Handling missing data for construction waste management: machine learning based on aggregated waste generation behaviors

Zhongze Yang[1], Fan Xue[2*] and Weisheng Lu[3]

Highlights

- A set of 821 waste generation behavioral features defined and analyzed for construction projects
- 'Missing not at random' project data handled by machine learning of aggregated behavioral data
- Adaboost selected based on experiments on 2,451 construction projects
- Satisfactory results in 10-fold cross-validation ($F_1 = 0.87$) and real-world tests ($F_1 = 0.80$)
- Effective and inexpensive approach to portraying project behaviors in waste big data

Abstract

In the era of big data, data is increasingly driving the construction waste management (CWM) for minimizing the impacts on the environment and recycling construction materials. However, missing data, led by various information barriers, often undermines the decision-making and hinders effective CWM. This paper applies aggregated behavior-based machine learning (ML)

[1] Zhongze YANG, Research Assistant, M.Sc.
Department of Real Estate and Construction, The University of Hong Kong, Pokfulam, Hong Kong SAR, China
Email: zhongze@hku.hk

[2] Fan XUE, Assistant Professor, Ph.D. Homepage: https://frankxue.com
Department of Real Estate and Construction, The University of Hong Kong, Pokfulam, Hong Kong SAR, China
Email: xuef@hku.hk, ORCiD: https://orcid.org/0000-0003-2217-3693

*: Corresponding author, Tel: +852 3917 4174, Fax: +852 2559 9457, Email: xuef@hku.hk

[3] Weisheng LU, Professor, Ph.D.
Department of Real Estate and Construction, The University of Hong Kong, Pokfulam, Hong Kong SAR, China
Email: wilsonlu@hku.hk, ORCiD: https://orcid.org/0000-0003-4674-0357

methods to handling the project-level 'missing not at random' (MNAR) data by using aggregated waste generation behaviors as a case study. First, we define a set of 821 waste generation behavioral features based on waste big data, then screen the indicative and decisive behaviors using automatic feature selection. Then, the most predictive ML method, trained via data of 2,451 construction projects in 2011-2016 in Hong Kong, is selected for handling the MNAR data. The experiments showed that the prediction of project missing data was satisfactory (validation $F_1 = 0.87$, test $F_1 = 0.80$). The contribution of this paper is to pinpoint the potential of waste big data in portraying project behaviors for more value-added applications, at the same time, to present a handling method for MNAR data that is automatic, fast, and low-cost from the CWM practitioner's perspective.

Keywords

Construction waste management, waste generation behavior, missing data handling, big data analytics, machine learning

1 Introduction

Construction wastes have great impacts on the environment through different aspects like ecosystem, environmental sustainability, and natural resources, while recycling of construction materials from wastes can loosen the pressure on constrained natural materials, such as running-out river sands (UNEP 2019; Osmani et al. 2008; Tang et al. 2020; Ma et al. 2020). To make construction waste management (CWM) more effective, the industry has adopted various informatization technologies including Geographic Information System (GIS), sensor technology, Internet of Things (IoT), computer vision (You & Wu 2019; You et al. 2020; Chen et al. 2021). Thanks to the rapid development of hardware and software, such as Building Information Modelling (BIM), embedded devices, and various sensors, the construction industry has entered the big data era for digitalizing construction waste handling processes (Kerzner 2017; Eastman et al. 2011; Bilal et al. 2016; Xue et al. 2021). The unprecedented amount of big data from the explosive growth of business corporates, government, and scientific databases also enabled the construction project management to fully embrace the big data analytics (Soibelman & Kim 2002).

The quality and completeness of data are crucial for data analytics (Sattari et al. 2017); however, the data of construction activities is often incomplete. Missing data, such as 'Null', 'N.A.', and sometimes misleading values led by information barriers, is an omnipresent challenge to the monitoring and evaluation in the CWM (Bilal et al. 2016; Callistus & Clinton 2016). The related information barriers in the construction industry arose from the distinctive and accelerating complexities in modern construction projects (Luo et al. 2017). First, a modern construction project often has multiple stakeholders, such as developer, contractor, designer, and suppliers, and subsequently complex intra-organizational structures and relationships,

inconsistent interest, and poor communications among the stakeholders (Olander 2007; Atkin & Skitmore 2008). Furthermore, construction work has applied diversified technologies, numerous trades, and complex processes, which are often overlapping or parallel, uncoordinated and highly variable (Kagioglou et al. 2000). In addition to complexity and flexibility, many construction processes are with uncertainties, such as the change of materials, mechanism, environment, and policies (Paslawski 2017). Other factors such as frequent turnovers of site personnel and temporary multi-organization also lead to missing data of construction projects (Love et al. 2002). From statistics and information perspectives, missing data is the missingness of information and hinders people from understanding or finding a phenomenon hiding in massive data (McKnight et al. 2007).

Mechanisms of missing data, which describe relationships between measured variables and the probability of missing data, were studied to describe, understand, and handle missing data based on specific analysis and assumptions (Baraldi & Enders 2010). Rubin (1976) classified such mechanisms into three types according to the reason why data is missing: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR means the missingness is completely independent of all observed and unobserved data; for example, a worker forgets to return the questionnaire by chance. If the missingness is independent of all unobserved data but may be dependent on the observed data, it is said to be MAR; for example, high-paid workers may be less willing to share their income information than those with lower salaries. As for MNAR, the missingness comes from non-random reasons, such as non-preferred values or ambiguous data collection instructions. Therefore, handling MNAR data requires additional conjectures and explorations for hidden reasons (Jia 2019); and MNAR is known more challenging than MCAR and MAR. However in the construction industry, numerous missing data falls in the challenging MNAR type, due to the information barriers in construction projects.

Researchers have applied several schools of statistical methods to handle general MNAR data (Baraldi & Enders 2010), such as deletion, regression imputation, multiple imputation, and maximum likelihood estimation (Peugh & Enders 2004; N.Baraldi & K.Enders 2010; Peeters et al. 2015). Statistically based approaches try to make reasonable and justifiable assumptions for model establishing, however, for MNAR data such assumptions are untestable and unverifiable, which is the main limitation for estimation (Rabe et al. 2018). After entering the big data era, advanced big data analytics and machine learning (ML) methods have thrown light upon many research domains (Ma et al. 2020), including MNAR data handling. Unlike the conventional statistics, ML methods' predictions are often more accurate due to the assumed nonlinear relationships of feature data and the ability of capturing high-order interactions between features (Jerez et al. 2010; Sattari et al. 2017; Nugroho et al. 2020).

However, most studies only focused on the numerical complement and prediction, but did not take the information barriers and the characteristics of target industries into consideration. Few studies have paid attention to the problem of missing data handling, especially MNAR data, in the construction industry.

There exist an increasing amount of well-structured feature data in CWM, which is essential to statistical analysis and ML. A structured feature referring to the knowledge of a partial subsystem could allow us easily guess the state of other parts of the same system, or a more detailed state of parts in the same classification category (Li et al. 2002). For example, the waste big data, and the aggregated waste generation behaviors of projects, waste facilities, companies, and individual trucks can offer such features. Some governments have extensively collected waste big data for environmental protection and cost control (Poon et al. 2004). For example, the Government of Hong Kong implemented the Construction Waste Disposal Charging Scheme (CWDCS) in 2006 to strengthen CWM. Based on millions of waste disposal records in the CWDCS dataset, Lu (2019) identified illegal dumping behaviors and found previously unknown characteristics of illegal dumpers. Xu et al. (2020a) also found that different types of construction projects had different waste generation patterns using passive bigger data in the same dataset. Through deep exploration of well-structured waste big dataset, MNAR data related to the information barriers could be predicted based on the characteristic waste generation behaviors with specific ML patterns.

This paper applies a ML approach based on big data-driven waste generation behavioral features to handle MNAR data for construction projects. Unlike the handling methods in quantity studies, this paper focuses on extracting and distilling the hidden behavioral features through excessive definitions and selection to facilitate existing ML methods. Besides, trained ML models provide new patterns and insights to understand the projects in the construction industry. The proposed approach was tested on the real data from CWDCS and preliminarily validated with a few developers and clients.

The remainder of the paper is organized as follows. After this introductory section, Section 2 is the literature review focusing on the missing data handling and waste generation behaviors. Section 3 presents the detailed research methods, including data collection and processing, feature definition and selection, training and evaluation of the ML methods, as well as the predictive handling of MNAR data. Section 4 reports the experimental results, sensitivity analyses, and findings from tests, followed by an in-depth discussion in Section 5. Conclusions are drawn in Section 6.

2 Literature review

2.1 Waste generation behaviors

Behavior is "anything an organism does in response to a particular situation or stimulus;" and classical theories state that "all behavior is due to a complex interaction between genetic influence and environmental experience" (Pierce & Cheney 2017). In the domain of waste management analysis, the generation laws or characteristics of construction waste can also be seen as "behaviors", since the waste generation and disposal schedules or flows are the response to the particular waste management regulation and construction process. Such behaviors, involving social, economic, environmental and human behavioral factors, can be heterogeneous in different regions and situations regarding to waste treatment (Bakshan et al. 2017; Luangcharoenrat et al. 2019; Alcay et al. 2020). For example, governments are developing and implementing new waste management systems, regulations, and techniques, which profound influence the waste generation and recycling (Tam & Tam 2006; Guerrero et al. 2013). Waste charging schemes, economic incentives of waste recycling, and personal awareness of environmental protection significantly affect waste generation and disposal (Corsini et al. 2018; Tamayo-Orbegozo et al. 2017). Many existing studies has aimed to identify the waste generation behavior to predict and explore the influence factors using dynamics and data-driven modeling techniques (Kontokosta et al. 2018). In general, current research trends on waste generation behaviors include: (i) how individuals' attitudes and decisions affect the waste generation behaviors (Lingard et al. 2000; Begum et al. 2009; Mattar et al. 2018); (ii) significant factors that influence the waste generation behaviors (Keser et al. 2012; Zhang et al. 2015); (iii) what models and techniques identify and predict the patterns of waste generation behaviors (Karadimas & Loumos 2008; Rimaitytė et al. 2012; Johnson et al. 2017). Among the studies in (iii), supervised ML methods, such as SVM, *K*-means clustering, and decision tree, have widely been adopted, sometimes with feature selection (Márquez et al. 2008; Kontokosta et al. 2018; Meza et al. 2019; Abbasi & El Hanandeh 2016).

Waste generation behaviors about solid construction wastes are profound and consistent project-level features forming different behavior patterns (Tonglet et al. 2004). Watanabe (1985) defined general patterns as an entity represented by a set of properties and feature variables. And pattern recognition is usually posed as a classification problem using supervised or unsupervised approaches (Jain et al. 2000). In the literature, abundant behavioral features can be defined on big datasets, to represent many waste generation patterns. Examples are trucks' behavioral features in Lu (Lu 2019), accumulative waste generation flows in Xu et al. (2020b), and combinations of prefabricated components in Lu et al. (2021). However, not every feature is essential, informative, or helpful for missing data handling. Feature selection aims to find the minimum subset automatically with some criteria to improve ML performance (Koller & Saharni 1996; Lu 2019). Thus, feature selection can improve the efficiency (lower time cost) and effectiveness (higher correctness) of an ML method at the same time by removing some less informative features (Vafaie & Jong 1992). Compared with principal component analysis

(PCA), which is widely used in studies and linearly transforms all the features into integrated principal components, feature selection pertains to the original meaning of features with no changes (Vafaie & Jong 1992). To sum up, waste generation behaviors of construction projects, together with dimensionality reduction, feature extraction, and feature selection, can improve ML methods for handling missing data (Tang et al. 2014).

