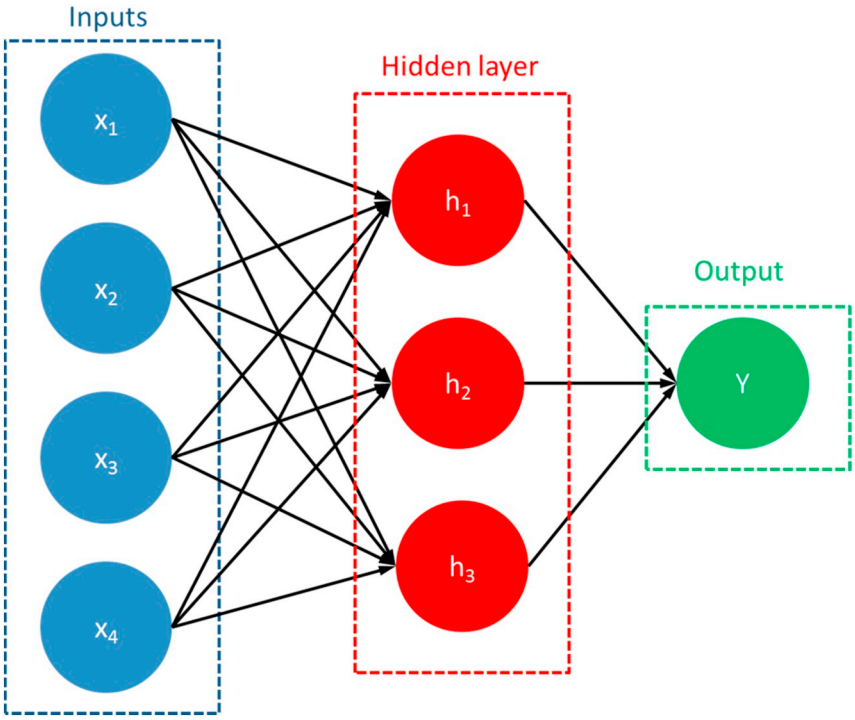

Fig. 1. A schematic of a typical feed-forward artificial neural network (ANN) with a single hidden layer.

can then be expressed as,

$$L(Y_i - \widehat{Y_i}) = \begin{cases} 0 & \text{if} \quad |Y_i - \widehat{Y_i}| < \epsilon \\ |Y_i - \widehat{Y_i}| - \epsilon & \text{if} \quad |Y_i - \widehat{Y_i}| > \epsilon \end{cases}. \tag{5}$$

SVMs seek to fit a model of the form,

$$\widehat{Y}(\mathbf{x}) = \sum_{i=1}^{N} c_i k(\mathbf{x}, \mathbf{x}_i) \tag{6}$$

where, the parameters c_i are referred to as choice coefficients, and the Gaussian kernel function $k(\mathbf{x}, \mathbf{x}_i)$ is defined as,

$$k(\mathbf{x}, \mathbf{x}_i) = e^{-\|\mathbf{x}-\mathbf{x}_i\|^2}. \tag{7}$$

This kernel function measures the similarity, quantified as a Gaussian distance, between the set of inputs $\mathbf{x}$ and those of the i^{th} point in the training set $\mathbf{x_i}$. Note that because such a quantity is used, the input data should be centered and scaled (i.e. subtracted from its mean and divided by its standard deviation) such that each variable is rendered unitless. Furthermore, Eq. (6) indicates that there are as many parameters (the N choice coefficients) in the model as points in the training set. To make the problem mathematically feasible, a regularization term (i.e., a penalty term for large choice coefficients) must be added to the objective function. The objective function can then be expressed as,

$$J = \sum_{i=1}^{N} L\left(Y_i - \sum_{j=1}^{N} c_j k(\mathbf{x}_i, \mathbf{x}_j)\right) + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j k(\mathbf{x}_i, \mathbf{x}_j). \tag{8}$$

The regularization factor λ determines how much the large choice coefficients are penalized and should be tuned appropriately via cross-validation [21].

2.2. Previous studies on concrete strength prediction using statistical methods

Numerous authors have applied statistical and machine learning techniques to predict the compressive strength of concrete based on its mixture proportions [10–19]. Notably, Yeh et al. [10] published a dataset consisting of 1031 measured compressive strengths as a function of w/c, cement, fly ash and blast furnace slag contents, coarse and fine aggregate contents, superplasticizer dosage, and age, and used artificial neural networks (ANNs) to develop a strength prediction model. The authors found that their ANN model offered a coefficient of determination (R^2) as high as 0.92 when comparing measured versus predicted strengths on their test data. Several other studies [11,12,18,26] have used the same dataset published by Yeh et al. [10] to evaluate the performance of other machine learning models in predicting compressive strength. These studies too have reported similar results, with various models including neural networks [11,12], boosted and bagged regression trees [11], and SVMs [12] resulting in $R^2 > 0.9$. In addition, other studies [13–17,19] have used different, smaller data sets to develop predictive models for compressive strength, primarily using neural networks [13–17,19]. Because most studies on compressive strength prediction have used the same data set, or very small data sets secured under carefully-controlled laboratory conditions, it is unclear how well these predictive models can estimate the strengths of industrially produced ready-mix concrete (RMC). This is significant as in industrial operations, variables such as the ambient temperature, type of mixing (central mixing, or truck mixing) or the moisture content of aggregates may be controlled imprecisely or not at all [27–29]. Therefore, one would expect the performance of predictive models to suffer from the extra noise introduced into the data by these unobserved and/or uncontrolled variables.

The present study aims to quantify the extent of such noise on model performance, and identify if such uncontrolled parameters may induce unexplainable variance in compressive strength predictions of industrially-produced concretes.

