Predictive Model for Construction Labour Productivity Using Hybrid Feature Selection and Principal Component Analysis

Sara EBRAHIMI*[1], Matin KAZEROONI*[2], Vuppuluri SUMATI, Ph.D.[3], and Aminah Robinson FAYEK, Ph.D., P.Eng., M.ASCE[4]

* These authors as co-first authors contributed equally to this work. Author ordering determined by agreement.

[1] M.Sc. Student and Graduate Research Assistant, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, AB, Canada, email: eb4@ualberta.ca

[2] M.Sc. Student and Graduate Research Assistant, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, AB, Canada, email: kazeroon@ualberta.ca

[3] Postdoctoral Fellow, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, AB, Canada, email: sumati.vuppuluri@ualberta.ca

[4] Tier 1 Canada Research Chair in Fuzzy Hybrid Decision Support Systems for Construction, NSERC Industrial Research Chair in Strategic Construction Modeling and Delivery, Professor, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, AB, Canada, email: aminah.robinson@ualberta.ca. (corresponding author)

Abstract: Construction labour productivity (CLP) is affected by numerous variables made up of subjective and objective factors. Thus, CLP modeling and prediction is a complex task, leading to high computational cost and the risk of overfitting of data. This paper proposes a predictive model for CLP by integrating hybrid feature selection (HFS), as a combination of filter and wrapper methods, with principal component analysis (PCA). This developed HFS-PCA method reduces the dimensionality and complexity of CLP data and obtains better prediction performance by identifying the most predictive factors. Identified factors are utilized as inputs for various classification methods to predict CLP. Finally, prediction error of the classification methods with and without using the proposed HFS-PCA method are compared, and the most accurate classification method is selected to develop the CLP predictive model. Experimental results show that using HFS-PCA for CLP prediction leads to better performances compared with past studies.
Keywords: Construction labour productivity prediction, hybrid feature selection, principal component analysis, genetic algorithm, support vector machine, ReliefF algorithm.

1. Introduction

As the construction industry accounts for the highest share of employment and labour costs comprise the majority of overall project cost in many countries (Heravi and Eslamdoost 2015), understanding construction labour productivity (CLP) as accurately as possible is key to improving project performance and directly affects construction companies' competitiveness and profitability. Therefore, a reasonably accurate predictive model of CLP is required to help organizations understand which factors most impact CLP (Moselhi and Khan 2012). In this study, CLP is defined as the ratio of units of output, expressed as installed quantity (in cubic meters), to units of input, expressed as total labour work-hours, as shown in Equation (1). The goal of the CLP system is to obtain higher CLP values.

$$CLP = \frac{Output\ (Installed\ quantity)}{Input\ (labor\ work-hours)} \tag{1}$$

The CLP environment is unpredictable and complex because a large number of parameters influence CLP directly or indirectly, and the process of tracking CLP is time consuming (Tsehayae and Fayek 2016). Various studies have identified numerous objective and subjective factors influencing CLP. These studies used questionnaire surveys to identify top factors influencing CLP (Dai and Goodrum 2012; Jarkas 2015; Montaser et al. 2018; Alaghbari et al. 2019; Kazerooni et al. 2020). While many studies focused on identifying CLP factors, fewer studies are found in the literature on predicting labour productivity (Agrawal and Halder 2020). Studies on predicting and modeling CLP can be classified as either statistical or artificial intelligence (AI) techniques (Golnaraghi et al. 2020). The most common statistical technique is regression analysis. Thomas and Sudhakumar (2014) developed several linear regression models to determine the effect of 11 influential factors on masonry labour productivity. Mohsenijam and Lu (2019) proposed a data-driven approach using multiple linear regression to select the most predictive project design factors affecting labour hours. However, regression models are limited by the number of influencing parameters and their capability of determining the combined impact of the influencing parameters (Song and AbouRizk 2008). Artificial neural network (ANN) methods are the most common AI techniques, and their capability to learn from experience to improve their performance and adapt themselves to changes make them useful methods for prediction (Mirahadi and Zayed 2016). Song and AbouRizk (2008) presented a CLP model based on ANN and discrete-event simulation by analyzing the historical data.

Notably, high-dimensional data may present different problems, such as reduced accuracy and increased complexity (Heravi and Eslamdoost 2015). CLP is set in a high-dimensional feature space where it is affected by numerous factors. Thus, CLP prediction imposes a high computational cost and the risk of overfitting. To address these two issues, it is necessary to reduce the dimensionality of CLP data and determine the factors most predictive of CLP. This can be done using dimensionality reduction methods, which are categorized into feature selection and feature extraction methods.

In any data mining process, feature selection (FS) is vital for reducing the number of features, removing redundant data, and identifying a relevant subset for prediction (Cao et al. 2018). FS methods are categorized into three primary groups: filter methods, wrapper methods, and hybrid methods. Filter methods offer less computational time to provide results and do not require a learning algorithm; rather, they rank and select features based on statistical measures such as correlation. Their main disadvantage is that they do not consider model prediction and feature interaction. Most filter methods are suitable only for developing mathematical equations by the statistical regression methods (Ghosh et al. 2019). Wrapper methods use the model prediction of a machine learning algorithm to determine the set of most suitable features. Thus, they are tuned to the specific interaction between a learning algorithm and its training data. However, their applications are limited by the high computational complexity when feature sets are wide (Piao and Ryu 2017).

A feature extraction (FE) method, such as principal component analysis (PCA), is used to transform the inputs onto a low-dimensional subspace, which preserves the majority of relevant information. According to Kavitha and Kannan (2016), FE methods are mainly grouped into two categories: (1) projection methods such as PCA and linear discriminate analysis for unsupervised learning, and (2) compression methods such as mutual information and information theory for supervised learning. PCA is a broadly used dimensionality reduction method that reduces computational complexity, distractive noise, and the risk of overfitting, with minimal loss of information when applied to correlated features (Salo et al. 2019). PCA identifies patterns in the dataset and preserves the most significant relationships between the features by calculating the eigenvalues and eigenvectors of the dataset's covariance matrix.

Hybrid feature selection (HFS) methods help resolve the problem of high computational complexity by merging a wrapper method with a suitable filter method to reduce deficiencies of both methods and thus is generally more efficient than single filter or wrapper methods. The general HFS approach has two stages. First, a filter method refines and selects the top-*n* features, then a wrapper method identifies the most discriminative subset from the top-*n* features (Ghosh et al. 2019).

The high degree of correlation between CLP factors is another challenge in predicting CLP, in addition to the generally complexity of construction projects. Thus, reducing the degree of correlation among the CLP factors and identifying key factors that significantly impact CLP is essential to predicting it with any reasonable accuracy. According to previous studies in other domains, using HFS and PCA enhanced accuracy of prediction and modeling (Piao and Ryu 2017; Salo et al. 2019). FS methods also provide a subset of original factors that can lead to identification of key CLP factors for improving CLP prediction. Notably, very few studies in labour productivity prediction used FS and FE to reduce dimensionality of CLP data and identify the most predictive factors affecting CLP. The main goal of this study was to develop a novel approach for predicting and modeling CLP. This goal was supported by (1) developing a novel approach using HFS-PCA for feature selection and extraction to select factors with the most influence on CLP, (2) developing an improved model for predicting and modeling CLP, and (3) ranking the factors most predictive of CLP. Thus, the major contribution of this paper is the presentation of a novel predictive model for CLP that integrates HFS and PCA as a hybrid method for determining the most predictive factors of CLP and reducing data dimensionality, computational time, and model complexity.

The rest of paper is organized as follows. Section 2 provides a review of past research on CLP modeling and FS and FE methods. Section 3 describes the proposed methodology. Section 4 presents the experimental results from using the proposed model to predict CLP, a summary of the results, and comparison of different classification methods. Section 5 offers conclusions and notes regarding future work.

2. Literature Review

2.1. Feature selection and extraction methods

Several studies combined PCA with FS techniques in order to (1) increase the advantages of both methods for providing improved classification performance with a minimum number of relevant and non-redundant features instead of using all affecting features and (2) present a hybrid method. Jain and Singh (2018) presented a new method consisting of ReliefF as a filter method and PCA for dimensionality reduction. Sahu et al. (2018) proposed a prediction model for breast cancer classification and diagnosis by integrating PCA and ANN as a hybrid approach. Abo El-Maaty and Wassal (2019) proposed a hybrid GA-PCA methodology in which GA was used as a FS wrapper technique to select a subset of n features from 561 features, and PCA was then utilized to reduce the subset into k orthogonal features. Salo et al. (2019) used PCA integrated with information gain (IG) as a filter method to decrease the search range in a predictive model for network intrusion detection. Mohammed and Ahmed (2019) developed a combined analysis of variance (ANOVA) and PCA technique on a dataset of 41 features. Correlation matrix technique was computed to show high correlation of the selected features. Thus, PCA was applied to transform and reduce data to a lower number of uncorrelated features. According to the literature, no study integrated ReliefF and support vector machine–genetic algorithm (SVM-GA) as an HFS method with PCA.