2.2 Missing data handling

Conventional missing data techniques, such as deletion and single imputation approaches, usually perform poorly for MNAR. Although maximum likelihood estimation and multiple imputation tend to perform better than most traditional techniques, these approaches still have obvious drawbacks in handling MNAR data (Meeyai 2016). The approaches to handling missing data in the literature can be divided into four different groups: complete case analysis, imputation, maximum likelihood, and embedded ML (García-Laencina et al. 2009). Complete case analysis, as the name suggests, is a procedure based on the complete data, which ignores the missing values and requires the support of a big volume of data. Both imputation and the maximum likelihood approaches aim at estimating; while imputation methods estimate the missing values directly and maximum likelihood methods estimate the model parameters to handle missing data (Little & Rubin 2019; Schafer 1997). In the MNAR case, the observed data are no longer representative of the population, which leads to selection bias in the sample, and therefore to bias in the parameters estimation. For imputation and maximum likelihood methods, which are to model missing data distribution, will have a computational burden and are often restricted to a limited number of MNAR variables (Sportisse et al. 2020). The evaluation of the methods is in Table 1 (Pigott 2010; Shylaja & Kumar 2018; N.Baraldi & K.Enders 2010; Nugroho et al. 2020).

Table 1. Assessment of statistical missing data approaches

Methods		Advantages	Disadvantages
Deletion	Complete case analysis (List wise deletion)	1) Simple to implement 2) Acceptable with small number of missing data 3) Suitable for MCAR data	1) Need big volume of data 2) Not efficient, may cause problems in prediction
	Available case analysis (Pair wise deletion)		
Imputation	Single imputation	1) Full sample size is preserved 2) Administrative data is not needed 3) Estimated variations that can be justified by a statistic. 4) More suitable for MAR and MCAR data	1) Single imputation may lead bias to the estimation 2) Limited by the MAR assumption 3) Not practical with large databases 4) Sensitive to missing rate
	Multiple imputation		
Maximum likelihood estimation		1) Could produce accurate estimates 2) Use all the available information	1) Limited by the MAR assumption 2) Model dependent

The embedded ML methods are ML approaches, such as decision trees and fuzzy neural networks (Ishibuchi et al. 1993), which have been proven to be much more effective in handling MNAR (Twala 2009; Song et al. 2008). Mitchell (2006) defined ML as 'machine learns with respect to a particular task *T*, performance metric *P*, and type of experience *E*, if the system reliably improves its performance *P* at task *T*, following experience *E*'. In the four different types of ML, i.e., unsupervised, supervised, semi-supervised, and reinforcement, supervised ML algorithms learn or approximate a behavior of a function which maps a vector into one of several classes by looking at several input outputs examples of the function (Osisanwo et al. 2017). Therefore, supervised ML algorithms can produce a general hypothesis of missing data from externally supplied features, which applies to missing data handling (Singh et al. 2016; Ge et al. 2017). Ding & Simonoff (2010) studied three types of missing data mechanisms and proved that supervised ML classification methods could provide accurate predictions and are more robust than traditional approaches, especially for the case when data in the test group is also uncompleted. Currently, there exist many supervised ML algorithms, such as Support Vector Machines (SVM), quadratic classifiers, *k*-means clustering, *k*-nearest neighbors (*k*-NN), boosting, decision tree, random forest (RF), Artificial Neural Networks, Bayesian networks, and ensemble methods combining them thereof.

Supervised ML methods have been integrated with non-ML methods to complete missing values and collaborate to improve prediction accuracy. For example, Twala & Cartwright (2010) ensembled decision trees with two imputation methods for predicting missing data, of which the results showed that the ensemble method improved the accuracy. Nanni et al. (2012) proposed a novel ensemble approach for completing missing data and tested its performance on different databases. Rahman & Islam (2013) presented two novel techniques based on Decision Tree impute missing values and used nine publicly available datasets to indicate the superiority of these techniques. Tran et al. (2017) combined multiple imputations and ensemble ML methods to handle missing data, and found it significantly better than decision tree and *k*-NN in terms of classification accuracy.

It is generally accepted that different characteristics of missing data can produce different effects on the completion, and some researchers have suggested that these characteristics should be taken into consideration when selecting handling methods (Garciarena & Santana 2017). Thus, the performances of different ML methods have also been compared for handling missing data based on different datasets in different domains. For example, Gavankar & Sawarkar (2015) reviewed different algorithms of decision trees for handling missing data, and employed a database to benchmark the performance of the algorithms. Abidin et al. (2018) compared three methods, i.e., *k*-NN, decision tree, and Bayesian networks, for handling

missing data on a medical dataset. Perkowski (2020) focused on the bagging and boosting ensemble ML methods on handling missing data and proved the significances. Therefore, for analyzing well-structured construction waste data, the characteristics of missing data should also be considered together with ML methods. In addition, features should be demanded based on well-structured data for handling missing data in the field of construction project management.

3 Research methods

Figure 1 shows the presented ML method for handling missing waste data. In general, there are four steps. After the data preparation, the second step is the aggregation of waste generation behaviors, which is the core characteristic that distinguishes this paper in the literature. The third step is iterated ML training with feature selection and cross-validation, while the last step is the tests with ground truth data obtained from interviews.

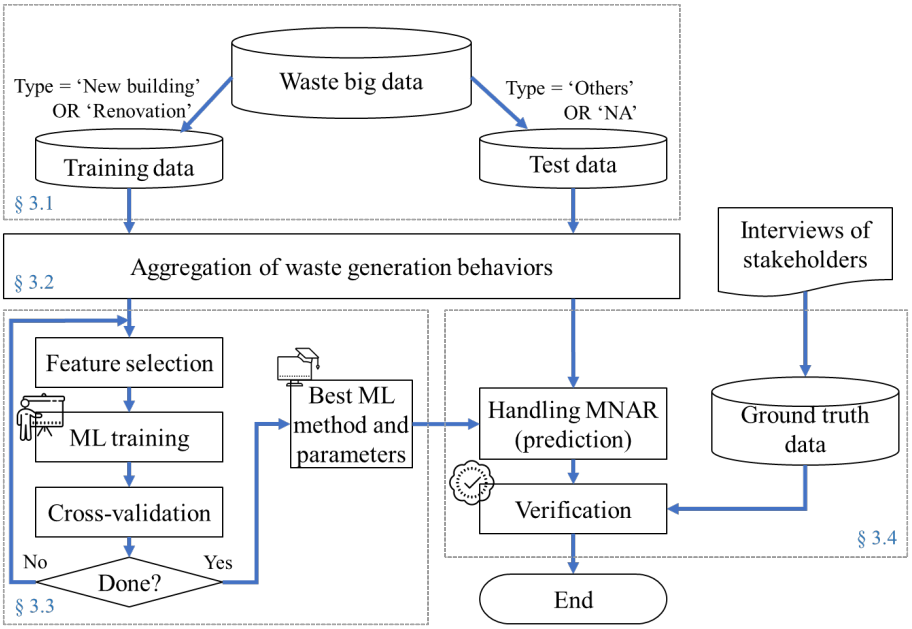

Figure 1. Flowchart of the presented ML method for handling missing waste data

3.1 Data source and preparation

The data source in this paper is the CWDCS dataset, which is a construction waste big data containing over 12 million transaction records of Hong Kong since 2011 (EPD 2020). Besides, the information of projects, facilities, vehicles, and their indexed relationships form a completed data structure regarding the construction waste transactions. The whole dataset is composed of structured data, such as numerical data and descriptive text of waste generation and disposal. In light of excluding the uncompleted projects, the study period is set to between January 2011 and December 2016 (both inclusive) including 19,626 projects. Figure 2 shows the structure of the data tables in the CWDCS dataset, where the four data tables are:

- Waste disposal data records every daily waste disposal information, including facility, account, date of transaction, vehicle, net weight, waste depth, etc.
- Facility data contains all government waste disposal facilities for construction waste, including facility type, name, and code. The abbreviation of the facility corresponds to that in the waste disposal database.
- Vehicle data contains information on vehicle plate numbers and the maximum weight capacities of 11,374 waste transportation trucks. The vehicle number corresponds to that in the waste disposal database.
- Project data contains detailed billing account information, such as contract number, contract name, site address, department, type of construction work, etc. The account number corresponds to that in the waste disposal database.

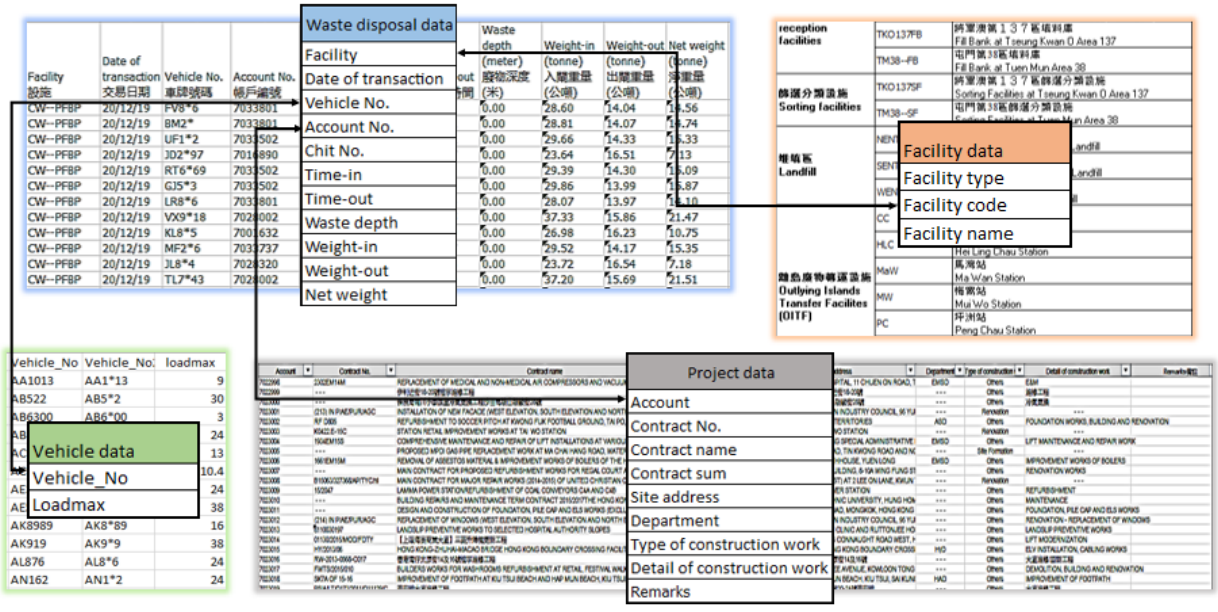

Figure 2. Excerpts and cross-references of waste disposal data, facility data, vehicle data, and project data in the CWDCS dataset

The missing data column to handle is the 'type of construction work' in the project data. According to the local regulations, the main contractor, who undertakes construction work under contracts with values of HK$1 million or above, must register a billing account in CWDCS solely for the contract. During the construction activities, construction waste will be transported and dumped into the facilities by truck. The 'type of construction work' is required to report the contract details, such as 'new building' , 'renovation' (general renovation, i.e., renovation, maintenance, and improvement (RMI)), 'demolition', 'site formation', and 'foundation work', along with project sum, site address, and contract number, as shown in Figure 2. The 'type of construction work' is essential for the government's waste charging policy reviews, strategic management of waste facilities, and regular statistical reports. For example, in Hong Kong, Monthly Digest from Buildings Department, Quarterly Report of

Employment and Vacancies at Construction Sites, and Report on the Quarterly Survey of Construction Output from the Census and Statistics Department all have specialized statistics of the performances of all construction works.