3. Analysis

3.1. Data collection

Two data sets were considered in this study. The first was the dataset published by Yeh et al. [10] consisting of 1031 measured compressive strengths, as previously discussed. Because this data set has been used by

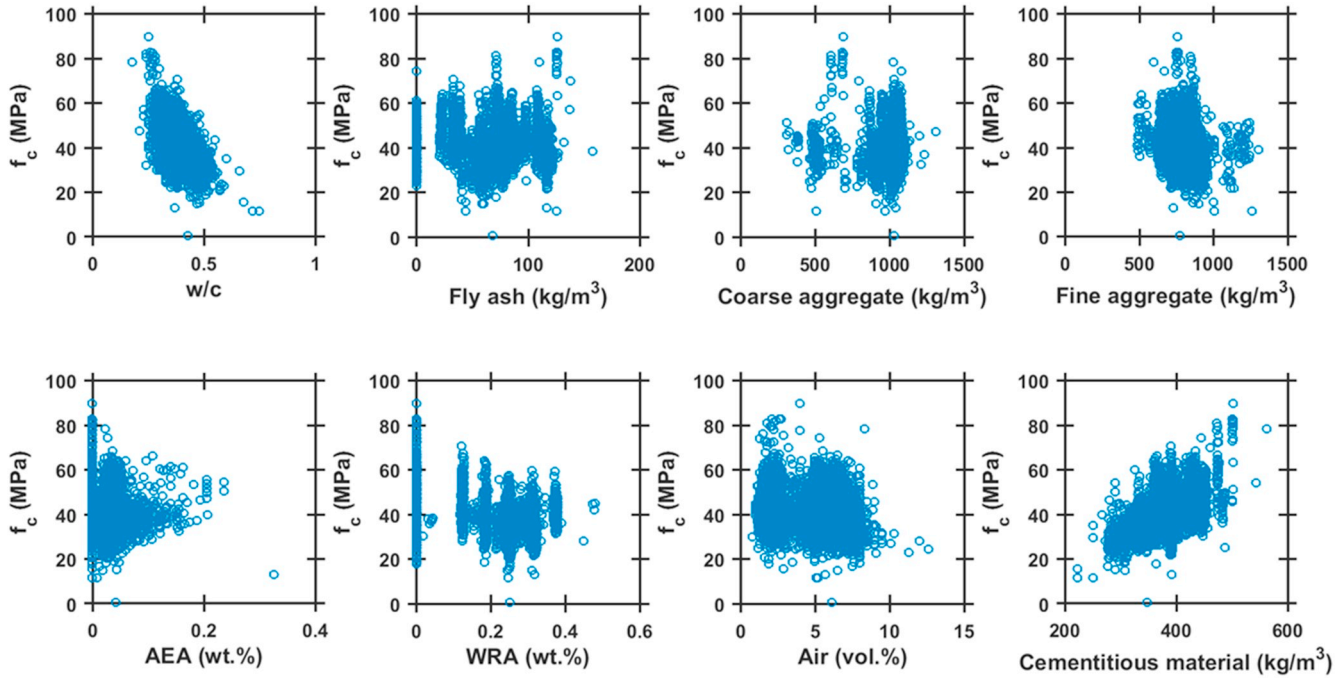

Fig. 2. The relationship between the eight input (i.e., mixture proportion) variables and the concrete's compressive strength in the VIP's data set.

several previous studies, it provided a useful benchmark by which to gauge model performance. Furthermore, as the present study is not focused on the effect of age on concrete strength (N.B.: the 28 days strength is most often captured during industrial production), only data for samples with ages greater than or equal to 28 days were considered. In total, 706 of the 1031 measurements were used from the data set of Yeh et al. while a subset of the data was neglected on account of having strengths measured before this 28-days reference age [10].

The second data set was provided by an international vertically-integrated cement/concrete producer (VIP) from across a range of different concrete production sites and consisted of 9994 measured compressive strengths from job-site mixtures that were sampled across various locations in the United States. Three samples corresponding to a given mixture was collected at the job-site (i.e., in the form of 6″ × 12″ cylinders) and cured following ASTM C39 for 28 days after which time their compressive strengths were measured. The data set contained mixture proportions in terms of: w/c, cement and fly ash contents (in kg per m^3 of concrete), water-reducing admixture (WRA) and air-entraining admixture contents (in kg per 100 kg of cementitious material), coarse and fine aggregate contents (in kg per m^3 of concrete), and fresh air content (in volume %) for each mixture. The mixture proportions reported reflect the actual mixture proportions, i.e., based on the batch weights. Furthermore, all mixtures in the industrial dataset used ASTM C150 compliant Type I/II OPC. In general, Class F fly ash compliant with ASTM C618 was used in all relevant cases. Finally, the aggregates used were compliant with ASTM C33.

Prior to model training and evaluation, the VIP data set was subject to a preprocessing step in which measurements that appeared to be corrupted by obvious recording/procedural errors were eliminated. However, measurements were not removed solely on the grounds of being classified as outliers, as doing so would provide an overly optimistic view of performance. Indeed, robustness to such outliers is an important quality of a useful predictive model, and care must be taken to ensure that training and validation data are not manipulated excessively prior to analysis.

To illustrate the variables in the data set, Fig. 2 plots the concrete's compressive strength (f_c, MPa) as a function of the eight input variables in the VIP data set. As expected, there is an inverse relationship between w/c and compressive strength, however, the range of measured compressive strengths for a given w/c is still very large. Furthermore, it is difficult to recognize the effect of each input variable on the compressive strength with seemingly no trends being visually evident in the case of several (supposedly influential) input variables including: the fly ash content, fine aggregate content, etc.

3.2. Data splitting

In order to obtain estimates of the generalization error for the predictive models developed in this study, each data set was partitioned into training and test sets. The data was partitioned at random, with around 80% of all data points used for training data and the remaining 20% used for validation. For the Yeh et al. data set, 564 data points were used as training points while the remaining 142 were used as validation points. For the VIP dataset, the training set consisted of 7995 measurements and the validation data set consisted of 1999 measurements. In each case, the training data set was used to develop and tune the strength prediction models, while the validation data was used only for final assessments of model performance.

3.3. Model tuning and cross-validation

Compressive strength prediction models were implemented following four distinct modeling approaches: neural network, gradient-boosted tree, random forest, and SVM. As discussed previously, each of these models involves a number of extra parameters that must be appropriately tuned to optimize performance. To accomplish this, a 5-fold cross-validation

Table 1
The parameters obtained following cross-validation for each compressive strength prediction model.

Model	Parameters
Neural network	2 hidden layers with 10 and 5 neurons each
Random forest	500 trees in ensemble, 3 input variables used for each tree
Gradient-boosted tree	500 trees in ensemble, learning rate = 0.09
SVM	ε = 2 MPa, λ = 13.5

Table 2
The estimated bulk density and cost of each concrete mixture component.