2.2. Identification of key factors influencing CLP

CLP is affected by numerous objective (e.g., crew size, crew average years of experience) and subjective (e.g., crew motivation, complexity of task) factors. Most previous studies used questionnaire surveys to identify top factors influencing CLP (Tsehayae and Fayek 2014; Jarkas 2015; Durdyev et al. 2018; Montaser et al. 2018; Alaghbari et al. 2019; Irfan et al. 2020; Agrawal and Halder 2020). Several studies identified top factors influencing CLP using statistical analyses such as relative importance index (RII), mean response (MR), and frequency index. Hafez (2014) used a questionnaire survey comprising 27 productivity factors and used RII to rank them and identify the most influential factors. Chigara and Moyo (2014) used a questionnaire that included 40 preselected CLP factors, which were ranked using RII and MR. Alaghbari et al. (2019) used a questionnaire comprising 52 predefined factors and used RII to identify the factors most influencing CLP from the perspective of structural engineers. A limitation to using questionnaire surveys, however, is that the selected factors highly rely on expert knowledge, which can be very changeable over time and between projects. Another limitation of evaluation indices such as RII is their lack of capability to consider interconnections among CLP factors. Several studies have attempted to identify the relative importance of CLP factors through the use of a data-driven approach such as feature selection (Moselhi and Khan 2012). Data-driven approaches are not dependent on expert knowledge and consider the dynamics of CLP factors and the interconnected relationships among them (Ebrahimi et al. 2021). Various studies in labour productivity used filter FS methods to identify top CLP factors. Tsehayae and Fayek (2016) used a correlation-based feature selection (CFS) filter method to identify key features influencing CLP. CFS is appropriate because of its capability to deal with a high-dimensional feature space. However, wrapper or HFS methods are more appropriate for predictive modeling because they use AI techniques, such as fuzzy inference system (FIS), ANN, and SVM to train predictive models (Piao and Ryu 2017). Several studies showed that using a wrapper or HFS method in the application, where the predictive model is developed, shows better results for accuracy (Ahmad and Pedrycz 2012; Gerami Seresht et al. 2020).

2.3. CLP modeling using AI techniques

Since many activities in the construction industry are labour dependent, numerous studies have focused on predicting and modeling labour productivity. More recently, most proposed productivity prediction models used AI techniques to increase prediction accuracy (Cheng et al. 2020). El-Gohary et al. (2017) used ANN and hyperbolic tangent as a transfer function to quantify and map the relationship between CLP and the relevant influencing factors. Their results showed an adequate convergence and more accurate and credible results compared with previous approaches. Khanzadi et al. (2017) proposed a hybrid simulation model of system dynamics and agent-based modeling to predict and improve CLP, which accounted for CLP factors with continuous behavior and the interaction between different agents involved in the project. Ghazi Al-Kofahi et al. (2021) developed a system dynamics model to investigate the impact of change orders on CLP and identify the causes of productivity loss. Golnaraghi et al. (2020) modeled expected CLP by using several ANN techniques, such as backpropagation neural network and adaptive neuro-fuzzy inference system (ANFIS), and compared their respective results to determine the best method for estimating expected labour productivity. Mirahadi and Zayed (2016) proposed a hybrid intelligent model using neural network–driven fuzzy reasoning to improve the accuracy of productivity prediction. Gerami Seresht and Fayek (2018) developed a predictive model of multifactor construction productivity using fuzzy system dynamics to address the subjective factors influencing productivity. Raoufi and Fayek (2018) integrated fuzzy logic and agent-based modeling to predict the performance of construction crews based on crew motivational and situational input variables. Nasirzadeh et al. (2020) proposed ANN-based prediction intervals as a method for forecasting CLP using historical data. Their model accounted for various sources of uncertainty affecting prediction. While these previous studies demonstrated the usefulness of using AI techniques in CLP prediction, the numerous objective and subjective factors affecting CLP provide a large number of inputs that may reduce the accuracy and increase the complexity of productivity prediction (Ebrahimi et al. 2020). Therefore, data mining approaches such as FS and FE, which reduce data dimensionality, computational time, and model complexity, can be used to increase the reasonable accuracy of predictive models for CLP.

3. Methodology

This paper presents a model that identifies the most predictive CLP factors and predicts CLP in a high-dimensional feature space where numerous factors affect CLP with the greatest accuracy possible. Figure 1 shows a general view of the proposed methodology, which includes two main phases: data preparation and data analysis. In the data preparation phase, the raw data is transformed into a form that can accurately be analyzed. In the data analysis phase, HFS-PCA is applied for analysis of the prepared dataset. The following sections are an overview of the CLP dataset used in this study and the stages of processing the CLP data.

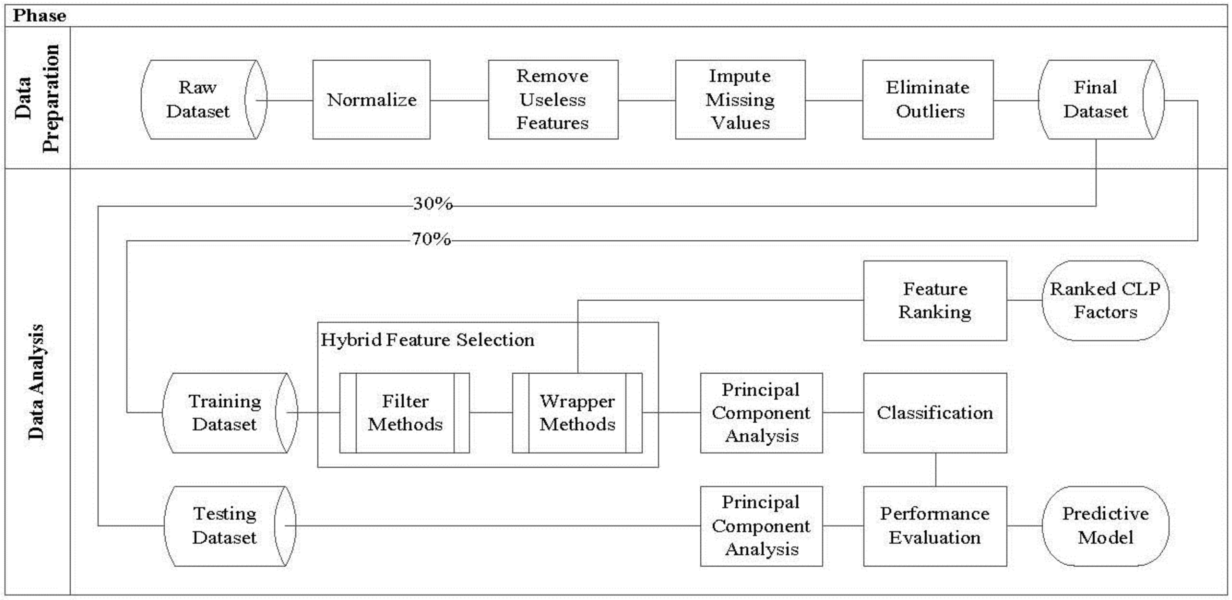

Fig. 1. A general view of the proposed predictive model for construction labor productivity.

3.1. CLP dataset overview

In this study, the proposed predictive model was developed for predicting the CLP of concrete placing activities using empirical data collected in three data collection cycles between June 2012 and October 2014 in collaboration with two partnering companies, in Alberta, Canada, in the context of four construction projects: industrial buildings, residential and commercial high-rise buildings, residential and commercial warehouse buildings, and institutional buildings (see Tsehayae and Fayek 2014; Tsehayae and Fayek 2016). The data were collected by documenting the value of CLP factors and CLP on a daily basis at the construction site. As a result, a total of 112 factors influencing CLP were identified and measured over 92 days. Therefore, the utilized CLP dataset in this study consists of 112 factors and 92 data points for each factor. All factors in the dataset are listed in Table S2 [see Supplementary Materials]. Due to the nature of the data collected, the factors addressed in this study focus on material-related and management-related factors affecting CLP. The effects of buildability factors (e.g., volume placed, concrete workability, rebar congestion) or other types of factors that may affect CLP are not addressed in the current paper.

3.2. Phase 1: Data preparation

Data preparation is the initial stage of processing data, with the goal of manipulating the raw data into a form that can accurately be analyzed. A CLP dataset is prepared as a raw dataset and transformed to a more informative form per the following data preparation stages, in order to make CLP data modeling and analysis more efficient.

3.2.1. Normalization

By adjusting the value range, normalization can lead to stable convergence and prevent biases in predictive models (Golnaraghi et al. 2020). The normal distribution, which subtracts the mean of the data from all values and divides them by the standard deviation, helps preserve the original distribution of the data (Frigerio et al. 2019). Thus, normalization with respect to normal distribution is used in the developed model to scale CLP data into an organized range.

