Estimation of construction waste composition based on bulk density: A big data-probability (BD-P) model

Liang Yuan, Weisheng Lu*, and Fan Xue

Department of Real Estate and Construction, Faculty of Architecture, the University of Hong Kong, Hong Kong, China (*: corresponding author email: wilsonlu@hku.hk)

This is the peer-reviewed post-print version of the paper:
Yuan, L., Lu, W., & Xue, F. (2021). Estimation of construction waste composition based on bulk density: a big data-probability (BD-P) model. *Journal of Environmental Management*, 292, Article ID 112822. Doi: 10.1016/j.jenvman.2021.112822.
The final version of this paper is available at: https://doi.org/10.1016/j.jenvman.2021.112822.

Abstract

Estimating the composition of construction waste is crucial to the efficient operation of various waste management facilities, such as landfills, public fills, and sorting plants. However, this estimating task is often challenged by the desire of quickness and accuracy in real-life scenarios. By harnessing a valuable data set in Hong Kong, this research develops a big data-probability (BD-P) model to estimate construction waste composition based on bulk density. Using a saturated data set of 4.27 million truckloads of construction waste, the probability distribution of construction waste bulk density is derived, and then, based on the Law of Joint Probability, the BD-P model is developed. A validation experiment using 604 ground truth data entries indicates a model accuracy of 90.2%, Area Under Curve (AUC) of 0.8775, and speed of around 52 seconds per load in estimating the composition of each incoming construction waste load. The BD-P model also informed a linear model which can perform the estimation with an accuracy of 88.8% but consuming 0.4 seconds per case. The major novelty of this research is to harmonize big data analytics and traditional probability theories in improving the classic challenge of predictive analyses. In the practical sphere, it satisfactorily solves the construction waste estimation problem faced by many waste management facility operators. In the academic sphere, this research provides a vivid example that big data and theories are not adversaries, but allies.

Keywords: Big Data; Probability; Big Data Enabled Probabilistic Analysis; Construction Waste; Composition Estimation

1. Introduction

Construction waste, sometimes also called construction and demolition (C&D) waste, is the solid waste generated from site clearance, soil excavation, new building, refurbishment, renovation, demolition, and other construction activities (HKEPD, 2019; Lu et al., 2020). It is often composed of both inert and non-inert construction materials (UK DEFRA, 2020;

Australian Government, 2011; HKEPD, 2008). The former includes clay, earth, concrete, rubble, and bitumen, while the latter includes wood, bamboo, paper, plastics, and vegetation (Aslam et al., 2020; EPA, 1997; EU, 2018). Prevalent construction waste management (CWM) systems normally stipulate different disposal destinations of C&D waste depending on its composition. For example, in the U.S., most C&D waste is lawfully destined for disposal in landfills regulated under Title 40 of the Code of Federal Regulations (CFR) (Allegri, 1986). In Hong Kong, non-inert waste goes to landfills or incinerators, inert waste to public fills or recycling plants, and mixed waste to sorting facilities (HKEPD, 2008). It is common that the facilities will have to determine admissibility or chargeable waste disposal levy, depending on waste composition. For example, in the Australian state of New South Wales, the Protection of the Environment Operations (Waste) Regulation (NSWEPA, 2020) and the Waste Levy Guidelines (NSWEPA, 2014) set out waste composition calculations and waste levy details. Recognition of waste composition also needs to be done swiftly, since many users are queuing outside the facilities. Therefore, quick and accurate estimation of construction waste composition is key to these CWM facilities or even the entire CWM system.

Previous methods for estimating waste composition can be roughly classified into three categories: statistical sampling, photogrammetry, and other non-invasive approaches. The statistical sampling approach classifies, weighs, and calculates the percentage of each type of waste material in a sample. It is widely used to analyze construction waste component characteristics in a specific region (Asgari et al., 2017; Cochran et al., 2007; Hoang et al., 2020), different project types (Villoria Sáez et al., 2012), or different construction stages (Wu et al., 2019). However, a drawback of the approach is its inefficiency due to requiring onerous manual operation. The photogrammetry-based approach uses algorithms to deal with the collected visual imagery of construction waste for recognizing its composition. Visual features recorded, measured and interpreted in photogrammetry include shape, pattern, gradation, and size (Chen et al., 2021; Califice et al., 2013; Paine & Kiser, 2012). The photogrammetry-based approach had received attention around ten years ago (Wagland et al., 2012) and is still topical nowadays (Davis et al., 2021) for academic studies and practical applications. The advantage of such approach is to reduce the onerous manual operation (Peddireddy et al., 2015; Wagland et al., 2012), but the downside is that construction waste must be spread evenly to a depth (generally no greater than 30 cm) for waste components to be visually recognised (Wagland et al., 2012). Based on an assumption that the visible surface composition reflects the total one, Chen et al., (2021) proposed a depth-controlling-free photogrammetry approach to gauge the composition by looking at the surface of truck-loaded construction waste dumps. However, more proofs are needed to prove the reliability and universality of the assumed condition. Other non-invasive approaches to waste composition estimation have also been explored, e.g., using fluorescence presence, microwave frequency measurement, and X-ray imaging (Vrancken et al., 2017). While these approaches can capture component type information, they lack accuracy in recording component weight and volume.