Mixture component	Subscript	Density (kg/m^3)	Cost ($/kg)	Ref.
Cement	*C*	3150	0.110	[32]
Water	*W*	1000	0.000	
Fly ash	*fla*	2500	0.055	[33]
Coarse aggregate	*ca*	2500	0.010	[32]
Fine aggregate	*fa*	2650	0.006	[32]
WRA	*WRA*	1350	2.940	[34]

procedure was used [20,21] which consisted of the following steps:

1) To randomly partition the training set into 5 "folds",
2) To train the model using data from 4 of the 5 folds and use the 5th to estimate the error,
3) To repeat this procedure such that each of the 5 folds is used once to estimate the error,
4) To average the estimated errors from each of the 5 folds to obtain the cross-validation error.

This procedure provides an estimate of the generalization error within the same training sample. It can then be used to optimize the model parameters by finding the values of the parameters that minimize the cross-validation error. Table 1 summarizes the parameters used by each model as tuned by cross-validation.

3.4. Concrete mixture optimization

The ability to estimate compressive strengths based on the mixture proportions would allow a concrete producer to design and proportion "optimal" mixtures that minimize monetary cost and/or embodied CO_2 impact, while still meeting a designer's strength requirements. Here, this is demonstrated by first identifying mixtures of minimal cost for a given range of strengths (i.e., a two-objective optimization problem) and then identifying mixtures that minimize cost and embodied CO_2 impact while still adhering to specified target strengths (three-objective optimization). Because there are multiple objectives in each optimization scenario, there is no single optimal solution but rather a series of optimal ("Pareto") solutions that emerge. Both optimization procedures were conducted using the artificial neural network (ANN) model trained on the VIP dataset. This model is better suited for use in a gradient-based optimization procedure because it is "smooth" with respect to the input parameters, as opposed to the tree-based models which may contain non-differentiable points. The optimization was carried out using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [30] within MATLAB's optimization toolbox. In an effort to identify global rather than local optima, a pattern search method was used to generate a large number of initial points. The optimization was performed for each of these points, and the best resulting solution was considered as the global optimum.

Because concrete mixtures with a higher cementitious content have higher compressive strengths, and because cementitious materials are more expensive than aggregates, there is a trade-off between the expected compressive strength and monetary cost of a mixture. The estimated monetary costs of each mixture component on a mass basis are listed in Table 2. The cost of mixing water was assumed to be negligible. The cost of fly ash can vary depending on location [31]; but to set a range, fly ash is proposed to cost between one-half to on par with cement (i.e., ordinary portland cement, OPC) as noted below.

The objective function is the total cost of the concrete mixture per m^3 of material, which can be expressed as,

$$\text{Total cost } (\$/m^3) = C_c p_c + C_{fla} p_{fla} + C_{fa} p_{fa} + C_{ca} p_{ca} + C_{WRA} p_{WRA} \quad (9)$$

where, C_i and p_i are the dosage (in kg/m^3) and cost (in $/kg) of mixture component *i*, respectively. The optimization problem is defined by a series of constraints, the first of which is that the compressive strength of the mixture must not be less than some target strength $f_{c,target}$, such that:

$$f_{c,pred}(\mathbf{x}) \geq f_{c,target} \quad (10)$$

where, $f_{c,pred}(\mathbf{x})$ is the compressive strength predicted by the neural network model as a function of the mixture parameters **x**. Next, it is necessary that the sum of the volume fractions of each mixture component add up to one,

$$\frac{C_c}{\rho_c} + \frac{C_w}{\rho_w} + \frac{C_{fla}}{\rho_{fla}} + \frac{C_{fa}}{\rho_{fa}} + \frac{C_{ca}}{\rho_{ca}} + \frac{C_{WRA}}{\rho_{WRA}} = 1 \quad (11)$$

where, ρ_i is the density of each component (in kg/m^3). Furthermore, upper and lower bounds were placed on some of the mixture parameters, as shown in Table 4. These bounds were based on the range of values observed in the training fraction of the VIP data set; extrapolation beyond them would likely lead to poor model performance and less certain results.

The dosage of WRA was not treated as a free parameter – instead, it was assumed to be directly proportional to the cementitious material content at a mass ratio of 0.0031 kg WRA per kg of cement [35]. Finally, in order to fully define the problem, the air content and AEA dosage (if present) are required as inputs to the strength prediction model. These parameters were assumed to be constant and to correspond to either air-entrained mixtures with an air content of 6 vol% and an AEA dosage of 0.03 mass % (by mass of cement), or to non-air-entrained mixes with an air content of 2 vol%; wherein no AEA is added. These values were based on the average values in the VIP dataset for air-entrained and non-air entrained concrete mixtures.

4. Results and discussion

4.1. Compressive strength prediction

Table 4 summarizes the performance of each model when applied to the test set data from each of the two data sets. Three error metrics are reported, namely (i) root-mean-square error (RMSE, in MPa), (ii) R^2-value (i.e., the strength of the linear relationship between predicted and observed values), and (iii) mean absolute percentage error (MAPE). For comparison, the performance of a simple linear regression model of the form,

$$\hat{Y} = \boldsymbol{\beta}^T \mathbf{x} + \beta_0 \quad (12)$$

is also reported. First, Table 4 establishes that the R^2-value of the predictive models was higher when applied to the data set published by Yeh et al. [10]. In other words, the models were able to explain more of the variance in compressive strength for this data set, as compared to the job-site data from the VIP. However, the RMSE was also larger, indicating that the absolute compressive strength was predicted less accurately on average than for the VIP data set. Taken together, these results show that, unsurprisingly, the VIP data set features more unexplainable variance, than the data set of Yeh et al. Furthermore, while the more advanced models clearly out-performed the linear regression model for the Yeh et al. data set, this difference was far less substantial for the VIP data set; perhaps due to the increased noise (variance).