3.2.2. Remove factors with zero standard deviation

Standard deviation as the square root of the variance is a measure of how spread out the values of each feature are in the dataset. ReliefF as a filter method uses correlation among features to filter the factors. If the standard deviation of a feature's data points equals zero, ReliefF is not capable to determine the existing correlations among the features. Accordingly, the features with zero standard deviation should be removed (Peker et al. 2020). In this study, 8 CLP factors had standard deviation equal to zero and consequently were removed from the CLP dataset. Thus, the total number of CLP factors was reduced to 110.

3.2.3. Impute missing values

Imputation is a technique of estimating the missing values of a dataset by applying various machine learning algorithms. Imputation methods based on K-nearest neighbors (KNN) use classification capacity to identify a subset of data points having the most similarity to the data points with missing values (Ma and Zhong 2016). Hence, in the presented model a KNN-based imputation method is utilized to impute missing values of the CLP dataset.

3.2.4. Eliminate outliers

Outliers in a dataset can significantly affect the performance of data analysis. The Tukey Test method is a commonly used outlier detector, in which a confidence interval is defined for each feature by utilizes the median, upper, and lower quartiles of a data set. Since quartiles are resistant to farthest data of the data set, Tukey's method is less sensitive compared to methods using mean and standard variance (Sandbhor and Chaphalkar 2019). In this study, after applying the Tukey Test method to the CLP dataset, 10 observations were identified as outliers. Hence, the total number of data points for each factor in the CLP dataset was reduced to 82.

3.3. Phase 2: Data analysis

The second phase of developing a model for CLP prediction is analyzing the final CLP dataset resulting from phase 1. First, the final CLP dataset is randomly divided into two subsets named Training Dataset and Testing Dataset. Of the final CLP dataset, 70 percent (in this study 69.5) is used for selecting the most predictive CLP factors and developing various classification models. The remaining 30 percent (in this study 30.5) of the final CLP dataset is used for estimating and comparing the performance of employed classifiers based on various performance measures. The dimensionality of the final CLP dataset was significantly high since it had 110 input features. Predicting CLP based on this dataset would thus lead to high computational complexity and low accuracy. Therefore, prior to predicting CLP, a new dimensionality reduction method was introduced by integrating HFS methods with PCA. HFS-PCA is used for identifying the most predictive CLP factors, reduce the feature space and computational complexity, and thus enhance the predictive model's performance. The following subsections explain the preliminary concepts used in HFS-PCA and describe the stages of CLP feature reduction and CLP prediction.

3.3.1. Preliminaries

The main concepts used in the proposed methodology's data analysis phase are as follows.

3.3.1.1. ReliefF algorithm (RFA)

The Relief algorithm as an individual evaluation filtering FS method assigns weights to each feature based on correlation between features and selects all features with greater weight compared with the threshold. Although Relief is an efficient method with reasonably accurate results, an important limitation of this algorithm is that it can handle only two-class classification problems. To manage this limitation and handle multiclass problems, Kononenko (1994) proposed ReliefF algorithm (RFA). Equation (2), which is the ReliefF function (RFF), shows the evaluation criteria of RFA, where *n* is the total number of features, *D* is distance measurement, $f_{t,j}$ is the value of instance x_j on feature f_j, and $f_{s(x_j)}$ and $f_{d(x_j)}$ denote the value of jth feature of the nearest point to x_j in the same and different class, respectively.

$$RFF(f_j) = 0.5 \sum_{j=1}^{n} \left(D\left(f_{t,j} - f_{s(x_j)}\right) - D\left(f_{t,j} - f_{d(x_j)}\right) \right) \tag{2}$$

3.3.1.2. Support vector machine (SVM)

An SVM is a supervised learning model that can solve two-class binary classification problems. SVMs are used for classification and regression analysis. The learning algorithm of SVM is based on statistical learning theory and structural risk minimization. Theoretically, SVMs experience less overfitting and better generalization than traditional techniques, such as ANN. The main approach of SVM is using the maximum margins between support vectors to build an optimal hyperplane. SVM shows great generalization performance, which represents the desired accuracy in classification and prediction of unseen samples (Fernández-Delgado et al. 2014). SVM is used for solving linear and non-linear problems. For non-linear classification, the mapping function is utilized to convert low-dimensional data to a high-dimensional dataset, which changes the non-linear problem to a linear and separable problem. Kernel functions are employed to make this process easier. There are various types of kernel function, namely, linear, polynomial, sigmoid, and Gaussian function. Gaussian function, presented in Equation (3), is the most common kernel function for solving classification problems, as it requires just one parameter, γ, which is a free parameter and has a significant influence on classification accuracy (Pai et al. 2021). Another important parameter in SVM is penalty factor C, which is the cost of misclassification. Based on the importance of these two parameters on the result of SVM, C and γ needed to be optimized for achieving the desired accuracy, which is accomplished by GA.

$$K(x, x') = \exp(-\gamma \parallel x - x' \parallel^2) \tag{3}$$

3.3.1.3. Genetic algorithm (GA) optimization

GA is a stochastic searching process based on the mechanism of natural selection and natural genetics, thus imitating the process of natural evolution. GA is a good approach to exploring feature space and can produce many alternative feature subsets through reproduction operations to obtain the best subset that includes the most predictive features. GA uses a fitness function to evaluate each candidate solution's fitness. The crossover and mutation functions randomly transfer chromosomes as two major operators with key impact on the fitness value. Crossover is a randomizing mechanism that exchanges features between two chromosomes using single-point, two-point, or homologue crossover (RazaviAlavi and AbouRizk 2017).

The three criteria for designing a fitness function are the number of selected features, classification accuracy, and cost. Based on these criteria, a chromosome with a small number of selected features, high classification accuracy, and low cost can produce a high fitness value. The GA optimization method maximizes the value of the fitness function, shown in Equation (4) where SVM_Error is a root mean square error (RMSE) of SVM classifier, W_f is a weight value for the number of features (n_f), f_i represents '1' if the feature i is selected or '0' if the feature i is not selected, and c_i is cost of feature i.

$$Fitness = (SVM\_Error \times (1 + W_f \times (\sum_{i=1}^{n_f} c_i \times f_i)))^{-1} \quad (4)$$

To achieve better performance, GA-based selected features are used as the inputs for PCA.

3.3.1.4. Principal component analysis (PCA)

PCA applies linear transformation on the original n features to convert them into a space where the new k features are linear independent. These k features are called principal components, which have three major properties (Faisal Elrawy et al. 2013): (1) the principal components are uncorrelated, (2) the first principal component (PC1) has the highest variance and each principal component that follows it covers the lesser value of variance, and (3) the total variance of the principal components is equal to the total variance of the original features. To be more specific, let X be the dataset with n features and m instances as shown in Equation (5), where X_i denotes the i_{th} feature as shown in Equation (6):

$$X = [X_1 \quad X_2 \quad \cdots \quad X_i \quad \cdots \quad X_n] \quad (5)$$

$$X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{im} \end{bmatrix} \quad (6)$$

where x_{ij} signifies the j_{th} instance of the i_{th} feature. Considering the above matrices, the steps of the developed PCA method are defined as:

Step 1 – Normalize the data to produce a dataset with zero mean.

Step 2 – Calculate the covariance matrix as follows:

$$CM = \begin{bmatrix} Cov[X_1, X_1] & Cov[X_1, X_2] & \cdots & Cov[X_1, X_n] \\ Cov[X_2, X_1] & Cov[X_2, X_2] & \cdots & Cov[X_2, X_n] \\ \vdots & \vdots & \vdots & \vdots \\ Cov[X_n, X_1] & Cov[X_n, X_2] & \cdots & Cov[X_n, X_n] \end{bmatrix}_{n\times n} \tag{7}$$

where $Cov[X_i, X_j]$ is the covariance between features X_i and X_j, which is computed as:

$$Cov\left[X_i, X_j\right] = \left(\frac{1}{m-1}\right) \sum_{z=1}^{m} (x_{iz} x_{jz}) \tag{8}$$

Step 3 – Extract the eigenvectors and eigenvalues from the covariance matrix using the following equations:

$$\det(\lambda_i [I]_{n\times n} - CM) = 0 \tag{9}$$

$$CM[v_i]_{n\times 1} = \lambda_i [v_i]_{n\times 1} \tag{10}$$

where *CM* denotes the covariance matrix, I is the identity matrix, "det" is the determinant of the matrix, λ_i signifies the i_{th} eigenvalue of the covariance matrix, and v_i is the corresponding eigenvector. The greater the eigenvalue, the more significant its corresponding eigenvector. Thus, by considering $\lambda_1 > \lambda_2 > \cdots > \lambda_n$, the principal components will be sorted in descending order in terms of significance.

Step 4 – Select the first k eigenvectors that correspond to the first k eigenvalues, and build the projection matrix V as follows:

$$V = [v_1 \quad v_2 \quad \dots \quad v_k]_{n\times k} \tag{11}$$

Since the eigenvalues of the ignored $n - k$ eigenvectors are small, the loss of information of the original dataset will be minimal. In order to determine k, it is suggested that the chosen number of principal components contain about 50 to 70 percent of the total variation of the original features (Faisal Elrawy et al. 2013).