This research aims to develop a quick and accurate approach to estimating construction waste composition. It is a direct response to a real-life challenge as faced by the HKEPD (Environmental Protection Department of Hong Kong), but we found parallels in many other places such as India, Taiwan, and the UK wherein their CWM facilities also need to estimate the composition of incoming waste bulks/loads quickly and accurately (see Figure 1).

Particularly, HKEPD is responsible for managing all C&D waste in the territory. It uses a tiered system whereby all construction waste must be disposed of at governmental waste disposal facilities, unless it is properly reused or recycled. A truckload of construction waste is sent to public fills or a recycling plant if it contains purely inert materials, to a sorting plant if it contains more than 50% inert materials by weight, or otherwise to landfill (Lu et al., 2016). Waste producers are charged a landfill disposal fee of HK$200 per tonne, a sorting plant fee of HK$175 per tonne, and HK$71 per tonne at public fills. Different levies and disposal destinations generate the need for construction waste composition estimation. Based on the empirical knowledge about the relationship between waste weight and composition, the HKEPD uses a 'quick and dirty' approach to screen waste loads at its CWM facilities, translating the weight and height of each load into a single index similar to a body mass index (BMI). This approach, however, has been found problematic by a Hong Kong Audit Commission review report (HKAC, 2016). It only had an accuracy of around 60%.

Figure 1. Equipment used to measure incoming waste loads (Sources: a: made by authors; b: https://bit.ly/2MgdREa; c: https://bit.ly/399WQoh; d: https://bit.ly/3c2Jqfs;)

Recent advancements in big data analytics suggest the potential for a solution to the waste composition recognition problem. The promise of big data has been widely documented. For example, it reveals that construction waste in a region is not entirely random in terms of its composition (Lu et al., 2021). Constrained by the prevailing construction materials, technologies, and waste recycling practices, the bulk density of different types of construction waste should converge into an interval and follow a certain probability distribution pattern. If we know the pattern, we should be able to tell the composition of an incoming truckload of construction waste by its bulk density. A strength of big data is its ability to reach a closer truth on the ground (LaValle et al., 2011; Lu et al., 2015). Having acquired a set of data concerning millions of truckloads of construction waste disposed of in Hong Kong over the past ten years, we are able to leverage this strength. The rest of the paper is organized as follows. Section 2 introduces bulk density and its probability distribution as informed by big data. Section 3 introduces the research methodology including the theoretical foundation and research methods. The composition estimation model is explained in Section 4. Section 5 describes the model validation experiment. Section 6 discusses the strengths, novelty, prospects, and challenges of the approach, and conclusions are drawn in Section 7.

2. Bulk density and its probability distribution

Bulk density is the mass of material divided by the total volume it occupies, where the total volume can include particle volume, inter-particle void volume, and internal pore volume (Lyon & Buckman, 1922). True density, apparent density, and bulk density are the three kinds of density in materials science. Figure 2 shows their differences and how they are calculated. Bulk density is not an intrinsic material property. Rather, it is changeable in line with constitutional material, pore, and inter-particle volume. Because of this, the bulk densities of construction waste appear to be totally random since the constitutional materials could be any combination of construction materials (e.g., concrete, clay, soil, rocks, paper, wood, vegetation), with any size of pore voids and inter-particle voids. Nevertheless, our hypothesis is that bulk density in a specific region should converge into an interval and follow a certain probability distribution pattern, as it is determined by prevailing regional construction materials, technologies, and waste recycling practices.