Overall, the contrast in model performance between data sets shows that while statistical/ML models are very effective in predicting compressive strength of laboratory-curated mixtures based on their mixture proportions, they are less effective in predicting the strength of concrete produced in an industrial setting. To improve predictive performance in the latter case, it may be necessary to expand the range of mixture (input) variables considered including: mixing temperature, mixing type (central, or truck mixing), relative humidity, aggregate source and type, moisture content, curing temperature profiles, potential job-site retempering of the concrete (if any), uncertainty in batch weights, changes in cement, or fly ash behavior upon silo restocking, etc. Despite

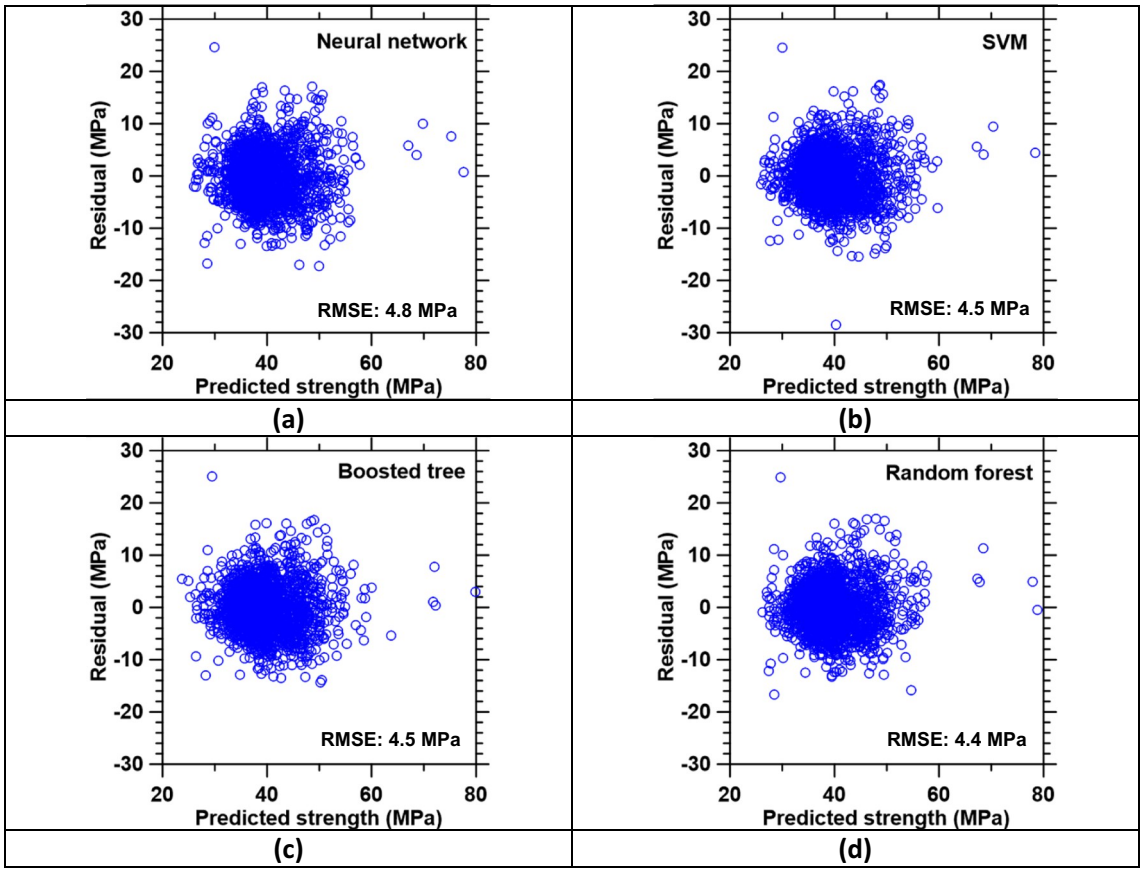

Fig. 3. The distribution of test set residuals from the VIP dataset as a function of predicted strengths for: (a) neural network, (b) SVM, (c) boosted tree, and (d) random forest models.

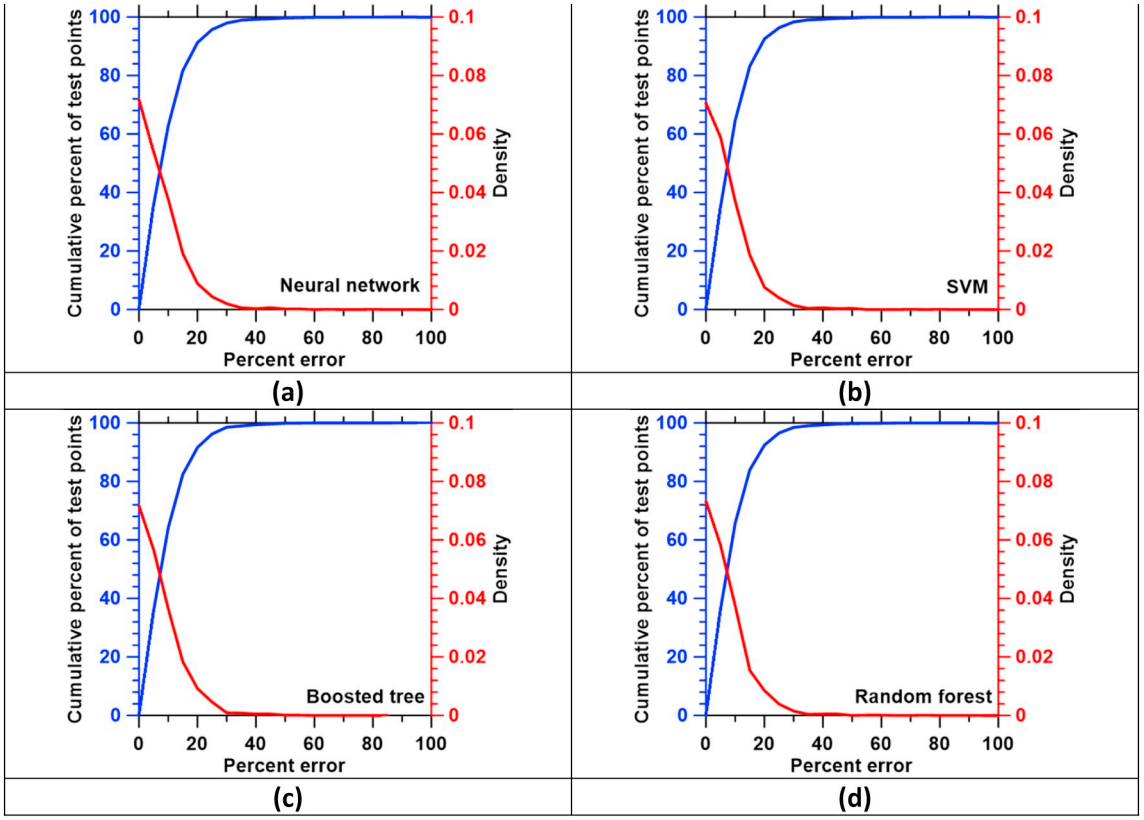

Fig. 4. The cumulative distribution of percentage error for: (a) neural network, (b) SVM, (c) boosted tree, and (d) random forest models.