Step 5 – Form the new dataset Y by transforming the original dataset X via V:

$$Y = X * V \tag{12}$$

where Y is the new dataset of k uncorrelated principal components and m instances.

3.3.2. HFS-PCA

An overview of the proposed HFS-PCA method is shown in Figure 2, which presents the process of integrating RFA as a filter method, GA and SVM as the wrapper method, and PCA as the FE method. HFS-PCA is used for selecting the most predictive factors affecting CLP and reducing their dimensionality in order to predict CLP more accurately.

As Figure 2 shows, detailed steps for developing the HFS-PCA method for the CLP dataset are as follows.

Step 1 – The RFA filter method evaluates the weight of each CLP factor according to the correlations between the factors and ranks them in terms of their weights. After the RFA process is complete, factor weights (w_r) are normalized from 0 to 1 to make the wrapper process more effective; by using a defined threshold (τ) in the range 0–1, any factors with a weight $w_r \geq \tau$ are selected.

Step 2 – GA generates the random initial population of chromosomes. Each chromosome in the population represents an available solution to the factor subset selection problem.

Step 3 –Selected factors that have weights greater than the threshold are the inputs of SVM.

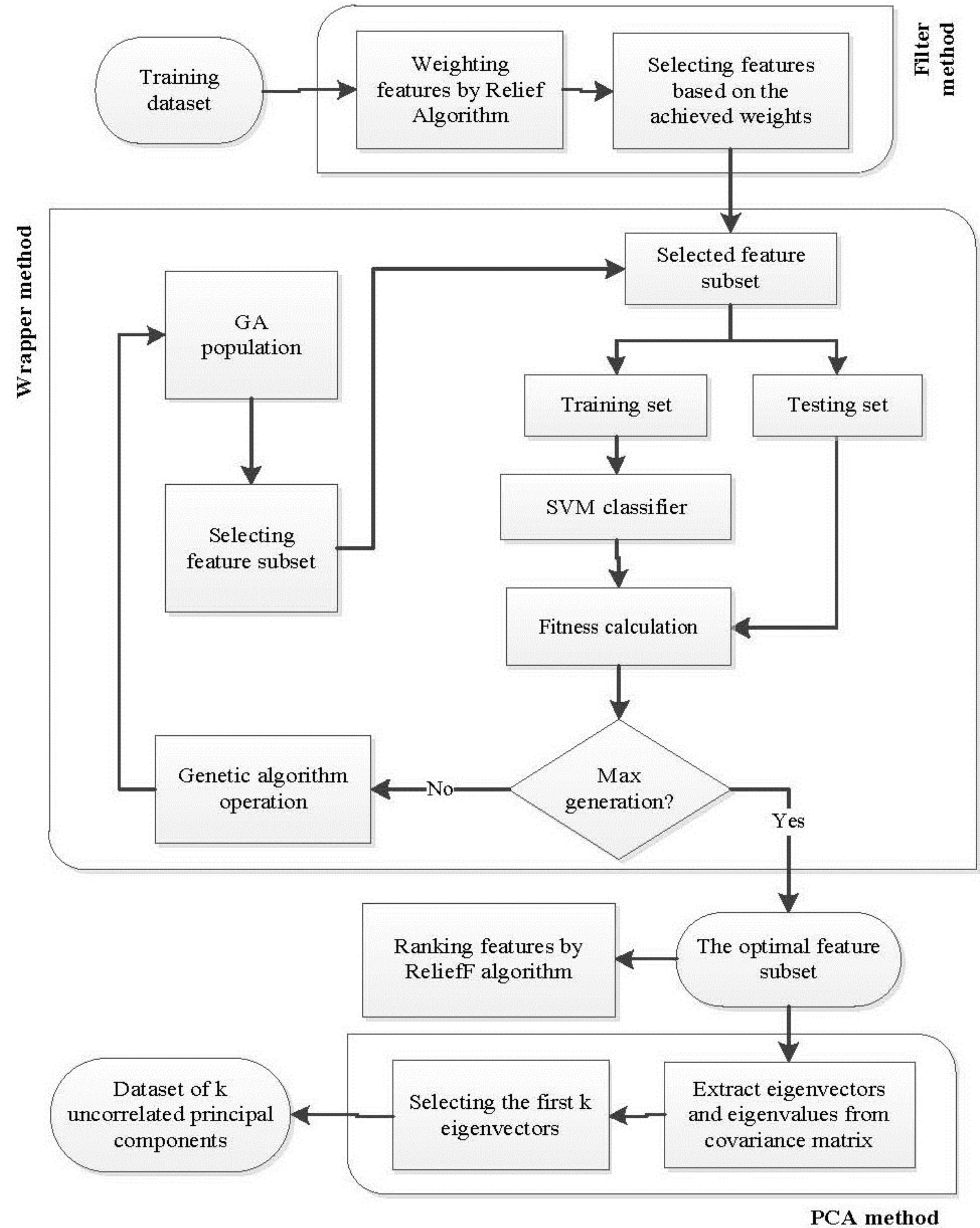

Fig. 2. Overview of the HFS-PCA method.

Step 4 – The training set and testing set are built from the selected CLP factor dataset. Then, using the training set, the process of training SVM begins, while the testing set is utilized to calculate the SVM error.

Step 5 – The fitness calculation process is completed using the calculated RMSE for SVM classification, based on Equation (13).

$$SVM\_Error = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - T_i)^2} \tag{13}$$

where n is the number of outputs, A_i is the actual output value of the ith output, and T_i is the target output value of the ith output. In this paper, there is one output, which is CLP. Note that a better fitness of the SVM requires a smaller error.

Step 6 – If termination criteria are satisfied, the process ends; otherwise, the process goes to the next generation by GA.

Step 7 – GA searches for better solutions by using crossover, mutation, elitism, and replacement. In this study, single-point binary crossover and binary mutation were performed. Also, per the elitism process the three best chromosomes are selected to be part of the population in the next generation.

Once the final generation meets termination criteria, the iteration stops, and the selected subset of factors is the one that has the best predictor of CLP among all subsets of factors. The termination criteria are: either the generation number reaches a determined value, or the fitness value does not improve during a specified number of generations. For this study, maximum generation was 150 and specified number of generations was 50.

Step 8 – Before employing PCA, RFA is used one more time to rank the selected factors and adjust factors' weights.

The CLP factors selected by HFS can be used directly for predicting CLP. However, considerable correlation among the factors will affect the performance of the predictive model. Furthermore, when managing a high-dimensional dataset such as a CLP dataset, the dimensionality reduction performed by HFS may not be enough to reduce computational complexity in classifiers. To alleviate these drawbacks, the factors selected by HFS are exposed for further reduction by applying the PCA method in the following steps.

Step 9 – The covariance matrix of the selected factors is calculated, then the eigenvectors and eigenvalues are extracted from it.

Step 10 – Since the eigenvectors are set in descending order in terms of significance, the first k eigenvectors are selected as the last step of HFS-PCA for forming a new dataset of k uncorrelated principal components, based on Equation (12).

The new dataset is then utilized in the following classification stage to develop various classifiers for CLP prediction.

3.3.3. Classification

After using the proposed HFS-PCA method to reduce CLP dataset dimensionality and thus identify the most predictive CLP factors using the proposed HFS-PCA method, four classifiers – KNN, ANN, random forest (RF), and ANFIS – are employed for CLP prediction and performance analysis. In order to avoid overfitting and manage the possible variations of input data, ten-fold cross validation is used for developing the classification models by partitioning the data into 10 random subsets. One subset is utilized to validate the model trained by the remaining subsets. This procedure is repeated 10 times such that each subset is used once for validation.

3.3.4. Performance evaluation

After development of the classification models, they are applied to the Testing Dataset for performance evaluation. The efficiency of the models is compared with three performance measures capable of managing numerical attributes such as CLP: (1) RMSE, which provides the same dimensions as the predicted value itself (Equation 14); (2) mean-absolute error (MAE), which is the average deviation of the predictions from the actual values without taking into account their sign (Equation 15); and (3) correlation coefficient, which is a statistical measurement that computes the correlation between the actual and predicted values (Equation 16). The correlation coefficient ranges from –1 to 1, where 0 signifies no correlation and –1 and 1 denote the highest negative and positive correlations, respectively. Higher correlation means better model performance. Unlike RMSE, MAE does not exaggerate the effect of instances whose prediction errors are larger than the other instances (Witten et al. 2017).

$$RMSE = \sqrt{\textstyle\sum_i (p_i - a_i)^2 / m} \tag{14}$$

$$MAE = \left(\textstyle\sum_i |p_i - a_i|\right)/m \tag{15}$$

$$Correlation\ coefficient = \frac{\sum_i (p_i - \bar{p})(a_i - \bar{a})/(m-1)}{\sqrt{\left(\sum_i (p_i - \bar{p})^2/(m-1)\right)\left(\sum_i (a_i - \bar{a})^2/(m-1)\right)}} \tag{16}$$

where a_i is the actual value and p_i is the predicted value for the ith instance, m is the number of instances, and $\bar{a}$ and $\bar{p}$ are the mean value over the actual and the predicted test data, respectively.