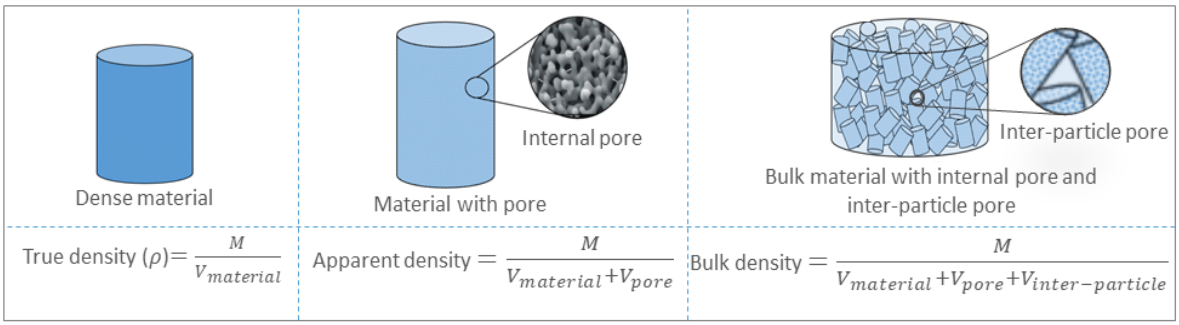

Figure 2. The three types of material densities and their calculation methods

In practice, many CWM facilities have devised methods to measure the mass and volume of waste loads on arrival. Figure 1 illustrates some equipment used. Specifically, at HKEPD's facilities (Figure 1a), the (net) weight of a load of C&D waste is calculated by weighing the vehicle at the in- and out-weighbridges and subtracting the two. Waste volume is captured and calculated using sensors above the in-weighbridge. Bulk density is then calculated using Equation (1):

$$\rho = \frac{W}{V} \tag{1}$$

where ρ is the bulk density of a load of construction waste, W is the net weight, and V is the total volume. As mentioned above, the concept is similar to that of BMI, which divides body mass by the square of height. Despite having its critics, BMI is an easy-to-obtain and comprehend indicator to measure whether a person is underweight, normal weight, overweight or obese.

Using a big data set of 4.27 million truckloads of construction waste recorded by the HKEPD from 2017 to 2020, the convergent bulk density intervals of inert and non-inert C&D waste materials, and their probability distributions are derived as shown in Figure 3. The bulk density of inert construction waste ranges from 0.207 tonnes/m^3 to 2.435 tonnes/m^3, and highly concentrates at around 1.300 tonnes/m^3 and 1.800 tonnes/m^3. The bulk density of non-inert construction waste is between around 0.045 tonnes/m^3 and 2.434 tonnes/m^3, and highly concentrates at around 0.300 tonnes/m^3. Although the exact cause of multiple crests is yet to be fully understood, the big data paints a clear distribution pattern of bulk densities of construction waste. This can be harnessed to estimate the composition of a new incoming waste load.

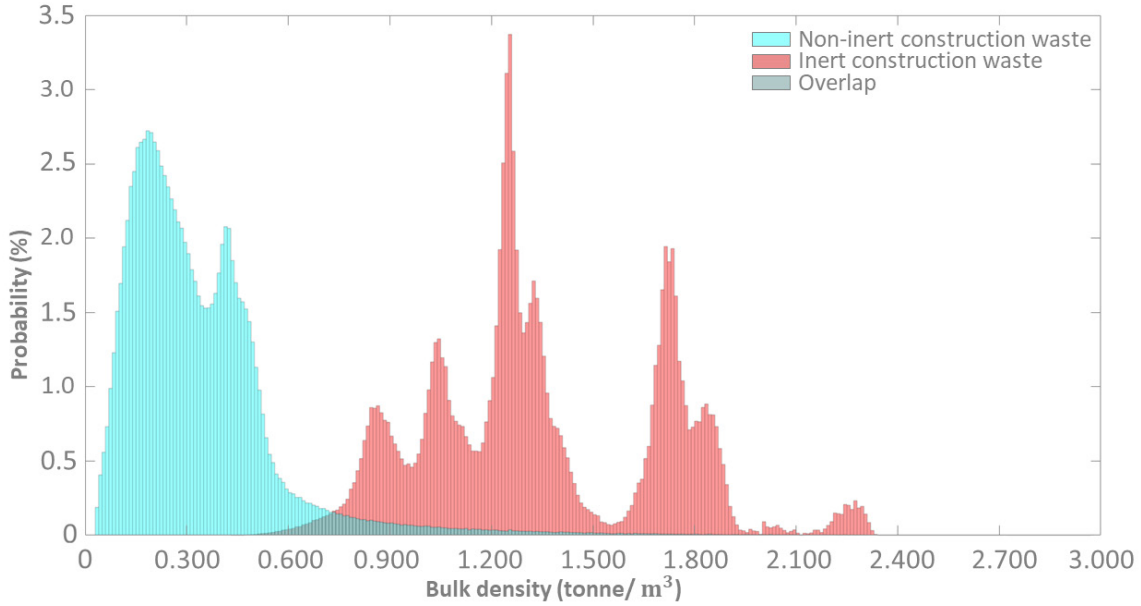

Figure 3. The bulk density probability distributions of inert and non-inert construction waste (Source: Adapted from Lu et al., 2021)