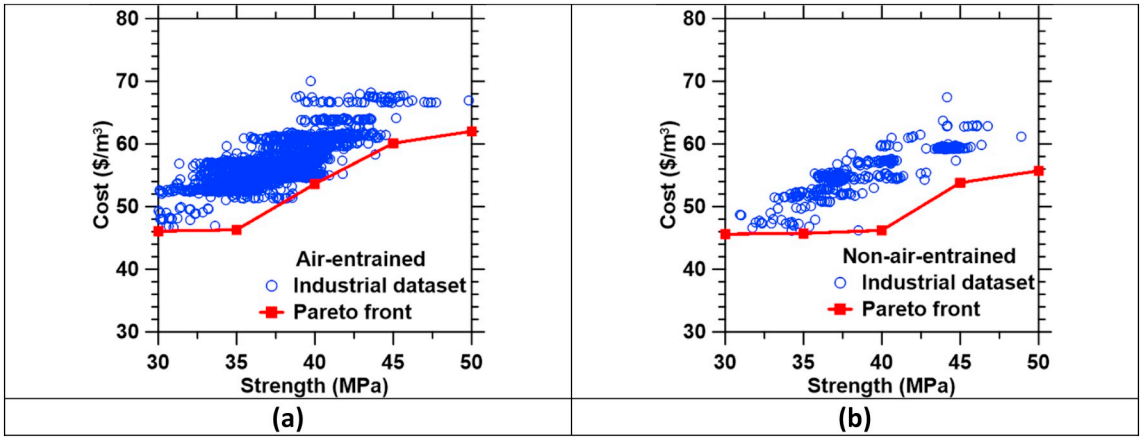

Fig. 5. The Pareto front showing the minimum possible (mixture) cost as a function of its target strength for: (a) air-entrained and (b) non-air entrained concrete mixtures. Also shown are the estimated strengths and costs of mixtures from the VIP data set. These mixture optimizations were carried out using the ANN model.

these shortcomings, each model was still able to predict compressive strength with an average relative error of < 10% – a very favorable outcome – even when applied to the industrial VIP data set. Further, while the performance of each machine learning method was similar, the random forest model exhibited the lowest RMSE and highest R^2-value for both data sets. However, such small differences in performance between models suggest that each is equally well-suited for predicting the compressive strength of industrial concrete, and therefore, selection between them can be made based on their ease of implementation when used for a procedure such as the mixture optimization examples that are demonstrated in the following section.

For clarity, Fig. 3 plots the test (validation data) set residuals from the VIP dataset (i.e., the difference between the measured and predicted strengths) as a function of predicted strengths for each of the four statistical models considered, namely: (a) neural network, (b) SVM, (c) boosted tree, and (d) random forest. First, Fig. 3 shows that there is no significant correlation between the predicted strengths and the residuals. Additionally, the distribution of residuals is similar across each of the four models. These observations are important because they suggest that the errors in each model's predictions are due to unexplainable variance in the data, rather than due to the models failing to recognize important interactions between the input variables and compressive strength. Finally, as an additional test of model performance, Fig. 4 plots the cumulative error distributions (i.e., showing the percentage of test set points that are predicted within a given error threshold) for each of the four models, along with the empirical probability density of the percentage errors. Indeed, each model was able to predict the strength within 10% relative error for over 60% of the test points, and within 20% error for over 80% of the test points. Once again, the performance of each model was similar, further reinforcing the conclusion that each was able to identify the underlying patterns represented in the training dataset.

Table 3
The upper and lower bound constraints placed on the mixture parameters.

Mix parameter	Expression	Lower bound	Upper bound
Cementitious material content	$C_c + C_{fla}$	300 kg/m^3	500 kg/m^3
w/c	$C_w/(C_c + C_{fla})$	0.20	0.60
Fly ash content	C_{fla}	0 kg/m^3	150 kg/m^3
Coarse aggregate content	C_{ca}	500 kg/m^3	1100 kg/m^3
Fine aggregate content	C_{fa}	600 kg/m^3	1200 kg/m^3
Fly ash/total cementitious material ratio	$C_{fla}/(C_c + C_{fla})$	0.00	0.30
Total volume fraction of aggregates	$\frac{C_{fa}}{\rho_{fa}} + \frac{C_{ca}}{\rho_{ca}}$	0.60	0.75
Coarse/fine aggregate ratio	$\frac{C_{fa}}{C_{ca}}$	0.50	1.00

Table 4
A comparison of model performance in predicting the compressive strength for each of the two data sets (Yeh et al. [10] and the VIP data set).

Model	Yeh et al. dataset			VIP data set		
	RMSE (MPa)	R^2	MAPE (%)	RMSE (MPa)	R^2	MAPE (%)
Linear regression	8.8	0.66	22	5.0	0.49	10
Neural network	6.3	0.82	14	4.8	0.54	9
Random forest	5.7	0.86	14	4.4	0.60	9
Boosted tree	5.8	0.85	13	4.5	0.59	9
SVM	6.4	0.83	15	4.5	0.59	9

4.2. Mixture optimization

4.2.1. Cost-strength optimization

Fig. 5 plots the minimum achievable mixture cost as a function of its target strength (i.e., the "Pareto front") for both: (a) air-entrained and (b) non-air entrained mixtures, computed via the previously described optimization procedure. Also shown are the predicted strengths and costs of the mixtures from the VIP data set. In general, concrete mixtures were considered to be air-entrained if they had an air content > 4 vol% and non-air entrained otherwise [36]. It should be noted that only mixtures that conformed to the imposed optimization constraints (shown in Table 3) within ± 5% are shown. In general, Fig. 5 establishes that all the job-site mixtures lie above the Pareto front, i.e., each mixture has a higher estimated cost than necessary to achieve its target strength. It should be noted, however, that some of the mixtures offered within the VIP dataset may have been subject to additional constraints depending on project specifications that were not considered in our optimization procedure.