4. Experimental Results and Discussion

In the data preparation phase of this study, the raw CLP dataset initially consisted of 112 factors and 92 data points for each factor. After data preparation phase, the final CLP dataset consisted of 110 factors and 82 normalized data points for each factor. The final CLP dataset with 110 factors still had too many factors for building a successful predictive model, because it would lead to high computational complexity and low prediction accuracy. The HFS-PCA method was implemented to reduce the dimensionality of CLP factors. This section illustrates the implementation of HFS-PCA in CLP prediction and presents an evaluation of the performance of various predictive models.

4.1. HFS-PCA results

For this study, factors that satisfied the threshold of 0.2 in Equation (2) were selected as key factors for the next stage of HFS. Of the 110 factors in the final CLP dataset, RFA selected 35 as key factors.

As noted in section 3.3.2, step 7, termination criteria for the GA-SVM method applied in this study were: a maximum generation of 150, or no improvement of the fitness value during the last 50 generations. SVM parameters C and σ were both set to 20, kernel type was radial, and kernel cache was 200. The parameter settings for GA were: population size of 100, crossover rate of 0.7, mutation rate of 0.02, one-point crossover, and tournament selection scheme. To reduce bias selection of the optimal subset of factors, 15 different local seeds were examined in order to identify the best possible subset of CLP factors. Considering these parameters, the proposed wrapper FS was developed, which selected 19 factors out of the 35 CLP factors specified by RFA. Table 1 shows RFA ranking of the 19 factors selected as the most predictive CLP factors.

Table 1. RFA ranking of the most predictive CLP factors

Factor index	CLP Factor	Normalized importance	RFA Rank
2	Fairness of work assignment	1.000	1
6	Complexity of task	0.793	2
7	Repetitiveness of task	0.706	3
16	Owner staff on site	0.568	4
10	Congestion of work area	0.535	5
19	Structural element	0.527	6
18	Concrete placement technique	0.476	7
1	Team spirit of crew	0.295	8
13	Weather (precipitation)	0.233	9
3	Crew participation in foreman's decision-making process	0.231	10
9	Location of work scope (distance)	0.229	11
5	Material movement practices (horizontal)	0.217	12
17	Availability of labour	0.199	13
12	Weather (temperature)	0.172	14
14	Variability of weather	0.168	15
4	Job security	0.071	16
8	Working conditions (dust and fumes)	0.041	17
15	Ground conditions	0.002	18
11	Cleanliness of work area	0.000	19

The CLP factors selected by HFS were used directly for classification purposes. However, some selected factors are highly correlated, which can affect the predictive model's performance. Figure S1 [see Supplementary Materials] shows the correlation matrix image of the 19 selected CLP factors in this study. Furthermore, despite the dimensionality reduction performed by HFS, the number of selected factors would still lead to computational complexity in classifiers such as ANFIS. To mitigate these limitations, PCA was applied to the selected CLP factors to reduce their dimensionality, as the final step of HFS-PCA.

PCA applied orthogonal transformation on the original 19 factors to convert them into k principal components that are linearly uncorrelated, without losing much information. The principal components were sorted from the highest variance to the lowest; hence, PC1 covered the highest variance, and other principal components covered the lesser values of variance. Table S1 [see Supplementary Materials] presents the variance between all the eigenvectors obtained from the covariance matrix of the 19 factors, using equations (10) and (11). All eigenvectors are linearly uncorrelated because their variances equal zero (see Table S1). Faisal Elrawy et al. (2013) recommended that chosen k principal components contain about 50–70% of total variation of the original factors and showed that having two principal components is better than having three. The first two, three, and four eigenvectors share 59%, 70%, and 75% of total variation of the selected 19 factors, respectively. Thus, 2 was chosen as the value of k, and the transformed dataset consisted of two uncorrelated principal components, PC1 and PC2 (second principal component).

4.2. Performance evaluation results

To evaluate the performance of the developed HFS-PCA method for predicting CLP, two parameters need to be satisfied: (1) the efficiency of the predictive model when HFS-PCA is applied compared with the models in which it is not employed, and (2) the efficiency of HFS-PCA when different classifiers are used for predicting CLP. To satisfy both parameters, several CLP predictive models were developed in four categories, as shown in Table 2. The same classifier was used in each category. The classifiers utilized in categories 1, 2, 3, and 4 were RF, ANN, KNN, and ANFIS, respectively. The symbols ✓ and × represent the use or non-use of a method in a given predictive model. In each category, three combinations of HFS and PCA were tested:

1. HFS-PCA along with the classifier—The two principal components (PC1 and PC2) were used for predicting CLP.
2. HFS along with the classifier—The 19 identified predictive factors were used for predicting CLP.

3. Classifier only—The final CLP dataset, which included 110 factors, was used for developing the CLP predictive model.

The Testing Dataset, which comprised the remaining 30.5 percent of the final CLP dataset, was used for estimating and comparing the performance of the developed CLP predictive models with respect to the mentioned performance measures, RMSE, MAE, and correlation coefficient. A sensitivity analysis considering different combinations of HFS, PCA, and the four classifiers was conducted to identify the best-performing CLP predictive model. The results of this analysis are presented in Table 2.

Table 2. The alternative CLP predictive models with their corresponding performances

Category	CLP predictive model	HFS	PCA	Classifier	Model performance		
					RMSE	MAE	Correlation
1	1	✓	✓	RF	0.668	0.516	0.609
	2	✓	×	RF	0.829	0.679	0.309
	3	×	×	RF	0.861	0.686	0.255
2	1	✓	✓	ANN	0.777	0.550	0.190
	2	✓	×	ANN	0.862	0.638	0.036
	3	×	×	ANN	1.538	1.287	0.165
3	1	✓	✓	KNN	0.785	0.673	0.555
	2	✓	×	KNN	1.033	0.834	0.116
	3	×	×	KNN	0.952	0.713	0.000
4	1	✓	✓	ANFIS	0.707	0.554	0.551
	2	✓	×	ANFIS	1.564	1.083	-0.177
	3	×	×	ANFIS	1.289	1.054	0.285

As Table 2 shows, the first CLP predictive model in each category, which includes the HFS-PCA method, outperformed the other models in the same category based on the three performance measures. Thus, (1) employing HFS-PCA for CLP prediction was better than employing some or no individual parts of this method, and (2) HFS-PCA performed successfully along with a variety of classifiers. Furthermore, the results emphasized the efficiency of the proposed prediction procedure, which reduced computational complexity of the high-dimensional CLP dataset by determining the most predictive CLP factors and reducing their dimensionality.

Among the predictive models listed in Table 2, the first model in the first category, which uses HFS-PCA with RF as the classifier, outperformed the other models with the following performance outputs: RMSE of 0.668, MAE of 0.516, and correlation coefficient of 0.609. Using the RF classifier with HFS-PCA dimensionality reduction reduced RMSE by 19.4% compared with using RF with HFS only. Similarly, using RF with HFS-PCA reduced RMSE by 22.42% compared with using RF with no dimensionality reduction method.

Based on RMSE and MAE values, the developed predictive models using the HFS method only (the second model in each category) do not necessarily achieve higher prediction accuracy than the models without a dimensionality reduction method (the third model in each category). In this sensitivity analysis, the prediction accuracy depended on the classification method used; when using RF and ANN as the classifiers, RMSE and MAE of the second model in each category are lower than RMSE and MAE of the third model in each category, which used no dimensionality reduction method. However, the reverse is true with KNN or ANFIS classification. The third performance measure, correlation coefficient, shows a statistical correlation between actual CLP and predicted CLP of the Testing Dataset.

The second predictive model in the fourth category provides a negative statistical correlation. According to Witten et al. (2017), negative correlation coefficient values should not occur for a reasonable predictive model, since negative correlation means that large values of the predicted CLP correspond to small values of the actual CLP and vice versa. Therefore, the negative correlation of the second predictive model in the fourth category indicates that this predictive model is deficient.

Comparing the correlation of the predictive models that use the HFS-PCA method, the first model in the second category that uses ANN as the classifier has a correlation of 0.190, while the correlation coefficients of the first models in the other categories are greater than 0.551. Thus, based on correlation coefficient, the predictive model that uses HFS-PCA with ANN as the classifier does not perform as accurately as the models that use HFS-PCA with the other classifiers (RF, KNN, ANFIS).

Comparing the results of the study with past studies indicated that HFS-PCA can better identify the most predictive CLP factors. Tsehayae and Fayek (2016) obtained 2.515 as the RMSE value, while in this study using the same dataset, the RMSE value of the best CLP predictive model was 0.668. Therefore, the CLP predictive model developed by HFS-PCA achieved better performance accuracy in CLP prediction compared with CLP prediction by Tsehayae and Fayek (2016). Thus, by transforming the high-dimensional data into data with a lesser number of factors, the developed HFS-PCA method leads to better identification of the most predictive factors for improving CLP and achieved better performance accuracy. Researchers can use HFS-PCA to identify the most predictive factors of CLP and avoid having to keep track of less-predictive factors. Furthermore, identification of the most predictive factors affecting CLP helps construction companies identify the improvement strategies that correspond to and can address specific identified CLP factors. Accordingly, this method has the potential to benefit construction companies in reasonably evaluating and predicting daily CLP while avoiding high computational complexity.