3. Research methodology

3.1 Big data-enabled statistical probability

The theoretical underpinning of the methodology here is a novel combination of traditional statistical probability theories with modern big data analytics. The classical statistical probability theories define an event's probability as the limit of its relative frequency in many repeated trials (Neyman, 1937). Conducting sufficient repeated trials to derive the frequencies and to ensure the trials are representative enough of the whole population is often too onerous to accomplish. Monte Carlo simulation and similar methods can be used to circumvent this problem, but such methods need to know the intervention of random variables. Modern analytics of big data, which is accumulated unintentionally as the business is done, offer a convenient means to this end. Big data can get researchers as close as possible to obtaining an event's probability, or even its totality. This has triggered a series of intellectual debate on the relationship between big data and scientific methods. Succi and Coveney (2019), for example, report that many researchers draw upon the Law of Large Numbers and foresee that with big enough data, errors (uncertainty) are bound to surrender to certainty. Big data-enabled correlation supersedes causation and science can advance even without coherent models, unified theories, or any mechanistic explanations (Anderson, 2008). This has even prompted the pronouncement that big data is the end of scientific theory.

Statistics and big data analytics are combined mainly in support of more accurate event prediction. Using big data-enabled probability distribution for prediction or estimation is a nascent field normally referred to as predictive analytics. Specifically, predictive analytics refers to using current and historical data to predict future or unknown events (Waller & Fawcett, 2013). Emerging studies have adopted it in several domains. Zuccolotto et al. (2018) analyzed the shooting performance of basketball players in high-pressure game situations by using the big data on Olympic Basketball Tournament. Zhang et al. (2015) estimated the probability distributions of safety distance of different-sized ships by using automatic identification system data collected in Singapore's port waters. Fu et al. (2018) proposed a big data-driven probability model to estimate the probability of insufficient gas supply incidents due to weather uncertainty. Li et al. (2017) developed a new method for tackling the problem of inaccurate disease probability prediction in the field of big health. These studies demonstrate the promise of using the big data-enabled probability distribution for predictive analysis. The key is to derive the probability distribution of the target event from the big data set, and then use a proper theoretical probability model (e.g., normal, lognormal, or Poisson distribution) to fit the distribution. Finally, the fitted probability model is used for predicting or estimating future events.

3.2 Harmonizing big data analytics and probability theories for waste composition estimation

The bulk density probability distributions of inert and non-inert construction waste are complex, showing two significant crests for non-inert waste and seven for inert waste (see Figure 3). Finding a theoretical probability model to fit these two distributions is not easy. In the end, inspired by Ross (2020), Scott (2018), and Torrecilla (2018), we decide not to fit theoretical probability models but to use the unfitted probability distribution directly. The Law of Joint Probability and Cumulative Distribution Function are adopted to build the core link of the composition estimation model. Firstly, we need to derive the joint probabilities of all possible composition forms (inert vs. non-inert composition) in a construction waste load. Then, we adopt the cumulative distribution function to calculate the cumulative probability of the composition belonging to a specific composition range or not.

The composition form of a construction waste load can be quantitatively expressed by the inert component weight proportion in the total weight, denoted as IWP (inert waste proportion). With a known IWP, the non-inert component weight can be easily derived because a waste bulk only consists of inert and non-inert parts. In this study, the output of construction waste composition estimation is defined as the range of IWP, such as $IWP > 80\%$.

The two bulk density probability distributions presented in Figure 3 can be expressed in Equation (2) and (3) as follows:

$$P(\rho_{inert} = \rho_j) = P_j;\ j=1, 2, 3, \ldots, n \quad (2)$$

$$P(\rho_{non-inert} = \rho_k) = P_k;\ k=1, 2, 3, \ldots, m \quad (3)$$

where ρ_j is one of the inert waste bulk densities (ρ_{inert}), and $\rho_{inert} \in [0.207, 2.435]$ tonne/m^3. ρ_k is one of the non-inert waste bulk densities ($\rho_{non-inert}$), and $\rho_{non-inert} \in [0.045, 2.434]$ tonne/m^3. n and m refer to the finite number of inert and non-inert composition bulk density values respectively, and they depend on the decimal place determination of weight and volume. Both the sum of all P_j and the sum of all P_k are 100%.