Many projects have missing project data in the CWDCS dataset. Due to the information barriers in the construction industry, the detailed information of the projects is usually vague and uncertain for information fillers, such as truck drivers, to fill confidently. Thus, fuzzy options, such as 'others', or even giving up filling, are a much preferable choice, see in Figure 3. It can be seen in Figure 3 that the majority of project type values are missing, which hinders many value-added analyses and applications for CWM.

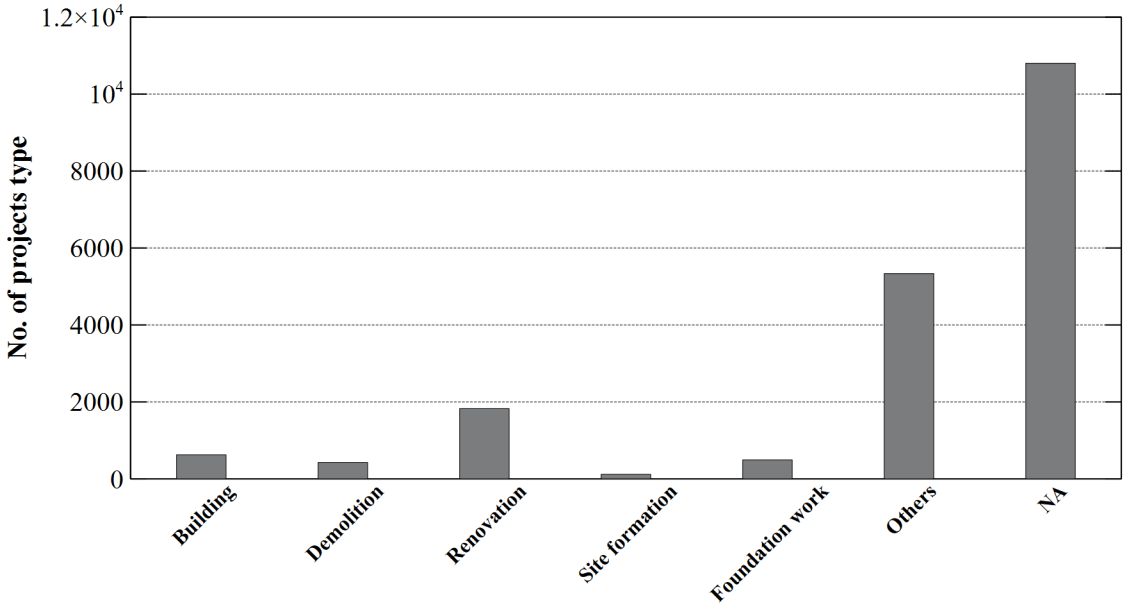

Figure 3. Project types in the CWDCS dataset (2011-2016), where 'NA' and 'others' are regarded as missing values

This paper focuses on the top two types, i.e., 'new building' and 'renovation,' of all registered projects in 2011-2016. As a result, 2,451 projects were selected to produce the training data, including 625 'new building' projects and 1,826 'renovation' projects. A data cleansing is conducted via the quartiles method, which removes data outliers and is neutral to non-normal distributions. After the data cleansing, there are 895,063 records of waste disposal transactions, to be aggregated for the 2,451 projects. The 10-fold cross-validation is used for the training ML model, instead of the conventional 80-20 training-test splitting, because of the less bias. Moreover, an extra test data set consists of our interviewee's projects originally reported as 'others' and 'NA' to further verify the selected ML method and explore the practical value, which is the final target of our research. The results of the test data set can at the same time, verify whether the model is overfitting or not.

3.2 Aggregation of waste generation behaviors

We define 821 behavioral features of waste generation for each construction project to represent the corresponding waste generation pattern thoroughly. The definitions are partially based on the pattern definition in Watanabe (1985) and partially by extending the 54 yearly indicators in Lu (2019). As listed in Table 2, there are 21 types of behavioral features clustered into three groups: (i) truck usage behavioral features, (ii) waste disposal behavioral features, and (iii) facility usage behavioral features. The definitions of the behavioral features well exploit the waste transactions, anonymous truck information, project information, and facility information interlinked in the CWDCS dataset.

Table 2. Definition of 821 features in 21 types of project waste generation behaviors

Group (Total no. of features)	Type id	Feature type definition	Unit	Features*				
				Daily (d)	Weekly (w)	Monthly (m)	Yearly (y)	Total (t)
Truck usage behaviors (226)	1	The truck usage for the pattern which the maximum load <=16t	1	$mean_j^i, stddev_j^i, max_j^i, min_j^i, pct5_j^i, pct25_j^i, pct50_j^i, pct75_j^i, pct95_j^i, extremum_j^i, IQR_j^i, i = 1,\dots,5, j = d,w,m,y$				$s_t^i = \frac{usage_i}{s_t^5}$ $i = 1,\dots,4$
	2	The truck usage for the pattern which the maximum load =24t	1					
	3	The truck usage for the pattern which the maximum load =30t	1					
	4	The truck usage for the pattern which the maximum load >=38t	1					
	5	The usage of all trucks for every waste disposal	1					s_t^5
	6	The usage count of distinct trucks for every waste disposal	1	/#				s_t^6
Waste disposal behaviors (55)	7	Activeness of waste disposal record	1	act_j^7, $j = d,w,m,y$				/#
	8	Amount of waste generation	t	/#	/#	/#	$weight_n$, $n = 11,\dots,16$	s_t^8
	9	Statistics of waste generation	t	$mean_j^i, stddev_j^i, max_j^i, min_j^i, pct5_j^i, pct25_j^i, pct50_j^i, pct75_j^i, pct95_j^i, extremum_j^i, IQR_j^i, i = 9,\dots,21, j = d,w,m,y$				/#
Facility usage behaviors (540)	10~18	The usage of the nine facilities	1					$s_t^i = \frac{usage_i}{\sum usage_i}$ $i = 10,\dots,21$
	19	The usage of all public facilities	1					
	20	The usage of all sorting facilities	1					
	21	The usage of all landfill facilities	1					

*: *Stddev*: standard deviation; *pct*: percentile; *extremum*: max − min; *IQR*: interquartile range.
#: Feature omitted. Reasons include insufficient variations from s_t^6 (Type 6), no definitions (Types 7 and 9), and represented by yearly features (Type 8).

For the first group of truck usage behaviors in Table 2, the 226 features are divided into six types by the maximum load: <=16t, =24t, =30t, >=38t, all trucks, and distinct trucks. For every behavioral feature type i, there exists a 2D array of features s_j^i, where s = mean, standard deviation, maximum, minimum, extremum, quartiles—i.e., 5th, 25th, 50th (median), 75th, 95th—or interquartile range is a statistic, and j = daily, weekly, monthly, and yearly is the period of behaviors. Besides, a total or percentage number is defined for each feature type.

The second group of 55 waste disposal behaviors represents a project's transaction activeness, yearly summation amount of waste, and the statistics with granularities j = daily, weekly, monthly, and yearly. In the third group of facility usage behaviors, the 540 features are grouped into 12 types. The first nine types represent the nine waste facilities operated in Hong Kong, while the remaining three types are the groups of the nine facilities, as listed in Table 3. Not that the 'outlying island transfer' facilities are omitted due to the low usage, which reflects construction solid wastes were rarely generated on the outlying islands.

Table 3. List of the waste handling facilities and associated feature type Ids

Facility group (feature type Id)	Feature type Id	Facility name with site	Required proportion of inert materials
Public fill (19)	10	Chai Wan Public Fill Barging Point	100%
	11	Mui Wo Temporary Public Fill Reception Facility	
	12	Fill Bank at Tseung Kwan O Area 137	
	13	Fill Bank at Tuen Mun Area 38	
Sorting (20)	14	Sorting Facilities at Tseung Kwan O Area 137	≥ 50%
	15	Sorting Facilities at Tuen Mun Area 38	
Landfill (21)	16	North East New Territories Landfill	≤ 50%
	17	South East New Territories Landfill	
	18	West New Territories Landfill	

In total, 821 features are defined for the 21 types to represent the behaviors of one of the 2,451 projects thoroughly, as listed in Table 2. Thus, after aggregation, the data set is a table of aggregated features in 2,451 rows and 821 columns of waste generation behaviors.

3.3 ML training with cross-validation and feature selection

The ML method in this paper is then applied to process the behavioral features and behavioral patterns represented by the features. Given a set $X = [f_1, f_2, \ldots, f_{821}]^T$ of waste generation behavioral features, the task of missing project data handling H in this paper is thus to find a prediction function:

$$H(X) \in C, \text{ for any } X \tag{1}$$

where C is the set of possible values of the missing data, i.e., 'new building' and 'renovation' in this study. The whole dataset for the supervised ML methods includes training input data in a 2,451 × 821 matrix and 2,451 × 1 class labels.

Four supervised ML methods are selected to represent the four popular groups of ML according to our pilot tests of the handling process H in Eq. (1). The 'decision tree' is selected from the group of trees and rules, 'k-NN' from the group of instance-based ML, 'SVM' from the group of function ML, and the 'ensemble ML' from the group of meta-models. Other ML methods,

such as Artificial Neural Networks and Bayesian Networks, are also tested but dropped, as shown in Tables A4 and A5 in the Appendix.

i. **Decision tree** learning is a data mining technique to classify data based on complete information. The structure of a decision tree contains three parts: a) root node, contains the complete set of samples; b) internal node, contains corresponding characteristic attribute tests; c) leaf node, represents the decision result (Myles et al. 2004).

ii. ***K*-Nearest Neighbors (*k*-NN)** is a basic ML algorithm that can compare features in a test set with those in a training set and find the top k most similar record, so that the classification of those inputted data is that with the most occurrences in k data (Laaksonen & Oja 1996).

iii. **Support vector machines (SVM)** is a binary classification model. Its basic model is a linear classifier with the most considerable interval defined in the feature space, and the main idea is to solve the separating hyperplane that can correctly divide the training dataset and have the largest geometric interval (Hearst et al. 1998).

iv. **Ensemble ML** combines multiple weakly-supervised models to obtain a better and more comprehensive strong-supervised model. The underlying idea of ensemble ML is that even if an individual weak classifier gets a wrong prediction, other weak classifiers can also correct it. Example ensemble ML algorithms are AdaBoost, RUSBoost, LogitBoost, GentleBoost, and Bag. (Dietterich 2000).