5. Conclusions and Future Work

Because construction is a labour-dependent industry, construction labour productivity (CLP) has long been a major research area in the construction engineering domain, and most previous studies have focused on this kind of productivity. The main challenge in predicting CLP is the large number of factors that directly or indirectly influence labour productivity. Additionally, most previous studies did not consider the dynamics, interconnection, and combined impact of CLP factors using a model that is not dependent on expert knowledge.

The main goal of this study was to develop a novel approach for predicting and modeling CLP. The proposed methodology consists of two phases, namely, (1) the data preparation stage and (2) the data analysis stage, which includes dimensionality reduction and CLP prediction. The integration of HFS and PCA was used as a dimensionality reduction method for selecting key CLP factors. Next, several classifiers were used for predicting CLP, and the results were compared. Implementation of the proposed model on a real case led to identification of the most predictive CLP factors. The results indicate that CLP prediction performance after employing HFS-PCA is better than employing some or no parts of this method. Also, the achieved error in this study indicates an improvement of the predictive model compared with past studies. Additionally, the proposed model's filter and PCA methods result in low computational complexity and computational time.

The contributions of this paper are threefold: (1) development of a novel approach using HFS-PCA for feature selection and extraction to select factors with the most influence on CLP, (2) development of an improved model for predicting and modeling CLP, and (3) ranking factors most predictive of CLP, such as *Fairness of work assignment*, *Complexity of task*, and *Repetitiveness of task*. The study results demonstrate that the proposed model enhanced the prediction of CLP. Better identification of predictive factors of CLP can lead to more effective management of productivity and project performance. Additionally, by implementing HFS-PCA, 19 factors were identified as the most predictive factors of labour productivity, and enhancing the level of these factors for similar future projects can lead to great improvement in the value of CLP of projects.

Utilizing HFS-PCA in user-friendly software could help construction practitioners identify the most predictive factors of CLP for their organizations. Future research on advancing CLP prediction modeling may focus on measuring improvements in productivity by improving each top-ranked CLP factor. Researchers may consider modeling CLP improvements based on a combination of improving factors both individually and simultaneously. As the factors addressed in this study focus on material-related and management-related factors affecting CLP, future studies may consider the effects of buildability factors (e.g., volume placed, concrete workability, rebar congestion) or other types of factors that may affect CLP in order to develop more generalized and reasonably accurate CLP predictive models. Further, future studies may use the proposed HFS-PCA approach to obtain better performance in modeling multifactor construction productivity, which includes labour, equipment, and material.

Competing Interests

The authors declare there are no competing interests.

Contributors' Statement

S.E.: Conceptualization, Methodology, Formal analysis, Software, Investigation, Writing - Original Draft, Review & Editing. **M.K.:** Conceptualization, Methodology, Formal analysis, Software, Investigation, Writing - Original Draft, Review & Editing. **V. S.:** Methodology, Formal analysis, Writing - Review & Editing. **A.R.F.:** Conceptualization, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

Funding

This research is funded by the Natural Sciences and Engineering Research Council of Canada

(NSERC) Industrial Research Chair (IRC) in Strategic Construction Modeling and Delivery (NSERC IRCPJ 428226–15), which is held by Dr. Aminah Robinson Fayek. The authors gratefully acknowledge the financial support provided by industry partners and NSERC through the IRC.

Data Availability

All data, models, and code generated or used during the study appear in the submitted article.

References

Abo El-Maaty, A.M., and Wassal, A.G. 2019. Hybrid GA-PCA feature selection approach for inertial human activity recognition. *In* Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI 2018), Bangalore, India, 18–21 November 2018. . Institute of Electrical and Electronics Engineers (IEEE), New York, pp. 1027–1032. doi:10.1109/SSCI.2018.8628702.

Agrawal, A., and Halder, S. 2020. Identifying factors affecting construction labour productivity in India and measures to improve productivity. Asian Journal of Civil Engineering, **21**(4): 569–579. doi:10.1007/s42107-019-00212-3.

Ahmad, S.S.S., and Pedrycz, W. 2012. Data and feature reduction in fuzzy modeling through particle swarm optimization. Applied Computational Intelligence and Soft Computing, **2012**: 347157. doi:10.1155/2012/347157.

Alaghbari, W., Al-Sakkaf, A.A., and Sultan, B. 2019. Factors affecting construction labour productivity in Yemen. International Journal of Construction Management, **19**(1): 79–91. doi:10.1080/15623599.2017.1382091.

Cao, Y., Ashuri, B., and Baek, M. 2018. Prediction of unit price bids of resurfacing highway projects through ensemble machine learning. Journal of Computing in Civil Engineering,

32(5): 04018043. doi:10.1061/(asce)cp.1943-5487.0000788.

Cheng, M.-Y., Cao, M.-T., and Jaya Mendrofa, A.Y. 2020. Dynamic feature selection for accurately predicting construction productivity using symbiotic organisms search-optimized least square support vector machine. Journal of Building Engineering, **35**: 101973. doi:10.1016/j.jobe.2020.101973.

Chigara, B., and Moyo, T. 2014. Factors affecting labor productivity on building projects in Zimbabwe. International Journal of Architecture, Engineering and Construction, **3**(1): 57–65. doi:10.7492/ijaec.2014.005.

Dai, J., and Goodrum, P.M. 2012. Generational differences on craft workers' perceptions of the factors affecting labour productivity. Canadian Journal of Civil Engineering, **39**(9): 1018–1026. doi:10.1139/L2012-053.

Durdyev, S., Ismail, S., and Kandymov, N. 2018. Structural equation model of the factors affecting construction labor productivity. Journal of Construction Engineering and Management, **144**(4): 04018007. doi:10.1061/(asce)co.1943-7862.0001452.

Ebrahimi, S., Fayek, A.R., and Sumati, V. 2021. Hybrid artificial intelligence HFS-RF-PSO model for construction labor productivity prediction and optimization. Algorithms, **14**(7): 214. doi:10.3390/a14070214.

Ebrahimi, S., Raoufi, M., and Fayek, A.R. 2020. Framework for integrating an artificial neural network and a genetic algorithm to develop a predictive model for construction labor productivity. *In* Construction Research Congress 2020. American Society of Civil Engineers, Reston, VA. pp. 58–66. doi:10.1061/9780784482865.007.

El-Gohary, K.M., Aziz, R.F., and Abdel-Khalek, H.A. 2017. Engineering approach using ANN to improve and predict construction labor productivity under different influences. Journal of Construction Engineering and Management, **143**(8): 04017045. doi:10.1061/(ASCE)CO.1943-7862.0001340.

Faisal Elrawy, M., Abdelhamid, T.K., and Mohamed, A.M. 2013. IDS in Telecommunication network using PCA. International journal of Computer Networks & Communications, **5**(4): 147–157. doi:10.5121/ijcnc.2013.5412.

Fernández-Delgado, M., Cernadas, E., Barro, S., Ribeiro, J., and Neves, J. 2014. Direct kernel perceptron (DKP): Ultra-fast kernel ELM-based classification with non-iterative closed-form weight calculation. Neural Networks, **50**: 60–71. doi:10.1016/j.neunet.2013.11.002.

Frigerio, L., de Oliveira, A.S., Gomez, L., and Duverger, P. 2019. Differentially private generative adversarial networks for time series, continuous, and discrete open data. *In* IFIP Advances in Information and Communication Technology. Springer, pp. 151–164.

Gerami Seresht, N., and Fayek, A.R. 2018. Dynamic modeling of multifactor construction productivity for equipment-intensive activities. Journal of Construction Engineering and Management, **144**(9): 04018091. doi:10.1061/(asce)co.1943-7862.0001549.

Gerami Seresht, N., Lourenzutti, R., and Fayek, A.R. 2020. A fuzzy clustering algorithm for developing predictive models in construction applications. Applied Soft Computing, **96**: 106679. doi:10.1016/j.asoc.2020.106679.

Ghazi Al-Kofahi, Z., Mahdavian, A., and Oloufa, A. 2021. A dynamic modelling of labor productivity impacts arising from change orders in road projects. Canadian Journal of Civil Engineering, **49**(2): 159–170. doi:10.1139/cjce-2020-0456.

Ghosh, M., Guha, R., Sarkar, R., and Abraham, A. 2019. A wrapper-filter feature selection technique based on ant colony optimization. Neural Computing and Applications, **32**: 7839–7857. doi:10.1007/s00521-019-04171-3.

Golnaraghi, S., Moselhi, O., Alkass, S., and Zangenehmadar, Z. 2020. Predicting construction labor productivity using lower upper decomposition radial base function neural network. Engineering Reports, **2**(2): 1–16. doi:10.1002/eng2.12107.