Furthermore, Table 5 shows the resulting "optimal" mixture parameters for various target strengths for both air-entrained and non-air-entrained mixes. As expected, the cementitious material content and thus the minimum achievable cost of each mixture increased with increasing target strength. Furthermore, the costs of air-entrained mixes were higher, as higher air contents decrease the compressive strength and thus must be compensated for by adding more cementitious material. Interestingly, most of the optimized mixtures contained the maximum allowable amount of fly ash, i.e., fly ash comprised 30% of the total cementitious material. This suggests that the reduction in strength arising from cement replacement by fly ash is more than compensated for by 28 days by the reduced cost of the fly ash as

Table 5
The optimized mixture parameters for concretes that achieve a series of target strengths, for both air-entrained and non-air entrained concrete formulations.

Target strength (MPa)	Min Cost ($/m^3)	Optimal mixture parameters						Constraint values	
		w/c	Fly ash (kg/m^3)	Coarse aggregate (kg/m^3)	Fine aggregate (kg/m^3)	Total cementitious material (kg/m^3)	Total agg. vol.%	Fly ash/cementitious ratio	Fine/coarse aggregate ratio
Air-entrained									
30	46.1	0.51	90	955	955	300	0.74	0.30	1.00
35	46.3	0.6	90	1095	759	300	0.72	0.30	0.69
40	53.6	0.6	58.3	1100	639	352	0.67	0.17	0.58
45	60.1	0.24	129	1100	837	429	0.75	0.30	0.76
50	62.7	0.25	138	1100	776	458	0.72	0.30	0.71
Non-air entrained									
30	45.6	0.60	90	946	899	300	0.72	0.30	0.95
35	45.7	0.57	90	933	933	300	0.73	0.30	1.00
40	46.2	0.49	90	963	963	300	0.75	0.30	1.00
45	53.8	0.33	112	978	952	373	0.75	0.30	0.97
50	55.7	0.29	117	965	965	392	0.75	0.30	1.00

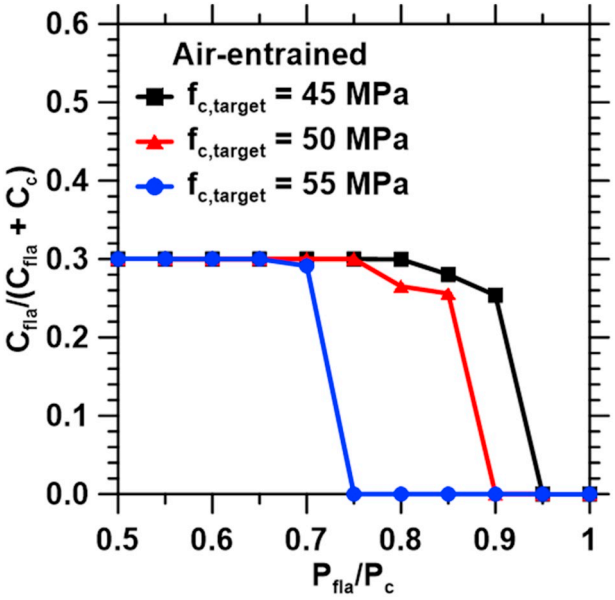

Fig. 6. The optimal fly ash content expressed as a fraction of the total cementitious content as a function of the price of fly ash scaled by the price of cement. Cement is assumed to cost $ 0.110 per kg in this example. These mixture optimizations were carried out using the ANN model.

compared to that of cement (N.B.: fly ash is assumed to cost 55% of cement in this example). However, this nature of compensation is unlikely as the cost of fly ash increases, and the cost of fly ash and cement achieve parity. This is even more complicated by the fact that often, constraints on fly ash contents in industrial concretes are imposed due to reduced rates of strength gain at early ages which results in reduced constructability; a project specific constraint that cannot be generically factored into approaches such as those demonstrated herein, i.e., without introducing a penalty factor that assesses the financial impact of reduced strength gain rates.

Because the cost of fly ash might vary between regions, and from season to season, it is important to ascertain a "critical" fly ash cost beyond which the inclusion of fly ash (whether Class C or Class F) in a concrete mixture is no longer an optimal choice – for cost minimization. To investigate these aspects, mixture optimization was performed across a range of possible fly ash costs. As such, Fig. 6 plots the optimal ratio of fly ash to total cementitious material as a function of the fly ash cost expressed as a percent of cement cost, for air-entrained concrete. Interestingly, for target strengths of 45 and 50 MPa, fly ash was still present in the optimal mixture even when it was 90% as expensive as cement. However, for a target strength of 55 MPa, fly ash was no longer included even when it was 75% as expensive as cement. This suggests that the inclusion of fly ash is a particularly effective cement dilution approach in conventional concrete mixtures, i.e., those which feature a compressive strength < 50 MPa.

4.2.2. *Strength-cost-embodied CO_2 optimization*

Another criterion that is expected to attain increasing prominence for concrete mixtures is their embodied CO_2 impact that is primarily attributed to its cement (OPC) content. In general, industrial by-products (IBPs) such as fly ash are not attributed any embodied CO_2 impact. The production of one metric ton of cement releases approximately 900 kg of CO_2 [37]. Therefore, the expected embodied CO_2 impact of a concrete mixture (i.e., in kg of CO_2 emitted per m^3 of concrete produced) can be expressed as,

$$CO_{2,embodied} = 0.9C_c \quad (13)$$

where, C_c is the cement content of the concrete formulation. Thus, the primary means of reducing embodied CO_2 for conventional concrete mixture is to substitute cement by fly ash (or another supplementary cementitious material, SCM, which in general have a lower embodied CO_2 impact than OPC). As shown in the previous section, substituting cement by fly ash can reduce the cost of the formulation. However, it is possible that in some cases, fly ash may be more expensive than cement. In this situation, there is a trade-off between minimizing cost and minimizing embodied CO_2. In such cases, the mixture optimization procedure can be modified to account for such complexities by introducing an additional constraint given by,

$$0.9C_c \leq CO_{2,max} \quad (14)$$

where $CO_{2,max}$ is the maximum embodied CO_2 (in kg/m^3). In this case, optimal mixtures are those that not only fulfill the imposed target strength but also the imposed maximum embodied CO_2 constraint.