We know that a bulk of incoming construction waste contains inert and non-inert materials with an unknown weight ratio. The total waste weight and volume can be obtained by using measuring equipment as shown in Figure 1. The possible inert composition weight ranges from zero to the total waste weight, and possible inert composition volume ranges from zero to the total waste volume. Randomly selecting a value from the weight interval to be the weight of the inert component of the construction waste load, and a value from the volume interval to be the volume of the inert component, we can calculate an inert composition bulk density value (ρ_j) by using the selected weight divided by the selected volume. The selected inert composition weight and volume naturally determine non-inert composition weight and volume, because the

waste load consists of inert and non-inert components only. We can thus calculate a non-inert composition bulk density value (ρ_k) by the same method. In this way, we can assume a possible ratio of inert and non-inert composition in a waste load and calculate their respective bulk densities by having a pair of weight and volume values.

If the derived ρ_j belongs to ρ_{inert} interval and the derived ρ_k belongs to $\rho_{non-inert}$ interval, the combination of ρ_j and ρ_k exactly represents one possible composition of the new incoming waste load, denoted as IWP_i. We can obtain the probability of ρ_j according to Equation (2) and the probability of ρ_k according to Equation (3). The joint probability of ρ_j and ρ_k can be calculated using the law of joint probability (Kelley, 1994), as Equation (4):

$$P_i = P\left(\rho_{inert} = P_j \cap \rho_{non-inert} = P_k\right) = P_j \times P_k \tag{4}$$

where P_i represents the probability of IWP_i. In other words, P_i is the probability of one possible composition.

Traversing all possible inert composition weight values in the interval between zero and the total waste weight of the construction waste load, and all possible inert composition volume values in the interval between zero and the total volume, we can obtain a data set of IWP and a data set of joint probabilities. Each IWP value solely corresponds to one joint probability, the two data sets thus form a new probability distribution regarding construction waste composition IWP_i and corresponding joint probability P_i, as shown in Equation (5):

$$P(IWP = IWP_i) = P_i;\ i=1, 2, 3, \ldots, h \tag{5}$$

where h refers to the finite number of IWP values. Its finiteness results from manual operations or the precision limitation of measuring equipment. The sum of all P_i is 100%. As the probability of IWP belonging to a specific inert composition range is a cumulative probability, and the probability distribution of all IWP values has been derived as Equation (5), we can calculate the probability of IWP belonging to the specific inert composition range by using the cumulative distribution function (Park & Park, 2018), as in Equation (6):

$$F(CR) = P(IWP_i > R) = \sum_{IWP_i > R} P_i \tag{6}$$

where CR is a specific inert composition range. $F(CR)$ is the cumulative distribution function of R.

Figure 4 illustrates how the cumulative distribution function works. The inert composition range CR in the example is arbitrarily set as IWP_i >53%. When a truckload of construction waste is incoming (a) with the different possible composition of inert and non-inert materials, the individual probabilities of inert and non-inert composition (b) will be calculated by referring to the probability distribution patterns from the big data analytics. Then, the joint probability of each possible composition state will be calculated (c). The combination of the joint probabilities of all possible composition states is visualized in (d). Finally, the cumulative probability of the composition states belonging to the set CR is calculated, as shown in the blue area of Figure 4 (e). Another part (namely, the orange area) represents the cumulative probability of the composition states not belonging to the set CR. The orange and blue areas in Figure 4 (e) together present the probability distribution of all possible composition forms in the new incoming construction waste load, and their sum is equal to 100%. By comparing the two cumulative probabilities, if '$F(CR)$' is greater than '1- $F(CR)$', the composition of the newly incoming construction waste load belongs to the composition range of IWP_i; otherwise, not belongs. The estimation of waste composition is completed.