The F_1 scores from 10-fold cross-validation to compare the ML methods. The evaluation of a trained ML model's performance is primarily based on the F_1 score, with reference to precision and recall rates, where

$$Precision = True\ Positive\ /\ (True\ Positive + False\ Positive) \quad (2)$$

$$Recall = True\ Positive\ /\ (True\ Positive + False\ Negative) \quad (3)$$

$$F_1 = 2 \times Precision \times Recall\ /\ (Precision + Recall). \quad (4)$$

In Eqs. (2) and (3), 'true positive' counts the missing data that is estimated as 'positive' while the prediction is 'true' according to the actual data (same as the predicted); similar rules apply to 'false positive' and 'false negative.' The F_1 score in Eq. (4) integrates the precision and recall rates into one harmonic mean, evaluates the two classes at the same time, and can solve the problem of the unbalanced sample sizes of the training classes.

Feature selection is applied to trim the noisy behavioral features prior to the ML training, as suggested in the literature (Koller & Saharni 1996; Lu, Big data analytics to identify illegal construction waste dumping: A Hong Kong study 2019). The feature selection algorithm is the permutation importance in the open-source library *scikit-learn* (ver. 0.24) in Python (ver. 3.9.1)

(Pedregosa et al. 2011). The parameters in the algorithm usually include *estimator*, *X*, *y*, *scoring*, and *n_repeats*. The *estimator* is the corresponding ML classifier; the *X* is the training data; the *y* was the prediction (handed MNAR data in this paper); the *scoring* function of permutation importance is the F_1 metric, and *n_repeats* indicates how many iterations to loop for. The permutation importance indicates the contribution of features to achieve a higher F_1 score and to choose useful features. The permutation importance is calculated as follows: i) first, a baseline metric, defined by *scoring*, is evaluated on *X* dataset; ii) next, a feature column from the validation set is permuted and the metric is re-evaluated. After calculating the difference between the baseline metric and metric from the feature column permutation, this function will return the importance score of the original ranking. Features with a positive importance score will be selected for ML.

Classification Learner in MATLAB (ver. R2019b) is also used to train the ML methods comparing with the results in *scikit-learn*. In order to obtain the best ML model with the best parameter settings, the Hyperparameter Optimization in MATLAB is applied to fine-tuning the ML parameters automatically. Meanwhile, the hyperparameter optimization could efficiently avoid the overfitting of model training, since it will cease automatically based on the performance of the learning curve. The four ML methods are, therefore, fine-tuned and compared in iterations using the F_1 of 10-fold cross-validation on the training dataset. The ML method with the top F_1 metric is finally determined.

3.4 Test

This step applies the trained and optimized ML method to predict the missing data in other construction projects. A set of construction project cases are selected based on three screening conditions:

i) Public construction projects with complete contract number and contract name for obtaining trustworthy and ethical interview data; and
ii) Sufficient (> 1,000) construction waste transaction records from 2011 to 2016 or having data records for at least three years from 2011 to 2016
iii) The ‘type of construction work’ in the dataset now is ‘Others’ or ‘NA’.

After the screening, 36 projects labeled as “Others” are filtered. All the data is organized by the same approach in Section 3.2 but used the optimized ML method with the same parameters directly. Many details and background information about the 36 projects are available freely on the internet. Furthermore, we can select typical cases for interviewing the developer and client, the Hong Kong Housing Authority (HKHA), on the ground truth and reasons of reporting as missing.

4 Experiments

4.1 ML training and validation results

The experiments were conducted on a laptop computer, with an Intel i5-5200U 2.20 GHz CPU and 8GB memory, where the hyperparameter optimizer was 'Bayesian' and the maximum number of iteration was 50. Table 4 compares the evaluation results of 10-fold cross-validation of the four ML methods. In general, all the four ML methods can produce > 0.8 F_1 scores with selected feature sets sized from 74 to 378. For the decision tree method, 124 features were selected and F_1 = 0.82. For the *k*-NN method, 222 features were selected, and F_1 = 0.81. There were 378 features selected for SVM, and F_1 = 0.82. The AdaBoost method received 74 features—the least in the methods, while F_1 = 0.87 was considerably higher than the rest three methods. Therefore, the AdaBoost method was applied for handling missing data in later steps.

Table 4. Average performance metrics of the ML models in 10-fold cross-validation, where the best value in each row is in bold.

		Decision tree	*k*-NN	SVM	AdaBoost
No. of selected features		124	222	378	**74**
Precision	New building	0.76	0.76	0.76	**0.89**
	Renovation	0.90	0.89	0.90	**0.91**
	Overall	0.83	0.83	0.83	**0.90**
Recall	New building	0.69	0.67	0.68	**0.72**
	Renovation	0.93	0.93	0.93	**0.97**
	Overall	0.81	0.80	0.81	**0.85**
Validation F_1 score	New building	0.72	0.71	0.72	**0.80**
	Renovation	0.91	0.91	0.92	**0.94**
	Overall	0.82	0.81	0.82	**0.87**

The decision tree concluded from the training data is shown in Figure 4, and the pruning level is 13 out of 21. Although the decision tree failed to return the best predictions of missing data, its interpretable presentation can reveal general patterns and behavioral characteristic of the two types of construction work. The root node in Figure 4 has two sub-trees. The upper sub-tree with the condition 'the mean of yearly waste generation ≥ 185.635t' represents the projects with high-level yearly waste generation. There were 453 projects labeled as 'New building', with higher weekly waste disposal records, less usage of trucks whose maximum load is below 16t, and smaller extremum. In other words, construction projects with continuous high-level waste generation amount and lower-level fluctuations are likely 'New buildings.' In the lower sub-tree in Figure 4, 1741 projects are labeled as 'Renovation', where the 'yearly 95% quantile of the usage of all trucks for every waste disposal <201.5' branch has 1723 'Renovation' projects. It means that with lower-level yearly waste generation, construction projects having less truck usage are likely 'Renovation.'

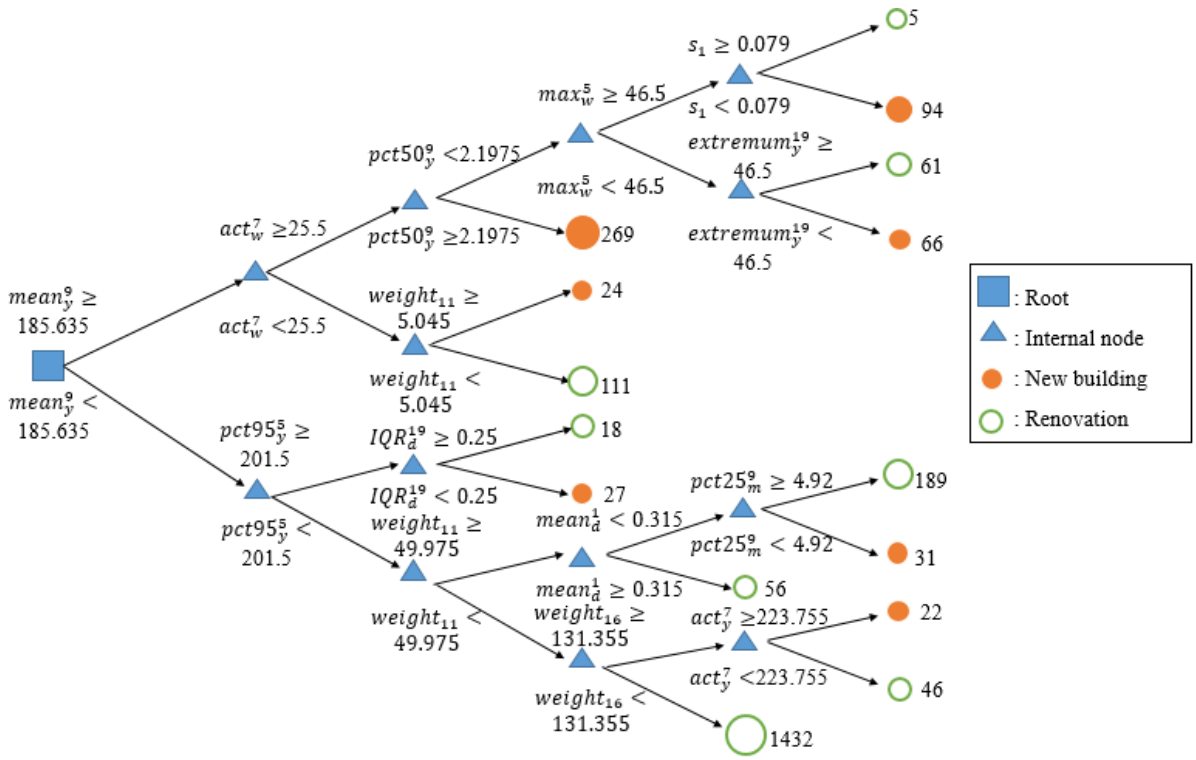

Figure 4. The decision tree learned by Classification Learner, where the number indicates the sample size of each leaf

AdaBoost method returned the best results on top of 74 features, as listed in Table 5 (other ML methods' features are in the Appendix). Among three groups of features, 6.2% (14 out of 226) truck usage features were selected, together with 38.2% (21 out of 55) waste disposal and 7.2% (39 out of 540) facility usage features. The importance scores of these 74 features were shown in Figure 5, and the statistical features of waste generation (Type 9) in the group of waste disposal behaviors was considerably the most critical than the rest in terms of predicting project type. Two following-up feature types fell in the rest two groups, respectively: one was the usage of Sorting Facilities at Tseung Kwan O Area 137 (Type 14) and the other was yearly statistics of waste generation amount (Type 8).