Hafez, M.S. 2014. Critical factors affecting construction labor productivity in Egypt. American Journal of Civil Engineering, **2**(2): 35–40. doi:10.11648/j.ajce.20140202.14.

Heravi, G., and Eslamdoost, E. 2015. Applying artificial neural networks for measuring and predicting construction-labor productivity. Journal of Construction Engineering and Management, **141**(10): 04015032. doi:10.1061/(asce)co.1943-7862.0001006.

Irfan, M., Zahoor, H., Abbas, M., and Ali, Y. 2020. Determinants of labor productivity for building projects in Pakistan. Journal of Construction Engineering, Management & Innovation, **3**(2): 85–100. doi:10.31462/jcemi.2020.02085100.

Jain, D., and Singh, V. 2018. An efficient hybrid feature selection model for dimensionality reduction. Procedia Computer Science, **132**: 333–341. doi:10.1016/j.procs.2018.05.188.

Jarkas, A.M. 2015. Factors influencing labour productivity in Bahrain's construction industry. International Journal of Construction Management, **15**(1): 94–108. doi:10.1080/15623599.2015.1012143.

Kavitha, R., and Kannan, E. 2016. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. *In* 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS 2016), Pudukkottai, India, 24–26 February 2016., pp. 1–5. doi:10.1109/ICETETS.2016.7603000.

Kazerooni, M., Raoufi, M., and Fayek, A.R. 2020. Framework to analyze construction labor productivity using fuzzy data clustering and multi-criteria decision-making. *In* Construction Research Congress 2020. American Society of Civil Engineers, Reston, Virg., pp. 48–57. doi:10.1061/9780784482865.006.

Khanzadi, M., Nasirzadeh, F., Mir, M., and Nojedehi, P. 2017. Prediction and improvement of labor productivity using hybrid system dynamics and agent-based modeling approach. Construction Innovation, **18**(1): 2–19. doi:10.1108/CI-06-2015-0034.

Kononenko, I. 1994. Estimating attributes: Analysis and extensions of RELIEF. *In* European Conference on Machine Learning (1994), Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence). pp. 171–182. *Edited by* Bergadano F., De Raedt L. doi:10.1007/3-540-57868-4_57.

Ma, X., and Zhong, Q. 2016. Missing value imputation method for disaster decision-making using K nearest neighbor. Journal of Applied Statistics, **43**(4): 767–781. doi:10.1080/02664763.2015.1077377.

Mirahadi, F., and Zayed, T. 2016. Simulation-based construction productivity forecast using Neural-Network-Driven Fuzzy Reasoning. Automation in Construction, **65**: 102–115. doi:10.1016/j.autcon.2015.12.021.

Mohammed, M.N., and Ahmed, M.M. 2019. Data preparation and reduction technique in intrusion detection systems: ANOVA-PCA. International Journal of Computer Science and Security (IJCSS), **13**(5): 167–182. Available from https://www.cscjournals.org/manuscript/Journals/IJCSS/Volume13/Issue5/IJCSS-1498.pdf [accessed 9 February 2022].

Mohsenijam, A., and Lu, M. 2019. Framework for developing labour-hour prediction models from project design features: Case study in structural steel fabrication. Canadian Journal of Civil Engineering, **46**(10): 871–880. doi:10.1139/cjce-2018-0349.

Montaser, N.M., Mahdi, I.M., Mahdi, H.A., and Rashid, I.A. 2018. Factors affecting construction labor productivity for construction of pre-stressed concrete bridges. International Journal of Construction Engineering and Management, **7**(6): 193–206. doi:10.5923/j.ijcem.20180706.01.

Moselhi, O., and Khan, Z. 2012. Significance ranking of parameters impacting construction labour productivity. Construction Innovation, **12**(3): 272–296. doi:10.1108/14714171211244541.

Nasirzadeh, F., Kabir, H.M.D., Akbari, M., Khosravi, A., Nahavandi, S., and Carmichael, D.G. 2020. ANN-based prediction intervals to forecast labour productivity. Engineering, Construction and Architectural Management, **27**(9): 2335–2351. doi:10.1108/ECAM-08-2019-0406.

Pai, S.G.S., Sanayei, M., and Smith, I.F.C. 2021. Model-class selection using clustering and classification for structural identification and prediction. Journal of Computing in Civil Engineering, **35**(1): 04020051. doi:10.1061/(asce)cp.1943-5487.0000932.

Peker, M., Balli, S., and Sagbas, E.A. 2020. Predicting human actions using a hybrid of ReliefF feature selection and kernel-based extreme learning machine. *In* Cognitive Analytics: Concepts, Methodologies, Tools, and Applications. IGI Global, pp. 307–325. doi:10.4018/978-1-7998-2460-2.ch017.

Piao, Y., and Ryu, K.H. 2017. A hybrid feature selection method based on symmetrical uncertainty and support vector machine for high–dimensional data classification. *In* Asian Conference on

Intelligent Information and Database Systems, Lecture Notes in Computer Science, volume 10191. *Edited by* N. Nguyen, S. Tojo, L. Nguyen, and B. Trawiński. Springer, Cham. pp. 721–727. doi:10.1007/978-3-319-54472-4_67.

Raoufi, M., and Fayek, A.R. 2018. Fuzzy agent-based modeling of construction crew motivation and performance. Journal of Computing in Civil Engineering, **32**(5): 04018035. doi:10.1061/(asce)cp.1943-5487.0000777.

RazaviAlavi, S., and AbouRizk, S. 2017. Site layout and construction plan optimization using an integrated genetic algorithm simulation framework. Journal of Computing in Civil Engineering, **31**(4): 04017011. doi:10.1061/(ASCE)CP.1943-5487.0000653.

Sahu, B., Mohanty, S., and Rout, S. 2018. A hybrid approach for breast cancer classification and diagnosis. ICST Transactions on Scalable Information Systems, **0**(0): 156086. doi:10.4108/eai.19-12-2018.156086.

Salo, F., Nassif, A.B., and Essex, A. 2019. Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. Computer Networks, **148**(1): 164–175. doi:10.1016/j.comnet.2018.11.010.

Sandbhor, S., and Chaphalkar, N.B. 2019. Impact of outlier detection on neural networks based property value prediction. *In* Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing, volume 862. *Edited by* S. Satapathy, V. Bhateja, R. Somanah, X.S.Yang, and R. Senkerik. pp. 481–495. doi:10.1007/978-981-13-3329-3_45.

Song, L., and Abourizk, S.M. 2008. Measuring and modeling labor productivity using historical data. Journal of Construction Engineering and Management, **134**(10): 786–794.

doi:10.1061/(ASCE)0733-9364(2008)134:10(786).

Thomas, A. V., and Sudhakumar, J. 2014. Modelling masonry labour productivity using multiple regression. *In* Proceedings 30th Annual Association of Researchers in Construction Management Conference (ARCOM 2014), Portmouth, UK, 1–3 September 2014. Association of Researchers in Construction Management, UK, pp. 1345–1354.

Tsehayae, A.A., and Fayek, A.R. 2014. Identification and comparative analysis of key parameters influencing construction labour productivity in building and industrial projects. Canadian Journal of Civil Engineering, **41**(10): 878–891. doi:10.1139/cjce-2014-0031.

Tsehayae, A.A., and Fayek, A.R. 2016. Developing and optimizing context-specific fuzzy inference system-based construction labor productivity models. Journal of Construction Engineering and Management, **142**(7): 04016017. doi:10.1061/(ASCE)CO.1943-7862.0001127.

Witten, I.H., Frank, E., Hall, M.A., and Pal, C.J. 2017. Chapter 3: Output: Knowledge representation. Data mining: Practical machine learning tools and techniques, fourth edition. Morgan Kaufmann Books, Burlington, Mass., pp. 67–89.