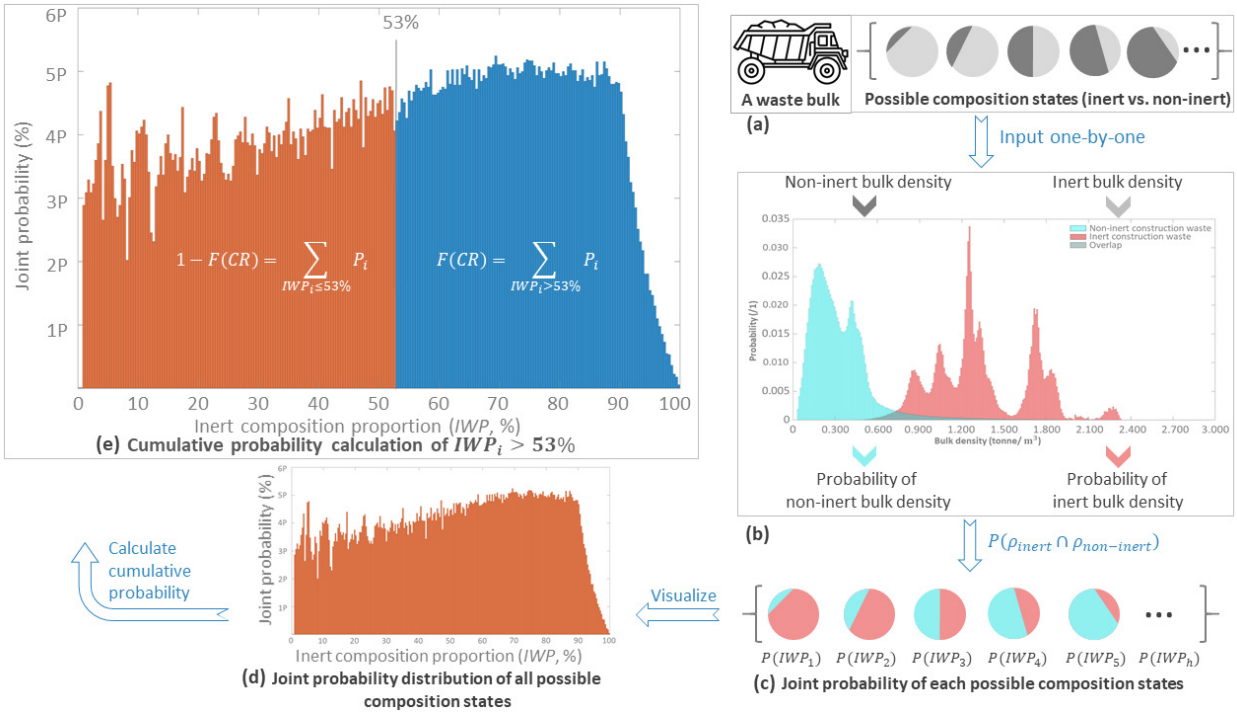

Figure 4. An example of cumulative probability calculations under a specific inert composition range

4. Development of the BD-P Model

Based on the research methodology as explained above, this section goes further to develop the construction waste composition estimation model. Given that the model is the bulk density probability distribution enabled by big data analytics, it is referred to as a big data-enabled probability model (BD-P model). As shown in Figure 5, the model consists of three parts: 1)

input collection to obtain the weight and volume of a construction waste load; 2) probability calculation to calculate the probability of each possible composition state; and 3) composition estimation to ascertain the composition of the construction waste load.

Using the BD-P model for composition estimation mainly includes two steps. The first step is measuring the weight and volume of construction waste. When there is an incoming truckload of construction waste, the volume and weight of construction waste can be easily obtained by using some on-site measuring equipment (see Figure 1) and combining pre-registered truck body specifications. According to a site observation for industrial practice, the collection of weight and volume generally takes one minute when using currently popular technologies and equipment. The second step is feeding the obtained weight and volume into the BD-P model for calculating construction waste composition.