Table 5. List of 74 decisive features selected for AdaBoost from the pool of 821 features

Group (subtotal)	Feature type id	Feature				
		Daily (d)	Weekly (w)	Monthly (m)	Yearly (y)	Total (t)
Truck usage behaviors (14)	1	stddev, IQR	pct95	pct75		
	2		mean	pct50, pct5	mean, pct50	
	3				pct75	
	4	pct95				
	5	stddev	stddev			
	6					s_t^6
	7			act		

Waste disposal behaviors (21)	8				*weight*$_n$, n = 11 to 14, 16	
	9	mean, max, pct25, pct75, IQR	min, extremum, pct5, pct95	mean, max, min, pct25	mean, pct95	
Facility usage behaviors (39)	10	stddev			pct50	s_t^{10}
	11					
	12	IQR		pct95	pct25	
	13	mean		pct5		s_t^{13}
	14	mean, stddev	stddev	pct50	IQR	s_t^{14}
	15		extremum, pct5			s_t^{15}
	16	stddev		max		s_t^{16}
	17	mean	pct25, pct75			
	18			pct95		
	19	stddev, pct75	pct75	pct5	pct95	s_t^{19}
	20		stddev	mean	pct75	s_t^{20}
	21	stddev		mean, pct5	stddev	
	Subtotal	19	14	17	16	8

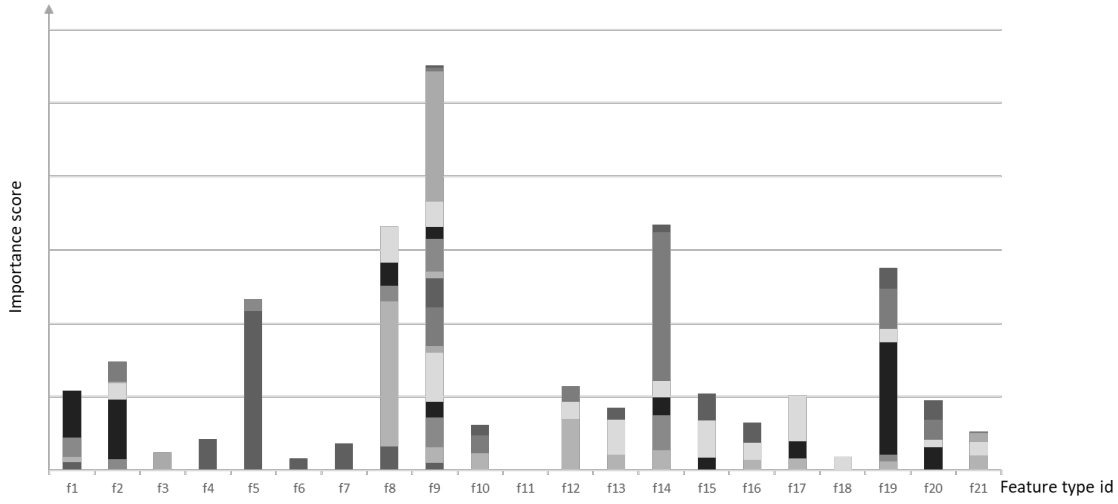

Figure 5. The importance scores of 74 decisive features, where the blocks from bottom to top represent the order from Daily (d) to Total (t).

As shown in Figure 5, the sub-feature with the highest importance score is the usage of trucks for daily waste generation, followed by the sub-features in the usage of the Sorting Facilities at Tseung Kwan O Area 137 (Type 14: yearly IQR), the amount of the waste disposal (Type 8: yearly weight in 2012), the statistics of the waste generation (Type 9: monthly pct25), and the usage of the Public Fill facilities (Type 19: weekly pct75). In other words, there were no distinct time periods of waste generation behaviors. Yet, there were unbalanced pairs in the truck types

and facilities in Table 5 and Figure 5. For the group of truck usage, the selection of the first two truck types stands for a disparity in using trucks with capacities ≤ 24t among four different truck types. As for the facility uses, Mui Wo Temporary Public Fill Reception Facility (Type 11) and West New Territories Landfill (Type 18) had less informative features. In our later investigation, it was found that the first facility received very low wastes from both new building and renovation, while the latter one had little differences between the two construction types. Sorting Facilities at Tseung Kwan O Area 137 (Type 14) and Public fill facilities (Type 19) have much more informative features. It was found that there were two disposal facilities at Tseung Kwan O Area 137, one is a public fill facility while the other is a sorting facility. It seems that the usage of these two facilities at Tseung Kwan O Area 137 contributes a lot to the distinction of the two construction types. In summary, the selected features are informative to correlate the targeted construction types.

4.2 Parameters sensitivity

Table 6 shows the F_1 scores from 10-fold cross-validations of the four ML methods on two ML libraries, i.e., *scikit-learn* and MATLAB. The ML libraries were found notably impactful to the results, even for the same ML method. In general, MATLAB's Classification Learner produced consistently higher F_1, which were preferred for handling the missing data in this study. The reason was that Classification Learner exploits an iterated 'Hyperparameter Optimization' improvement for the ML methods.

Table 6. Comparison of F_1 scores of same ML methods implemented in different scientific packages, the best value in each row in bold

ML library	Language	Decision tree	*k*-NN	SVM	AdaBoost
scikit-learn	Python	0.76	0.72	0.79	0.84
MATLAB (iter=50)	Object C	**0.82**	**0.81**	**0.82**	**0.87**

Table 7 lists the F_1 scores of the AdaBoost method, with the number of iterations increased from 1 to 70. The key parameter of the Hyperparameter Optimization in MATLAB is the number of iterations. The scores hint at an inverse U-shape, i.e., highest in the middle. A series of 30 to 50 iterations in total should train AdaBoost for the best predictions. Similarly, the minor parameters of AdaBoost were determined as 'maximum number of splits' = 16, 'number of learners' = 492, and 'learning rate' = 0.3145.

Table 7. Validation F_1 scores of AdaBoost against different number of iterations, the best value in each row in bold

No. of iterations		1	10	30	50	70
Validation F_1 score	New building:	0.76	0.78	**0.80**	**0.80**	0.79
	Renovation:	0.93	0.93	**0.94**	**0.94**	**0.94**
	Overall:	0.85	0.86	**0.87**	**0.87**	0.87

4.3 Results of tests

Application values are tested by further tests through new construction projects. First, Adaboot's selected 74 features were calculated for the test data of 36 public projects. Then, the missing data can be predicted by the trained AdaBoost model. As a result, there were 26 projects classified as 'new buildings', and other 10 projects are predicted as 'renovations'. The public project documents and reports available on the Internet claimed that there were 24 'new buildings', 12 general renovations (or RMI), as listed in Table 8. Thus, the test results was satisfactory, with *precision* = 0.82, *recall* = 0.79, F_1 = 0.80. Therefore, it is plausible to complete the missing data of construction project types using the presented ML method.

Table 8. Confusion matrix of the test results on the 36 public projects, where correct prediction are in bold

Actual / Predicted	New building	General renovation (RMI)	
		Renovation	Maintenance and improvement
New building	**22**	0	4
Renovation	2	**1**	**7**

The HKHA, as the developer and client of the test cases, was contacted for double-checking the predictions. Two construction projects, one predicted as a new building and the other predicted as renovation, were selected. We found one reason for the missing data was the inconsistent definition of the 'renovation' type. In EPD's CWDCS system, 'renovation' stands for the general renovation, or RMI; however, in some other systems, like HKHA's, the general renovation has subcategories, such as 'maintenance and improvement', a narrowly-defined 'renovation', 'commercial building', and 'civil engineering'. The inconsistent definitions are also one prevailing information barrier in the industry.

Table 9. Two project cases in the tests

Proj. ID	Project Description	Project value (HK$ million)	Period	Type of construction work		
				Our prediction	Ground truth from client	Correct ?
1	"The works comprise the building work on total construction area 132,047 m^2 with retail area, estate management office, carpark, refuse collection facilities, etc."	2,821.8	2014 – 2016	New building	New building	Yes
2	"The works comprise the maintenance and repairs of and alterations and additions to any properties, site and slopes	374.0	2014 – 2017	(General) renovation	Maintenance and improvement	Yes

	including vacant flat refurbishment."

5 Discussion

For practitioners in the construction industry, the methodology presented in this paper can summarize quantitative waste generation behaviors, predict missing data for each project, and help eliminate the information barriers. Although construction projects are known to have distinctive characteristics, their aggregated waster generation behaviors are relatively stable for analyses and applications. For example, the findings in this paper show 'New building' and 'renovation' projects had different behavioral patterns in Hong Kong (2011-2016). 'New building' projects had higher yearly waste generation amount, less usage of trucks with lower maximum load, and more usage of all the trucks; while the 'Renovation' showed opposite trends. From the perspective of CWM, the waste generation behaviors utilize the existing big data and enable new data analytics like missing data handling. Participators in CWM can explore the essential and subtle behavioral characteristics of the waste generation according to the feature selection and ML results. The completion of the construction work type may also benefits further construction waste management and studies, e.g., construction work type-specific waste facility policies. In the experiments, the waste generation behaviors showed correlations with the vehicles and facilities, and further concluded in the interpretable decision trees and non-readable ML methods.

The proposed aggregated waster generation behavior-based ML method has several advantages in handling missing project data. First, it is much more efficient (less time and human resource) than interviews to complete the missing information. The predictions were satisfactory (F_1 = 0.87) and provided a quantitative basis for further studies. Meanwhile, the decisive attributes can explain the core differences in the aggregated behaviors regarding the missing data. From the perspective of construction management, the quantitative behavioral features can help understand the projects and characteristics of different construction work types. The excessive feature generation is inclusive for the features that are easily overlooked. Based on the investigation on the predicted results, possible data management policies can be recommended, e.g., renaming the 'renovation' to 'general renovation' or 'RMI' in EPD's system. Last but not least, the proposed methods are general, so that they can be applied to other types of project data elsewhere, not limited to 'new building' and 'renovation' in Hong Kong.

However, there exist a few limitations in this study. First, only two significant values of missing data were selected to illustrate the presented methods. More extensive ranges of missing data can be tested in future studies. Secondly, one ML method, decision tree, can be interpreted in a human-readable fashion. More interpretable ML methods should be investigated in the future.

There was a slight unbalance between the missing data, e.g., 'Renovation' got higher precision, recall, and F_1 scores in Table 3. Thus, the balance, apart from correctness, should be noted in the predictions as well. Lastly, the waste big data in this paper was from Hong Kong during 2011 and 2016, the socio-industrial environments and project behaviors may change in another time in another place.

6 Conclusion

Information barriers in the construction industry lead to missing data that hinders CWM from digitalization and data-driven decision-making. This paper presents a machine learning (ML) method to handle quantitative missing not at random (MNAR) data based on aggregated projects' waste generation behaviors. Experiments on waste big data, including 895,063 rows of disposal transactions data in 2,451 construction projects, confirmed the presented method. The ensemble ML was selected as the best prediction method with F_1 = 0.87. The critical features were recognized and analyzed, besides the characteristics of the waste generation of two types of construction projects were summarized. As shown in Table 5, the No. 9 statistical feature "Statistics of waste generation" (by time) in the group "waste disposal behavior" was the most critical characteristic; the second-tier characteristics included one "truck usage behavior" (No. 2), one "waste disposal behavior" (No. 8) and two "facility usage behaviors" (Nos. 14 and 19). The major parameters in the ML method were analyzed, and the results of tests were preliminarily validated by the true values from clients.

The contribution of this paper is two-fold. From the theoretical perspective, it pinpoints that waste big data has the potential to articulate projects' waste generation behaviors for more value-added applications. From the CWM practitioner's perspective, the presented handling method is an automatic, fast, and low-cost pipeline to complete MNAR data and understand the projects' behaviors. Meanwhile, the more suitable and consistent classification category is beneficial to decrease the confusion due to the information barriers and improve the completion and efficiency of the big data system in CWM. The recognized and summarized waste generation behaviors of the two construction types, like the truck usage and the usage preference of some specific facilities, could also help CWM practitioners and policymakers to have a deeper understanding of the waste generation characteristics, which further benefits the rulemaking.