Fig. 7 shows the Pareto front cost of optimal concrete mixtures as a function of both the target strength and maximum embodied CO_2 impact. As bounding cases, in Fig. 7(a), the price of fly ash was taken as one-half that of cement, whereas in Fig. 7(b) it was taken as 1.5 times that of cement. In Fig. 7(a), the curves overlap because both cost and embodied CO_2 can be simultaneously minimized by replacing as much cement as possible, by fly ash. This is because increasing the allowable embodied CO_2, and allowing for increased cement content, does nothing to reduce mixture cost. In contrast, Fig. 7(b) shows that increasing the embodied CO_2 impact – and thus requiring less fly ash – leads to a lower cost. Furthermore, there were some cases when the target strength and embodied CO_2 impact could not be achieved. For example, it was not possible to identify a mixture which featured a strength of 45 MPa but had an embodied CO_2 impact of < 250 kg of

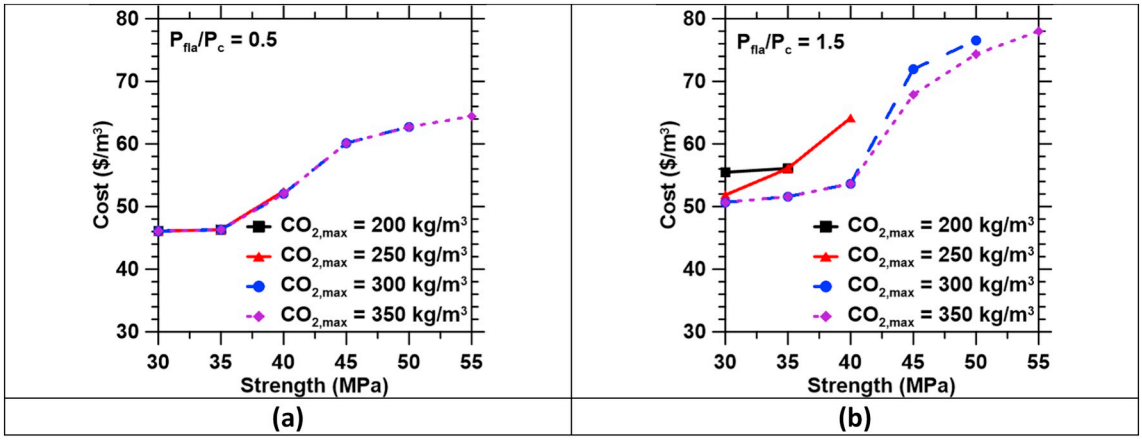

Fig. 7. The cost of optimized concrete mixes as a function of both their target strength and embodied CO_2 impact for different minimum cement contents. To set bounds on the analysis, in (a) the price of fly ash was set as one-half that of cement, while in (b) it was set as 1.5 times higher than cement. The mixture optimizations were carried out using the ANN model.

CO_2 per m^3 of concrete produced. This nature of analyses offers the means to rapidly screen, and eliminate unfeasible mixtures, and thereby constrain and restrict (validatory) trial batching activities only to those mixture formulations which are expected to feature the lowest cost, and embodied CO_2 impact.

5. Summary and outlook

This study has examined the use of statistical and machine learning methods to predict the compressive strength of concrete as a function of its mixture proportions, so as to consequently improve the practice of concrete mixture design, quality control, and quality assurance. A large dataset with over 10,000 measured compressive strengths was obtained from a vertically integrated cement/concrete producer (VIP) across a range of concrete production sites and used to train the predictive models. While the models have been shown to be very effective in predicting compressive strength of laboratory-produced concrete samples, they were less accurate when applied to the job-site data, possibly due to additional noise introduced by uncontrolled or unreported process variables, and variance within the formulation, proportioning, mixture and casting, and testing process. However, despite this, the models could still predict compressive strength with an average relative error of $< 10\%$.

It should be noted however that the VIP data which spans a substantial time-period of production includes diversity in: ambient weather conditions, the batches (and hence behavior) of cement and fly ash used, mixing action (truck or central-plant mixing) and aggregate composition and grading as may be expected for large volume production operations. This is only further complicated by differences in how water corrections are carried out and aggregate moisture content is measured which may be handled differently across different concrete production sites. These styles of differences may explain why the modeling approaches used herein are less effective at estimating job-site based concrete compressive strengths as compared to laboratory sourced data. This suggests a need to: (a) expand the size of the data sets used for training and testing the models, as doing so simultaneously improves the ability of models to distinguish meaningful patterns in the data from noise and allows for more refined hyperparameter tuning through cross-validation, and (b) incorporate a wider range of input variables for their influences on affecting concrete strength.

Finally, a mixture optimization procedure was demonstrated using an ANN based strength prediction model to identify mixtures that minimized the cost for a given target strength, while imposing limits on the estimated embodied CO_2 footprint of the mixture. These approaches offer a means to rapidly screen promising formulations for more intensive trial batching based evaluations – while reducing the labor and time intensity of concrete batching/trial operations. The outcomes of this paper are significant since they demonstrate a mathematical basis to estimate the strength of concrete – without a need for cement hydration, or microstructure models. Rather, the paper highlights that access to carefully curated large volumes of data, wherein the input variables are well-known (i.e., without needing to be carefully controlled) is a powerful means to apply big data analytics to rationalize, improve and accelerate concrete production operations: from the perspective of performance, quality control, and robustness – while reducing material (over)use, and wastage, and limiting overdesign. Each of these aspects are valuable to enable the design of cost-efficient and environmentally-friendly concrete mixtures for the construction of buildings and infrastructure which are often substantially overdesigned on account of poor predictability of in-place engineering performance.

Acknowledgements

The authors acknowledge financial support for this research provisioned by: Infravation ERA-NET grant (ECLIPS: 31109806.0001), and, U.S. National Science Foundation (CAREER: 1253269). The authors also acknowledge financial support provided by the Office of the Vice-Chancellor for Research at UCLA via the 'Sustainable L.A. Grand Challenge'. The contents of this paper reflect the views and opinions of the authors who are responsible for the accuracy of data presented. This research was carried out in the Laboratory for the Chemistry of Construction Materials (LC^2) and Molecular Instrumentation Center at UCLA. As such, the authors gratefully acknowledge the support that has made these laboratories and their operations possible.