Supplementary Materials

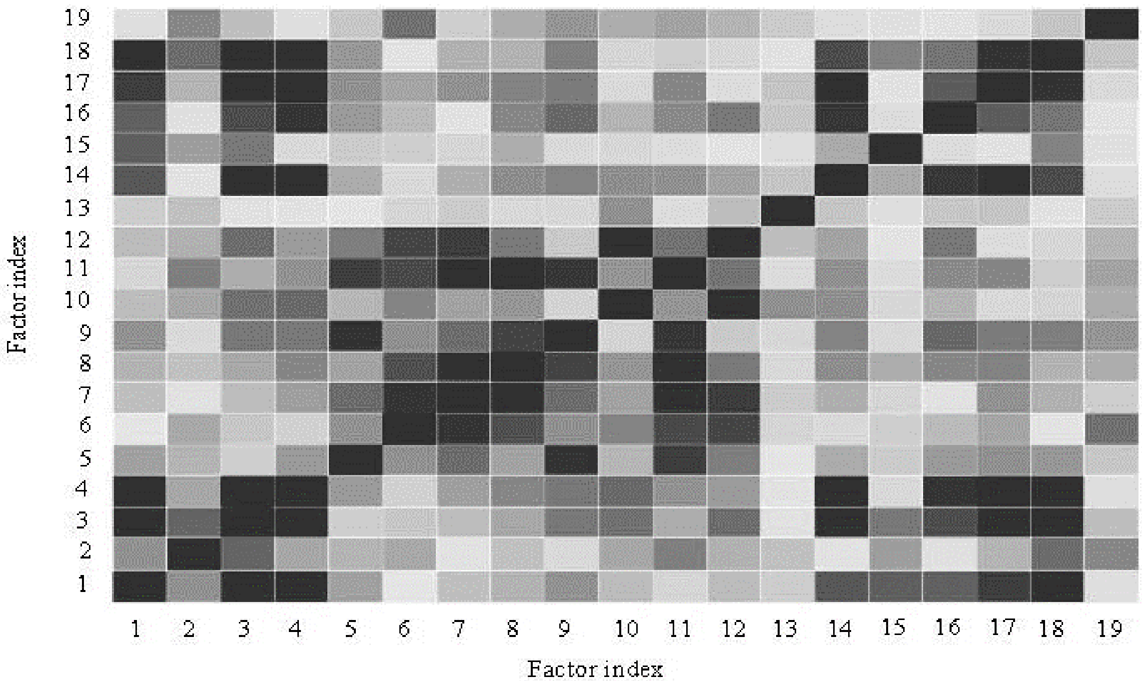

Fig. S1. Correlation matrix image of the selected CLP factors. (Note: The lighter the cell, the lower the correlation; so, black–and–dark-gray cells represent the highest correlations.)

Table S1. Variance between the eigenvectors of the factors selected by HFS

Eigenvector Number	1	2	3	4	5	6	7	8	9	10	11	12
1	6.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	3.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	1.84	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.84	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.76	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.61	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.49	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.41	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12

Table S2. CLP Factors of Dataset

No.	Factor	Scale of Measure
1	Crew size	Integer (total number of crew members)
2	Craftsperson education	Categorical (most frequent category)
3	Craftsperson on-job training	Real number (number of trainings attended × duration of training, hrs)
4	Craftsperson technical training	Real number (number of trainings attended × duration of training, hrs)
5	Crew composition	Proportion (ratio of journeyman to apprentice and helper, 1 JR / 2 AP)
6	Crew experience (seniority)	Real number (crew average years of experience)
7	Number of languages spoken	Integer (number of languages spoken, total for a crew)
8	Co-operation among craftspersons	1–5 Predetermined rating
9	Treatment of craftsperson by foreman	1–5 Predetermined rating
10	Craftsperson motivation	1–5 Rating
11	Craftsperson fatigue	Real number (ratio of total worked hours per week to regular work hours per week)
12	Craftsperson trust in foreman	1–5 Predetermined rating
13	Team spirit of crew	1–5 Predetermined rating
14	Level of absenteeism	Percentage (% average number of absent crew members to total crew size, daily average)
15	Crew turnover	Turnover rate (% of crew)
16	Discontinuity in crew makeup	Real number (average occurrence of crew member change)
17	Level of interruption and disruption	Integer (number of interruptions and disruptions per day)
18	Fairness of work assignment	1–5 Predetermined rating
19	Crew participation in foreman's decision-making process	Categorical (decision type)
20	Crew flexibility	1–5 Rating
21	Job site orientation program	Categorical
22	Job security	Integer (average length of unemployment period, months)

23	Availability of craftspersons	Integer (average number of unmet labour demands per crew for a given task)
24	Availability of task materials	Real number (average waiting time for getting materials, person-hours)
25	Quality of task materials	1–5 Predetermined rating
26	Material unloading practices	Real number (average unloading time, min.)
27	Material movement practices (horizontal)	Real number (average distance, m)
28	Material movement practices (vertical)	Real number (average distance, m)
29	Availability of work equipment (crane, forklift)	1–5 rating
30	Availability of transport equipment (person lift)	1–5 rating
31	Equipment breakdown	Integer (equipment type and average number of breakdown occurrences per week)
32	Availability of tools	Real number (average waiting time, min.)
33	Sharing of tools	Real number (number of crews sharing a tool)
34	Quality of tools	Real number (average number of tool breakdowns per week)
35	Misplacement of tools	Real number (average number of misplacements per day)
36	Availability of electric power	Real number (average waiting time, min.)
37	Availability of extension cords	Real number (average waiting time, min.)
38	Complexity of task	1–5 Predetermined rating
39	Repetitiveness of task	Real number (ratio of identical work tasks quantity to the total work task quantity)
40	Total work volume	Real number (approved quantity for construction)
41	Level of rework	Real number (construction field rework index)
42	Frequency of rework	Real number (number of rework occurrences per scope of work)
43	Task change orders – Extent	Real number (ratio of approved total volume of change orders to total work volume)
44	Task change orders – Frequency	Real number (number of occurrences per scope of work)
45	Working condition (noise)	1–5 Predetermined rating

46	Working condition (dust and fumes)	1–5 Predetermined rating
47	Location of work scope (distance)	Real number (distance, m)
48	Location of work scope (elevation)	Real number (distance, m)
49	Congestion of work area	Real number (ratio of actual peak manpower to actual average manpower)
50	Cleanliness of work area	Integer (number of cleaning operations per day)
51	Foreman skill and responsibility	1–5 Rating
52	Fairness in performance review of crew by foreman	1–5 Predetermined rating
53	Change of foremen	Turnover rate (number of turnovers per month)
54	Span of control	Integer (average total number of crews per foreman)
55	Response rate with RFIs	Real number (response time, hrs)
56	Concrete placement technique	Categorical
57	Structural element	Categorical
58	Change in design drawings	Real number (ratio of number of changed drawings to total number of drawings per discipline)
59	Change in specifications	Real number (ratio of number of changed specifications to total number of specification clauses on specific scope)
60	Changes in contract conditions	Real number (ratio of number of contract conditions changes to total number of contract clauses on specific scope)
61	Lack of information	Real number (number of RFI's per month per discipline)
62	Approval for building permit	Real number (average process time for work or permit approval, months)
63	Year of construction (to identify relation)	Integer (year of construction)
64	Project level rework	Real number (project overall construction field rework index)
65	Project level change order	Real number (ratio approved total cost of change order overall project to original approved project cost)
66	Weather (temperature)	Real number (˚C)
67	Weather (precipitation)	Real number (mm)

68	Weather (humidity)	Real number (%)
69	Weather (wind speed)	Real number (km/hr)
70	Variability of weather	1–5 Rating
71	Ground conditions	1–5 Predetermined rating
72	Site congestion	Real number (ratio of free site space to total site area)
73	Width of site access	Real number (width of access, m)
74	Queue time to access site	Real number (average queue time to access time, min.)
75	Project work times	1–5 Rating
76	Owner staff on site	Integer (total number of owner staff on site)
77	Approval of shop drawings and sample materials	Real number (average time taken to approve, days)
78	Support and administrative staff	Real number (ratio of support to technical staff)
79	Level of paper work for work approval	1–5 Rating
80	Treatment of foremen by superintendent and project manager	1–5 Predetermined rating
81	Uniformity of work rules by superintendent	1–5 Predetermined rating
82	Availability of labour	Real number (unmet labour requirement for the given trade)
83	Labour disputes (legal cases involving a worker on a project)	Real number (average number of cases per project)
84	Project cost control	1–5 Rating
85	Labour productivity measurement practice	1–5 Predetermined rating
86	Quality audits	Real number (number of inspections per month)
87	Inspection delay	Real number (average delay for inspection, min)
88	Interference	Real number (average number of interruptions due to interference)
89	Out-of-sequence inspection or survey work	Real number (number of occurrences per week)
90	Project safety plan execution	1–5 rating
91	Safety training	Real number (number of trainings attended × duration of training, hrs)

92	Safety inspections	Real number (number of inspections per month)
93	Safety audits	Real number (number of audits per month)
94	Safety incidents	1–5 Predetermined rating
95	Equipment/property damage	Integer (number of reported equipment/ property damage incidents per month)
96	Safety incident investigation	1–5 Rating
97	Project safety administration and reporting	1–5 Predetermined rating
98	Risk monitoring and control	1–5 Predetermined rating
99	Crisis management	1–5 Predetermined rating
100	Communication between different trades	1–5 Predetermined rating
101	Availability of communication devices	Real number (ratio of communication radios to number of crews, %)
102	Hiring practices (open shop)	1–5 Predetermined rating
103	Project team development	1–5 rating
104	Project team closeout	1–5 rating
105	Project environmental assurance	1–5 Predetermined rating
106	Environmental audits	Real number (number of inspections per month)
107	Sorting of waste materials	1–3 Predetermined rating
108	Project environmental control	1–5 Predetermined rating
109	Oil price	Real number (dollars/barrel)
110	Oil price fluctuation	Real number (weekly price change, %)
111	Natural gas price	Real number (dollars/GJ)
112	Natural gas fluctuation	Real number (weekly price change, %)