Future work to extend this study lies in several aspects. First, missing data having three or more values should be studied to validate the methodological scalability. Furthermore, interpretable ML should be studied besides correctness, such as the F_1 scores. The balance of predictions across different target values is another issue to investigate. Meanwhile, researchers should be reminded to pay attention to ethical and privacy issues while processing more information.

Acknowledgments

The work presented in this paper was supported by the Policy Innovation and Co-ordination Office (PICO) of the Government of Hong Kong SAR under the Strategic Public Policy Research (SPPR) (No.: S2018.A8.010.18S). The funding agency had no role in research design, data collection and analysis, decision to publish, or preparation, of the manuscript.

Data Availability

The training datasets, including the full set consisting of 821 features and those selected for the four ML methods, are provided as supplemental materials of this paper. Other data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request.

References

Abbasi, M. & El Hanandeh, A. (2016). Forecasting municipal solid waste generation using artificial intelligence modelling approaches. *Waste management, 56*, 13-22. doi:10.1016/j.wasman.2016.05.018

Abidin, N. Z., Ismail, A. R. & Emran, N. A. (2018). Performance analysis of machine learning algorithms for missing value imputation. *International Journal of Advanced Computer Science and Applications (IJACSA), 9*(6), 442-447. doi:10.14569/IJACSA.2018.090660

Alcay, A., Montañés, A. & Simón-Fernández, M. B. (2020). Waste generation in Spain. Do Spanish regions exhibit a similar behavior? *Waste Management, 112*, 66-73. doi:10.1016/j.wasman.2020.05.029

Atkin, B. & Skitmore, M. (2008). Stakeholder management in construction. *Construction Management and Economics, 26*(6), 549-552. doi:10.1080/01446190802142405

Bakshan, A., Srour, I., Chehab, G., El-Fadel, M. & Karaziwan, J. (2017). Behavioral determinants towards enhancing construction waste management: A Bayesian Network analysis. *Resources, Conservation and Recycling, 117*, 274-284. doi:10.1016/j.resconrec.2016.10.006

Baraldi, A. N. & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of school psychology, 48*(1), 5-37. doi:10.1016/j.jsp.2009.10.001

Begum, R. A., Siwar, C., Pereira, J. J. & Jaafar, A. H. (2009). Attitude and behavioral factors in waste management in the construction industry of Malaysia. *Resources, Conservation and Recycling, 53*(6), 321-328. doi:10.1016/j.resconrec.2009.01.005

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., ... & Pasha, M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. *Advanced engineering informatics, 30*(3), 500-521. doi:10.1016/j.aei.2016.07.001

Callistus, T. & Clinton, A. (2016). Evaluating Barriers to Effective Implementation of Project Monitoring and Evaluation in the Ghanaian Construction Industry. *Procedia engineering, 164*, 389-394. doi:10.1016/j.proeng.2016.11.635

Chen, J., Lu, W. & Xue, F. (2021). "Looking beneath the surface": A visual-physical feature hybrid approach for unattended gauging of construction waste composition. *Journal of Environmental Management, 286*, 112233. doi:10.1016/j.jenvman.2021.112233

Corsini, F., Gusmerotti, N. M., Testa, F. & Iraldo, F. (2018). Exploring waste prevention behaviour through empirical research. *Waste Management, 79*, 132-141. doi:10.1016/j.wasman.2018.07.037

Dietterich, T. (2000). Ensemble Methods in Machine Learning. *International workshop on multiple classifier systems* (pp. 1-15). Berlin: Springer. doi:10.1007/3-540-45014-9_1

Ding, Y. & Simonoff, J. S. (2010). An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research, 11*(1), 131-170. doi:abs/10.5555/1756006.1756012

Eastman, C. M., Eastman, C., Teicholz, P., Sacks, R. & Liston, K. (2011). *BIM handbook: A guide to building information modeling for owners, managers, designers, engineers and contractors* (2nd ed.). John Wiley & Sons.

EPD. (2020). *Monitoring of Solid Waste in Hong Kong: Waste Statistics for 2019.* Hong Kong: Environmental Protection Department, Government of Hong Kong SAR. Retrieved from https://www.wastereduction.gov.hk/en/assistancewizard/waste_red_sat.htm

García-Laencina, P. J., Sancho-Gómez, J. L., Figueiras-Vidal, A. R. & Verleysen, M. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing, 72*(7-9), 1483-1493. doi:10.1016/j.neucom.2008.11.026

Garciarena, U. & Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications, 89*, 52-65. doi:10.1016/j.eswa.2017.07.026

Gavankar, S. & Sawarkar, S. (2015). Decision Tree: Review of Techniques for Missing Values at Training, Testing and Compatibility. *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)* (pp. 122-126). Kota Kinabalu: IEEE. doi:10.1109/AIMS.2015.29

Ge, Z., Song, Z., Ding, S. X. & Huang, B. (2017). Data mining and analytics in the process industry: The role of machine learning. *Ieee Access, 5*, 20590-20616. doi:10.1109/ACCESS.2017.2756872.

Guerrero, L. A., Maas, G. & Hogland, W. (2013). Solid waste management challenges for cities in developing countries. *Waste management, 33*(1), 220-232. doi:10.1016/j.wasman.2012.09.008

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications* (pp. 18-28). IEEE. doi:10.1109/5254.708428

Ishibuchi, H., Miyazaki, A., Kwon, K. & Tanaka, H. (1993). Learning from incomplete training data with missing values and medical application. *Proceedings of 1993 International Conference on Neural Networks(IJCNN-93-Nagoya, Japan)*, (pp. 1871-1874). Nagoya, Japan. doi:10.1109/IJCNN.1993.717020

Jain, A. K., Duin, R. P. & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence, 22*(1), 4-37. doi:10.1109/34.824819

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M. & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine, 50*(2), 105-115. doi:10.1016/j.artmed.2010.05.002

Jia, Z. (2019). *A Nonparametric Multiple Imputation Approach For MNAR Mechanism Using the Sample Selection Model Framework.* Arizona: The University of Arizona. Retrieved from http:arizona.edu/handle/10150/632554

Johnson, N. E., Ianiuk, O., Cazap, D., Liu, L., Starobin, D., Dobler, G. & Ghandehari, M. (2017). Patterns of waste generation: A gradient boosting model for short-term waste prediction in New York City. *Waste management, 62*, 3-11. doi:10.1016/j.wasman.2017.01.037

Kagioglou, M., Cooper, R., Aouad, G. & Sexton, M. (2000). Rethinking construction: the generic design and construction process protocol. *Engineering, construction and architectural management, 7*(2), 141-153. doi:10.1108/eb021139

Karadimas, N. V. & Loumos, V. G. (2008). GIS-based modelling for the estimation of municipal solid waste generation and collection. *Waste Management & Research, 26*(4), 337-346. doi:10.1177/0734242X07081484

Kerzner, H. (2017). *Project management: a systems approach to planning, scheduling, and controlling* (12th ed.). John Wiley & Sons.

Keser, S., Duzgun, S. & Aksoy, A. (2012). Application of spatial and non-spatial data analysis in determination of the factors that impact municipal solid waste generation rates in Turkey. *Waste management, 32*(3), 359-371. doi:10.1016/j.wasman.2011.10.017

Koller, D. & Saharni, M. (1996). *Toward Optimal Feature Selection.* Stanford InfoLab. Retrieved from http://ilpubs.stanford.edu:8090/208/

Kontokosta, C. E., Hong, B., Johnson, N. E. & Starobin, D. (2018). Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities. *Computers, Environment and Urban Systems, 70*, 151-162. doi:10.1016/j.compenvurbsys.2018.03.004

Laaksonen, J. & Oja, E. (1996). Classification with learning k-nearest neighbors. *Proceedings of International Conference on Neural Networks (ICNN'96)* (pp. 1480-1483). Washington, DC: IEEE. doi:10.1109/ICNN.1996.549118

Li, R. P., Mukaidono, M. & Turksen, I. B. (2002). A fuzzy neural network for pattern classification and feature selection. *Fuzzy Sets and Systems, 130*(1), 101-108. doi:10.1016/S0165-0114(02)00050-7

Lingard, H., Graham, P. & Smithers, G. (2000). Employee perceptions of the solid waste management system operating in a large Australian contracting organization: implications for company policy implementation. *Construction Management & Economics, 18*(4), 383-393. doi:10.1080/01446190050024806

Little, R. J. & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). New York: John Wiley & Sons. doi:10.1002/9781119482260

Love, P. E., Holt, G. D., Shen, L. Y., Li, H. & Irani, Z. (2002). Using systems dynamics to better understand change and rework in construction project management systems. *International journal of project management, 20*(6), 425-436. doi:10.1016/S0263-7863(01)00039-4

Lu, W. (2019). Big data analytics to identify illegal construction waste dumping: A Hong Kong study. *Resources, Conservation and Recycling, 141*, 264-272. doi:10.1016/j.resconrec.2018.10.039

Lu, W., Lee, W. M., Xue, F. & Xu, J. (2021). Revisiting the effects of prefabrication on construction waste minimization: A quantitative study using bigger data. *Resources, Conservation and Recycling, 170*, 105579. doi:10.1016/j.resconrec.2021.105579

Luangcharoenrat, C., Intrachooto, S., Peansupap, V. & Sutthinarakorn, W. (2019). Factors influencing construction waste generation in building construction: Thailand's perspective. *Sustainability, 11*(13), 3638. doi:10.3390/su11133638

Luo, L., He, Q., Jaselskis, E. J. & Xie, J. (2017). Construction project complexity: research trends and implications. *Journal of construction engineering and management, 143*(7), 04017019. doi:10.1061/(ASCE)CO.1943-7862.0001306

Ma, J., Cheng, J. C., Jiang, F., Chen, W., Wang, M. & Zhai, C. (2020). A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data. *Energy and Buildings, 216*, 109941. doi:10.1016/j.enbuild.2020.109941

Ma, M., Tam, V. W., Le, K. N. & Li, W. (2020). Challenges in current construction and demolition waste recycling: A China study. *Waste Management, 118*, 610-625. doi:10.1016/j.wasman.2020.09.030

Márquez, M. Y., Ojeda, S. & Hidalgo, H. (2008). Identification of behavior patterns in household solid waste generation in Mexicali's city: Study case. *Resources, Conservation and Recycling, 52*(11), 1299-1306. doi:10.1016/j.resconrec.2008.07.011

Mattar, L., Abiad, M. G., Chalak, A., Diab, M. & Hassan, H. (2018). Attitudes and behaviors shaping household food waste generation: Lessons from Lebanon. *Journal of Cleaner Production, 198*, 1219-1223. doi:10.1016/j.jclepro.2018.07.085

McKnight, P. E., McKnight, K. M., Sidani, S. & Figueredo, A. J. (2007). *Missing data: A gentle introduction.* New York: Guilford Press. Retrieved from https://psycnet.apa/record/2007-06639-000