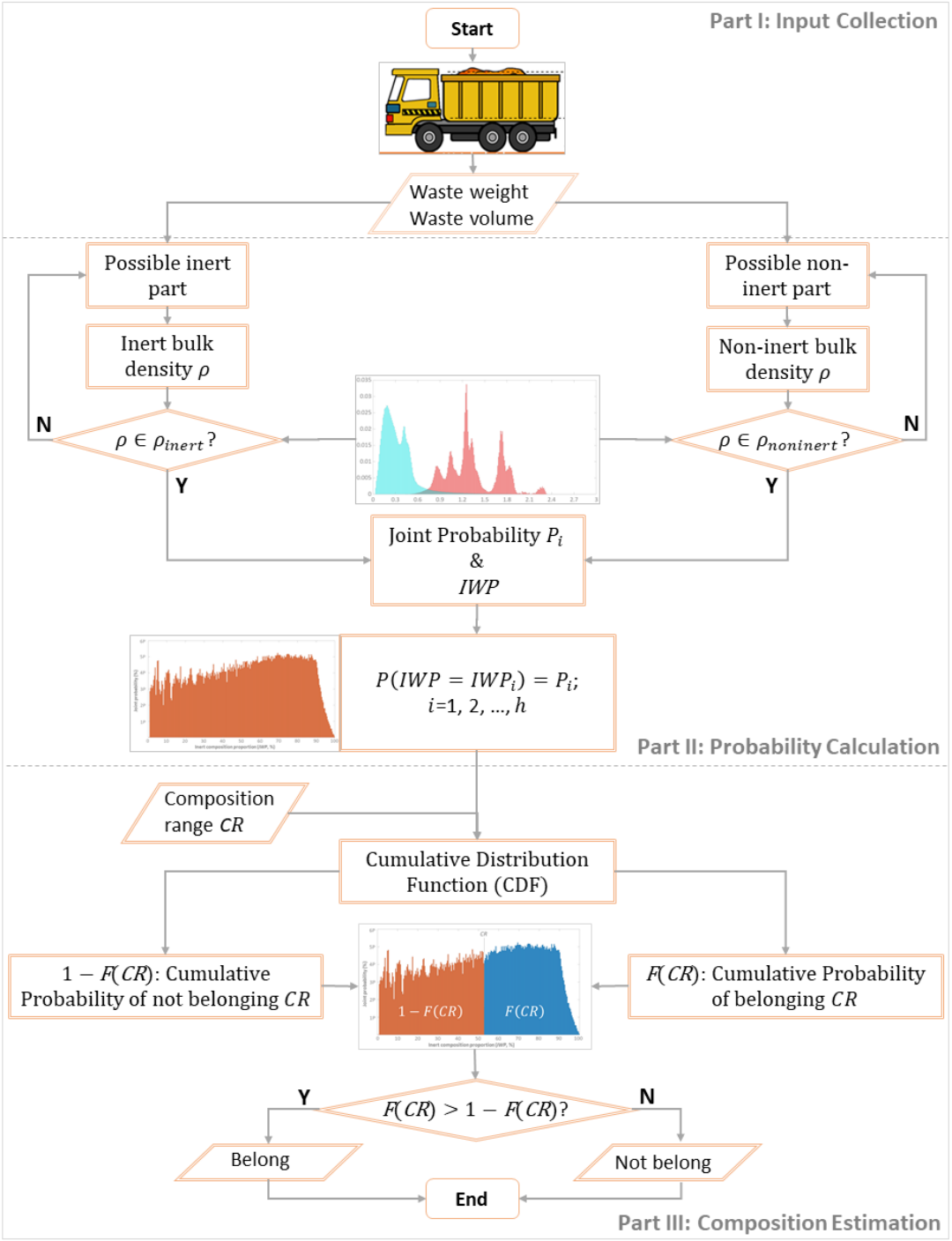

Figure 5. The architecture of the construction waste composition estimation model (BD-P model)

5. Model Validation

This section describes the experiments conducted to validate the BD-P model. It includes constructing ground truth data set, determining evaluation criteria for model performance, and analyzing experimental results.

5.1 Constructing the "ground truth"

Performance testing of the BD-P model requires multiple composition-known construction waste loads as ground truth data. In this study, raw ground truth data is collected from a construction waste sorting facility (TKO137SF) in Hong Kong. This facility is designated to receive and sort construction waste loads containing more than 50% inert composition by weight (HKEPD, 2008). The admission system installed at the entrance of the facility captures information about an incoming waste load to judge whether it contains more than 50% inert composition by weight, and therefore whether to accept or reject the waste load. The admission system will integrate all information regarding the waste load into a disposal record no matter the waste load being rejected or accepted. A complete record contains 37 variables, including total weight when the truck enters, weight when the truck exits, vehicle number, final decision about acceptance or rejection, waste bulk height, waste load photos, permitted gross vehicle weight (PGVW) of the hauling truck, facility name, date, and so on. We collected 22,800 such records, covering all construction waste load disposal data at TKO137SF from September to November 2019. The data set comprises 2,278 records for rejected waste loads (i.e., inert composition weight less than 50% of total waste weight), and 20,522 records for accepted waste loads (i.e., inert composition weight more than 50% of total waste weight). It is important to note that the compositions of 251 loads among the accepted construction waste loads were manually checked at the facility by experienced inspectors.

The collected data is used to construct a ground truth data set including positive and negative data. In this study, the positive data refers to waste loads containing more than 50% inert material by weight (i.e., $IWP > 50\%$), and the negative data refers to waste loads containing not more than 50% inert material by weight (i.e., $IWP \leq 50\%$). In general, a larger size of the ground truth dataset can promise a more reliable model accuracy (Krig, 2016), but collecting more ground truth data is usually limited by time and cost viability, e.g., the manual separation and inspection as at TKO137SF. A proper size is thus more feasible than a large size. Referring to the method proposed by Sharma et al. (2017), this study determines the proper ground truth dataset size when the model accuracy no longer changes significantly, i.e. stabilized or saturated. Our preliminary experimental analyses found that the model accuracy starts to be stable when the ground truth dataset size increases to around 600 data points with balanced positive and negative data. Therefore, this study decided to construct a ground truth dataset containing around 300 positive data and 300 negative data.