References

[1] M.L. Wilson, S.H. Kosmatka, Design and control of concrete mixtures, Skokie, Ill.: Portland Cement Assn, 15 edition, 2011.
[2] S. Mindess, J.F. Young, D. Darwin, Concrete, 2nd edition, Pearson, Upper Saddle River, NJ, 2002.
[3] J.J. Thomas, et al., Modeling and simulation of cement hydration kinetics and microstructure development, Cem. Concr. Res. 41 (12) (Dec. 2011) 1257–1278.
[4] K.O. Kjellsen, R.J. Detwiler, Reaction kinetics of portland cement mortars hydrated at different temperatures, Cem. Concr. Res. 22 (1) (Jan. 1992) 112–120.
[5] J.W. Bullard, et al., Mechanisms of cement hydration, Cem. Concr. Res. 41 (12) (Dec. 2011) 1208–1223.
[6] D.P. Bentz, Three-dimensional computer simulation of portland cement hydration and microstructure development, J. Am. Ceram. Soc. 80 (1) (Jan. 1997) 3–21.
[7] Z.C. Grasley, D.A. Lange, Constitutive modeling of the aging viscoelastic properties of portland cement paste, Mech. Time-Depend. Mater. 11 (3–4) (Dec. 2007) 175–198.
[8] R. Alizadeh, J.J. Beaudoin, L. Raki, Viscoelastic nature of calcium silicate hydrate, Cem. Concr. Compos. 32 (5) (May 2010) 369–376.

[9] X. Li, Z.C. Grasley, J.W. Bullard, E.J. Garboczi, Computing the time evolution of the apparent viscoelastic/viscoplastic Poisson's ratio of hydrating cement paste, Cem. Concr. Compos. 56 (Feb. 2015) 121–133.
[10] I.-C. Yeh, Modeling of strength of high-performance concrete using artificial neural networks, Cem. Concr. Res. 28 (12) (Dec. 1998) 1797–1808.
[11] J.-S. Chou, C.-K. Chiu, M. Farfoura, I. Al-Taharwa, Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques, J. Comput. Civ. Eng. 25 (May 2011) 242–253.
[12] K. O. Akande, T. O. Owolabi, S. Twaha, S.O. Olatunji, Performance comparison of SVM and ANN in predicting compressive strength of concrete, IOSR J. Comput. Eng. 16 (5) (2014) 88–94.
[13] M.H.F. Zarandi, I.B. Türksen, J. Sobhani, A.A. Ramezanianpour, Fuzzy polynomial neural networks for approximation of the compressive strength of concrete, Appl. Soft Comput. 8 (1) (2008) 488–498.
[14] U. Atici, Prediction of the strength of mineral admixture concrete using multivariable regression analysis and an artificial neural network, Expert Syst. Appl. 38 (8) (2011) 9609–9618.
[15] J. Kasperkiewicz, J. Racz, A. Dubrawski, HPC strength prediction using artificial neural network, J. Comput. Civ. Eng. 9 (4) (1995) 279–284.
[16] H. Ni, J. Wang, Prediction of compressive strength of concrete by neural networks, Cem. Concr. Res. 30 (8) (2000) 1245–1250.
[17] A. Öztaş, M. Pala, E. Özbay, E. Kanca, N. Caglar, M.A. Bhatti, Predicting the compressive strength and slump of high strength concrete using neural network, Constr. Build. Mater. 20 (9) (2006) 769–775.
[18] M.H. Rafiei, W.H. Khushefati, R. Demirboga, H. Adeli, Supervised deep restricted Boltzmann machine for estimation of concrete, ACI Mater. J. 114 (2) (2017).
[19] I.B. Topcu, M. Sarıdemir, Prediction of compressive strength of concrete containing fly ash using artificial neural networks and fuzzy logic, Comput. Mater. Sci. 41 (3) (2008) 305–311.
[20] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer, New York, NY, 2013.
[21] Z.Q. John Lu, The elements of statistical learning: data mining, inference, and prediction, J. R. Stat. Soc. A. Stat. Soc. 173 (3) (Jul. 2010) 693–694.
[22] H.B. Demuth, M.H. Beale, O. De Jess, M.T. Hagan, Neural Network Design, 2nd ed., Martin Hagan, USA, 2014.
[23] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen (Eds.), Classification and Regression Trees, UK ed., Chapman and Hall/CRC, Boca Raton, 1984.
[24] M.C. Mozer, M.I. Jordan, T. Petsche, Advances in Neural Information Processing Systems 9: Proceedings of the 1996 Conference, MIT Press, 1997.
[25] The Nature of Statistical Learning Theory Vladimir Vapnik Springer.
[26] N.K. Nagwani, S.V. Deo, Estimating the concrete compressive strength using hard clustering and fuzzy clustering based regression techniques, Sci. World J. 2014 (Oct. 2014) e381549.
[27] K. Wang, J. Hu, Use of a moisture sensor for monitoring the effect of mixing procedure on uniformity of concrete mixtures, J. Adv. Concr. Technol. 3 (3) (2005) 371–383.
[28] Z. Bofang, Thermal Stresses and Temperature Control of Mass Concrete, Butterworth-Heinemann, 2013.
[29] G.H. Tattersall, Workability and Quality Control of Concrete, CRC Press, 2003.
[30] "Wiley: Practical Methods of Optimization, 2nd edition - R. Fletcher." ([Online]. Available: http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471494631.html. Accessed: 15-Sep-2017).
[31] M. Ahmaruzzaman, A review on the utilization of fly ash, Prog. Energy Combust. Sci. 36 (3) (Jun. 2010) 327–363.
[32] U.S. Geological Survey, Mineral Commodity Summaries, (2017).
[33] Ashgrove Cement Company, ASTM Class F Fly Ash Information Sheet, (1999).
[34] M. Collepardi, Admixtures used to enhance placing characteristics of concrete, Cem. Concr. Compos. 20 (2) (Jan. 1998) 103–112.
[35] B.A.S.F. Corporation, MasterGlenium 7500 Product Data Sheet, (2015).
[36] Portland Cement Association, Control of Air Content in Concrete, (Apr. 1998).
[37] N. Mahasenan, S. Smith, K. Humphreys, The cement industry and global climate change: current and potential future cement industry CO_2 emissions, in: J. Gale, Y. Kaya (Eds.), Greenhouse Gas Control Technologies - 6th International Conference, Pergamon, Oxford, 2003, pp. 995–1000.