Meeyai, S. (2016). Logistic Regression with Missing Data: A Comparison of Handling Methods, and Effects of Percent Missing Values. *Journal of Traffic and Logistics Engineering, 4*(2). doi:10.18178/JTLE.4.2.128-134

Meza, J. K., Yepes, D. O., Rodrigo-Ilarri, J. & Cassiraga, E. (2019). Predictive analysis of urban waste generation for the city of Bogotá, Colombia, through the implementation of decision trees-based machine learning, support vector machines and artificial neural networks. *Heliyon, 5*(11), e02810. doi:10.1016/j.heliyon.2019.e02810

Mitchell, T. M. (2006). *The discipline of machine learning (Vol. 9).* Pittsburgh: Carnegie Mellon University. Retrieved from https://web.cs.wpi.edu/~kmlee/cs539/MachineLearning

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A. & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society, 18*(6), 275-285. doi:10.1002/cem.873

N.Baraldi, A. & K.Enders, C. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5-37. doi:10.1016/j.jsp.2009.10.001

Nanni, L., Lumini, A. & Brahnam, S. (2012). A classifier ensemble approach for the missing feature problem. *Artificial intelligence in medicine, 55*(1), 37-50. doi:10.1016/j.artmed.2011.11.006

Nugroho, H., Utama, N. P. & Surendro, K. (2020). Comparison Method for Handling Missing Data in Clinical Studies. *Proceedings of the 2020 9th International Conference on Software and Computer Applications* (pp. 46–50). New York: ICSCA 2020. doi:10.1145/3384544.3384594

Olander, S. (2007). Stakeholder impact analysis in construction project management. *Construction management and economics, 25*(3), 277-287. doi:10.1080/01446190600879125

Osisanwo, F. Y., Awodele, O., Hinmikaiye, J. O., Olakanmi, O. & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT), 48*(3), 128-138. doi:10.14445/22312803/IJCTT-V48P126

Osmani, M., Glass, J. & Price, A. D. (2008). Architects' perspectives on construction waste reduction by design. *Waste Management, 28*(7), 1147-1158. doi:10.1016/j.wasman.2007.05.011

Paslawski, J. (2017). Flexible approach for construction process management under risk and uncertainty. *Procedia engineering, 208*, 114-124. doi:10.1016/j.proeng.2017.11.028

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesna. (2011). Scikit-learn: Machine Learning in

Python. *Journal of Machine Learning Research, 12*, 2825--2830. doi:10.5555/1953048.2078195

Peeters, M., Zondervan-Zwijnenburg, M., Vink, G. & Van de Schoot, R. (2015). How to handle missing data: A comparison of different approaches. *European Journal of Developmental Psychology, 12*(4), 377-394. doi:10.1080/17405629.2015.1049526

Perkowski, E. (2020). *Impact of ensemble machine learning methods on handling missing data.* Twente: University of Twente. Retrieved from http://purl.utwente.nl/essays/82210

Peugh, J. L. & Enders, C. K. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research, 74*(4), 525-556. doi:10.3102/00346543074004525

Pierce, W. D. & Cheney, C. D. (2017). *Behavior Analysis and Learning, Sixth Edition (6th ed.).* New York: Routledge. doi:10.4324/9781315200682

Pigott, T. D. (2010). A Review of Methods for Missing Data. *Educational Research and Evaluation , 7*(4), 353-383. doi:10.1076/edre.7.4.353.8937

Poon, C. S., Yu, A. T., Wong, S. W. & Cheung, E. (2004). Management of construction waste in public housing projects in Hong Kong. *Construction Management & Economics, 22*(7), 675-689. doi:10.1080/0144619042000213292

Rabe, B. A., Day, S., Fiero, M. H. & Bell, M. L. (2018). Missing data handling in non-inferiority and equivalence trials: A systematic review. *Pharmaceutical statistics, 17*(5), 477-488. doi:10.1002/pst.1867

Rahman, M. G. & Islam, M. Z. (2013). Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques. *Knowledge-Based Systems, 53*, 51-65. doi:10.1016/j.knosys.2013.08.023

Rimaitytė, I., Ruzgas, T., Denafas, G., Račys, V. & Martuzevicius, D. (2012). Application and evaluation of forecasting methods for municipal solid waste generation in an eastern-European city. *Waste Management & Research, 30*(1), 89-98. doi:10.1177/0734242X10396754

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592. doi:10.1093/biomet/63.3.581

Sattari, M. T., Rezazadeh-Joudi, A. & Kusiak, A. (2017). Assessment of different methods for estimation of missing data in precipitation studies. *Hydrology Research, 48*(4), 1032-1044. doi:10.2166/nh.2016.364

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* New York: Chapman and Hall/CRC. doi:10.1201/9780367803025

Shylaja, B. & Kumar, R. S. (2018). Traditional versus modern missing data handling techniques: An overview. *International Journal of Pure and Applied Mathematics, 118*(14), 77-84. Retrieved from https://acadpubl.eu/jsi/2018-118-14-15/articles/14/12.pdf

Singh, A., Thakur, N. & Sharma, A. (2016). A review of supervised machine learning algorithms. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1310-1315). New Delhi: IEEE. Retrieved from https://ieeexplore.ieee.org/abstract/document/7724478

Soibelman, L. & Kim, H. (2002). Data preparation process for construction knowledge generation through knowledge discovery in databases. *Journal of Computing in Civil Engineering, 16*(1), 39-48. doi:10.1061/(ASCE)0887-3801(2002)16:1(39)

Song, Q., Shepperd, M., Chen, X. & Liu, J. (2008). Can k-NN imputation improve the performance of C4. 5 with small software project data sets? A comparative evaluation. *Journal of Systems and software, 81*(12), 2361-2370. doi:10.1016/j.jss.2008.05.008

Sportisse, A., Boyer, C. & Josses, J. (2020). Estimation and imputation in probabilistic principal component analysis with missing not at random data. *34th Conference on Neural Information Processing Systems.* Canada: NeurIPS 2020. Retrieved from https://proceedings.neurips.cc/paper/2020/file/4ecb679fd35dcfd0f0894c399590be1a-Paper.pdf

Tam, V. W. & Tam, C. M. (2006). Evaluations of existing waste recycling methods: a Hong Kong study. *Building and Environment, 41*(12), 1640-1660. doi:10.1016/j.buildenv.2005.06.017

Tamayo-Orbegozo, U., Vicente-Molina, M. A. & Villarreal-Larrinaga, O. (2017). Eco-innovation strategic model. A multiple-case study from a highly eco-innovative European region. *Journal of Cleaner Production, 142*, 1347-1367. doi:10.1016/j.jclepro.2016.11.174

Tang, J., Alelyani, S. & Liu, H. (2014). Feature selection for classification: A review. *Data Classification: Algorithms and Applications* (pp. 37-64). CRC Press. doi:10.1201/b17320

Tang, Z., Li, W., Tam, V. W. & Xue, C. (2020). Advanced progress in recycling municipal and construction solid wastes for manufacturing sustainable construction materials. *Resources, Conservation & Recycling: X, 6*, 100036. doi:10.1016/j.rcrx.2020.100036

Tonglet, M., Phillips, P. S. & Bates, M. P. (2004). Determining the drivers for householder pro-environmental behaviour: waste minimisation compared to recycling. *Resources, conservation and recycling, 42*(1), 27-48. doi:10.1016/j.resconrec.2004.02.001

Tran, C. T. (2017). Multiple imputation and ensemble learning for classification with incomplete data. In *Intelligent and Evolutionary Systems* (pp. 401-415). Springer. doi:10.1007/978-3-319-49049-6_29

Twala, B. & Cartwright, M. (2010). Ensemble Missing Data Techniques for Software Effort Prediction. *Intelligent Data Analysis, 14*(3), 299-331. doi:10.3233/IDA-2010-0423

Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence, 23*(5), 373-405. doi:10.1080/08839510902872223

UNEP. (2019). *Sand and sustainability: Finding new solutions for environmental governance of global sand resources.* Geneva, Switzerland: United Nations Environment Programme. Retrieved from https://wedocs.unep.org/handle/20.500.11822/28163

Vafaie, H. & Jong, K. D. (1992). Genetic algorithms as a tool for feature selection in machine learning. *Proceedings Fourth International Conference on Tools with Artificial Intelligence TAI '92. 1*, pp. 200-203. Arlington: IEEE. doi:10.1109/TAI.1992.246402

Watanabe, S. (1985). *Pattern recognition: human and mechanical.* New York: John Wiley & Sons, Inc.

Xu, J., Lu, W., Ye, M., Webster, C. & Xue, F. (2020a). An anatomy of waste generation flows in construction projects using passive bigger data. *Waste Management, 106*, 162-172. doi:10.1016/j.wasman.2020.03.024

Xu, J., Lu, W., Ye, M., Xue, F., Zhang, X. & Lee, B. F. (2020b). Is the private sector more efficient? Big data analytics of construction waste management sectoral efficiency. *Resources, Conservation and Recycling, 155*, 104674. doi:10.1016/j.resconrec.2019.104674

Xue, F., Wu, L. & Lu, W. (2021). Semantic enrichment of building and city information models: A ten-year review. *Advanced Engineering Informatics, 47*, 101245. doi:10.1016/j.aei.2020.101245

You, Z. & Wu, C. (2019). A framework for data-driven informatization of the construction company. *Advanced Engineering Informatics, 39*, 268-277. doi:10.1016/j.aei.2019.02.002

You, Z., Wu, C., Zheng, L. & Feng, L. (2020). An Informatization Scheme for Construction and Demolition Waste Supervision and Management in China. *Sustainability, 12*(4), 1672. doi:10.3390/su12041672

Zhang, G., Lin, T., Chen, S., Xiao, L., Wang, J. & Guo, Y. (2015). Spatial characteristics of municipal solid waste generation and its influential spatial factors on a city scale: a case study of Xiamen, China. *Journal of Material Cycles and Waste Management, 17*(2), 399-409. doi:10.1007/s10163-014-0257-7

Appendix

Table A1. List of 124 decisive features selected for Decision Tree from the pool of 821 features

Group (subtotal)	Feature type id	Feature				
		Daily (*d*)	Weekly (*w*)	Monthly (*m*)	Yearly (*y*)	Total (*t*)
Truck usage behaviors (28)	1	mean	mean	extremum	pct50	s_t^1
	2	mean, extremum	mean	extremum, pct5	mean,pct25,IQR	s_t^2
	3		pct95	mean	max, min, IQR	
	4					
	5	stddev, max	mean, stddev, max	stddev, max, min	pct95	
	6					

Waste disposal behaviors (62)	7		act		act	
	8				*weight*$_n$, n = 11,14, 16	s_t^8
	9	max,pct50, pct5,pct25, pct95, extremum, IQR, stddev	min, pct5,pct25,p ct75, pct95,IQR	mean, min, extremum, pct25,pct75, pct95,IQR	mean, stddev, pct50,extremum, pct75,pct95,IQR	
Facility usage behaviors (34)	10		stddev	mean, IQR		s_t^{10}
	11	max				
	12	pct95	IQR	extremum		
	13	pct75			stddev	
	14	pct25	mean, stddev, pct75	IQR		s_t^{14}
	15	stddev	IQR	mean, max	extremum	
	16	stddev,pct95		stddev, max, extremum	mean, max, pct75	
	17	pct95,IQR	stddev, pct25, pct75	min,pct50, pct5,pct95	min,pct5, pct95,IQR	
	18		pct75,IQR		extremum	
	19	mean,pct95, IQR	max		pct50, extremum	s_t^{19}
	20	max	mean, stddev	stddev,pct95	extremum	s_t^{20}
	21				pct,pct75	s_t^{21}
	Subtotal	26	27	29	34	8

900

Table A2. List of 222 decisive features selected for KNN from the pool of 821 features

Group (subtotal)	Feature type id	Feature				
		Daily (d)	Weekly (w)	Monthly (m)	Yearly (y)	Total (t)
Truck usage behaviors (42)	1				stddev, min, pct50, pct5, pct25	
	2		stddev	mean, stddev, max, pct50, pct75, pct95	mean, stddev, max, min, pct50, extremum, pct5, pct25, pct75, pct95, IQR	
	3			stddev	mean, stddev, max, min, pct50, pct5, pct25, pct75, pct95, IQR	
	4					
	5			stddev, max, extremum	stddev, max, pct50, extremum, pct95	
	6					
Waste disposal behaviors (44)	7				act	
	8				$weight_n$, n = 11,…,16	s_t^8
	9	mean, stddev, max, min, pct50, extremum, pct5, pct25, pct75, pct95, IQR	stddev, max, min, pct50, extremum, pct5, pct25, pct95, IQR	stddev, max, min, extremum, pct5, pct25, pct75, pct95, IQR	mean, stddev, max, extremum, pct75, pct95, IQR	
Facility usage behaviors (136)	10	max	max, pct95	mean, stddev, max, extremum, pct75, pct95	mean, stddev, max, min, pct50, extremum, pct5, pct25, pct75, pct95, IQR	
	11					
	12	stddev, max, extremum, pct95	stddev, max, extremum, pct95	mean, stddev, max, extremum, pct75, pct95, IQR	mean, stddev, max, min, pct50, extremum, pct5, pct25, pct75, pct95	
	13			stddev	stddev, max, pct50, pct95	
	14				stddev	
	15		stddev	stddev	mean, stddev, max, pct50, extremum, pct75, pct95, IQR	
	16			stddev, max, extremum,	mean, stddev, max, pct50, extremum,	

			pct75, pct95, IQR	pct75, pct95, IQR	
17			mean, max, min, pct25, pct95	stddev	
18			stddev	stddev	
19	stddev, max, extremum, pct95, IQR	mean, stddev, max, pct50, extremum, pct75, pct95, IQR	mean, stddev, max, pct50, extremum, pct75, pct95, IQR	mean, stddev, max, min, pct50, pct5, pct25, pct75, pct95, IQR	
20		stddev	stddev	mean, stddev, max, min, pct50, pct5, pct25, pct75, pct95	
21			mean, max, pct75, pct95, IQR	mean, stddev, max, pct50, pct75, pct95	
Subtotal	21	26	60	114	1

Table A3. List of 378 decisive features selected for SVM from the pool of 821 features

Group (subtotal)	Feature type id	Feature				
		Daily (*d*)	Weekly (*w*)	Monthly (*m*)	Yearly (*y*)	Total (*t*)
Truck usage behaviors (105)	1	mean, stddev, max, extremum, pct75, pct95, IQR	stddev, max, min, extremum, pct5, pct25, pct95, IQR	stddev, max, pct50, extremum, pct75, pct95	mean, stddev, max, min, pct50, extremum, pct5, pct75, pct95, IQR	
	2	stddev, min, pct5, pct25	stddev, pct50, pct25, pct95	mean, stddev		
	3	mean, stddev, min, pct50, pct5, pct25, pct75, IQR	mean, stddev, min, pct50, pct5, pct25, pct75, pct95, IQR	mean, stddev, max, min, pct50, extremum, pct5, pct25, pct75, IQR	mean, stddev, max, min, pct50, extremum, pct5, pct25, pct75, pct95, IQR	
	4	min, pct50, pct5, pct25	min, pct50, pct5, pct25	min, pct50, pct5, pct25	pct50	s_t^4
	5	stddev	stddev, pct25	stddev, min, pct5, pct25	stddev, min, pct50, pct5, pct25	
	6					
Waste disposal behaviors (23)	7		act	act	act	
	8				$weight_n$, n = 13,15,16	
	9	stddev, min, pct50, pct5, pct25, pct75, IQR	mean, stddev, min, pct50, pct25, pct75, pct95, IQR	mean, stddev, min, pct50, pct5, pct25, pct75, IQR	stddev, min, pct5, pct25	
Facility usage behaviors (250)	10	min, pct5, pct25, IQR	mean, stddev, min, pct50, pct5, pct25, pct75, IQR	mean, stddev, max, min, extremum, pct5, pct25, pct75, pct95, IQR	mean, stddev, max, min, pct50, extremum, pct5, pct25, pct75, pct95, IQR	
	11	mean, stddev, max, pct50, extremum, pct25, pct75, pct95, IQR	mean, stddev, max, pct50, extremum, pct25, pct75, pct95, IQR	mean, stddev, max, min, pct50, extremum, pct5, pct25, pct75, pct95, IQR	mean, stddev, max, min, pct50, extremum, pct5, pct25, pct75, pct95, IQR	
	12	mean, stddev, min, pct50, extremum, pct25, pct75, pct95, IQR	mean, stddev, min, pct50, pct5, pct25, pct75, pct95, IQR	mean, stddev, max, min, pct50, extremum, pct5, pct25, pct75, pct95, IQR	mean, stddev, max, min, pct50, extremum, pct5, pct25, pct75, IQR	s_t^{12}
	13	mean, min, pct50, pct5, pct25	mean, stddev, min, pct50, pct5,	mean, stddev, min, pct50,		

			pct25, pct75, IQR	pct5, pct25, pct75		
	14	mean, min, pct50, pct5, pct25, IQR	min, pct5, pct25			s_t^{14}
	15	stddev, max, min, extremum, pct5, pct25	stddev, max, min, extremum, pct5	stddev, min, pct5, pct25	mean, stddev, min, pct5, pct25	s_t^{15}
	16	mean, pct50, pct75	pct50	min, pct5, pct25	min, pct5	s_t^{16}
	17	pct50, pct25	pct50	pct25	stddev	s_t^{17}
	18	stddev, max, min, extremum, pct5, IQR	stddev, pct95	max, extremum	stddev, min, pct50, pct5, pct25	
	19	mean, stddev, min, pct50, pct5, pct25, pct75, IQR	mean, stddev, min, pct50, pct5, pct25, pct75, pct95, IQR	mean, stddev, min, pct50, pct5, pct25, pct75, IQR	stddev, min, pct5, pct25, pct75, IQR	s_t^{19}
	20	min, pct50, pct5, pct25	min, pct5, pct25	min	stddev	
	21	pct50, pct25		pct25, pct75		
	Subtotal	95	94	95	87	7

Table A4. List of 162 decisive features selected for Neural-network from the pool of 821 features

Group (subtotal)	Feature type id	Feature				
		Daily (*d*)	Weekly (*w*)	Monthly (*m*)	Yearly (*y*)	Total (*t*)
Truck usage behaviors (43)	1			stddev	stddev, max, extremum, pct75, pct95, IQR	
	2		stddev	stddev	max, min, extremum, pct5, pct25, pct75, IQR	
	3	stddev	mean, stddev, max	extremum	stddev, max, pct95	
	4			stddev	stddev, max, pct75, IQR	
	5	max, extremum	stddev, max, extremum	stddev	stddev, max, min, extremum, pct5, pct75, pct95	
	6					s_t^6
Waste disposal behaviors (88)	7		act			
	8				$weight_n$, n = 11,12,14,16	s_t^8
	9	mean, stddev, pct5, pct25, pct75, pct95	mean, pct50, extremum, pct75, pct95, IQR	mean, stddev, max, pct50, extremum, pct75, pct95, IQR	mean, stddev, max, pct50, extremum	
Facility usage behaviors (31)	10		stddev, max, extremum	stddev, max, extremum	stddev, max, extremum, pct75, pct95	
	11				stddev	
	12	max, extremum	stddev, max, extremum	stddev, max, extremum, pct75, IQR	stddev, max, min, pct50, extremum, pct5, pct25, pct95	
	13	stddev	stddev	stddev, extremum	stddev, max, min, pct5, pct25, pct95, IQR	
	14		stddev	stddev, pct95	mean, stddev, max, extremum, pct75, pct95, IQR	
	15				stddev	
	16			stddev	stddev, min, pct50	
	17			stddev	max	
	18					
	19	stddev, max, extremum	stddev, max, extremum	stddev, max, extremum, pct75, IQR	mean, stddev, max, min, extremum, pct5,	

					pct25, pct75, pct95, IQR	
	20			stddev	mean, stddev, max, extremum, pct75, pct95, IQR	
	21				stddev	
	Subtotal	15	25	33	87	2

910

Table A5. Average performance metrics of the ML models in 10-fold cross-validation, where the best value in each row is in bold

		Decision tree
No. of selected features		162
Precision	New building	0.77
	Renovation	0.89
	Overall	0.83
Recall	New building	0.66
	Renovation	0.93
	Overall	0.80
F_1 score	New building	0.71
	Renovation	0.91
	Overall	0.81

915 Table A6. List of 45 decisive features selected for Bayesian Networks from the pool of 821 features

Group (subtotal)	Feature type id	Feature				
		Daily (*d*)	Weekly (*w*)	Monthly (*m*)	Yearly (*y*)	Total (*t*)
Truck usage behaviors (17)	1				stddev	
	2		stddev	stddev	max, extremum, pct75, pct95, IQR	
	3			stddev	stddev	
	4			stddev	stddev	
	5			stddev	max, extremum, pct95	s_t^5
	6					
Waste disposal behaviors (20)	7					
	8				$weight_n$, n = 11,15,16	
	9	IQR	stddev	pct75, IQR	stddev	
Facility usage behaviors (8)	10					
	11					
	12			stddev	stddev	
	13			stddev	stddev, max, extremum, pct75, pct95, IQR	
	14					
	15				stddev	
	16				stddev	
	17					
	18				stddev	
	19			stddev	stddev, max, extremum, pct75, pct95, IQR	
	20					
	21				stddev	
	Subtotal	1	2	9	31	1

Table A7. Average performance metrics of the ML models in 10-fold cross-validation, where the best value in each row is in bold

		Decision tree
No. of selected features		45
Precision	New building	0.34
	Renovation	0.98
	Overall	0.66
Recall	New building	0.86
	Renovation	0.81
	Overall	0.84
F_1 score	New building	0.49
	Renovation	0.89
	Overall	0.74