We can confidently adopt the 251 manually checked loads as positive ground truth data. For the negative data, we selected 450 construction loads from the chronological 2,278 rejected loads by the Equidistant Sampling method. However, we cannot directly use them as negative ground truth data due to the auditing report finding that the admission system is not 100% accurate. Therefore, using the admission system judgement as a reference, we combine the

waste weight, waste bulk height, waste load photos, and domain knowledge to manually check each rejected waste load's state, and correct it when finding the admission system's judgment wrong (see Figure 7 for an example). Through these, 339 raw positive data records and 362 raw negative data records are included in the ground truth data set. Finally, the data is further cleansed and a ground truth dataset containing 604 records is generated for model validation (see Figure 8).

Basic information:

- Waste weight: 3.20 tonnes;
- Max waste height: 0.52 m;
- Min waste height: 0.12 m;
- Average waste height: 0.40 m;
- PGVW: 24 tonnes;
- Visible waste compositions: soil (inert), concrete (inert), rock (inert);
- Type of work: foundation construction;
- Admission system decision: inert composition **less than 50%;**

Inference and correction:

- Possible rejection reason: the waste weight ratio (13%) is less than the required minimum value (18%), waste weight ratio=waste weight ÷PGVW;
- Manually corrected result: inert composition **more than 50%.**

Figure 7. An example of the manual check and correction (Note: for this truck type, the HKEPD admission system makes the >50% inert composition decision on the basis of waste height <1.6m and waste weight ratio >18% [HKEPD, 2019])

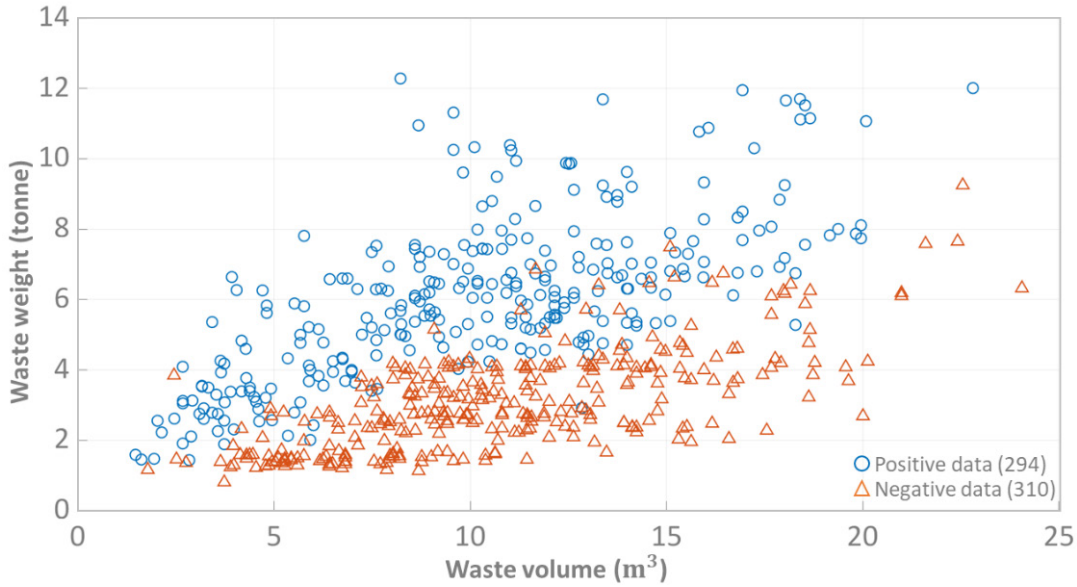

Figure 8. The weight and volume distribution of 604 ground truth data

5.2 Determining evaluation criteria for model performance

Accuracy, speed, complexity, and receiver operating characteristic (ROC) curve are widely used criteria for prediction model performance evaluation (Alpaydin, 2020; Yuan et al., 2020). For construction waste composition estimation, a model with higher estimation accuracy and speed is preferred for practical applications. Therefore, three indicators, namely accuracy, ROC curve, and speed are chosen as the BD-P model performance evaluation metrics in this study.

Estimation accuracy has two calculation methods. Mean absolute percentage error (De Myttenaere et al., 2016) is used when the estimated result is numerical, while Confusion Matrix method (Alpaydin, 2020) is suitable when the estimated result is not numerical. The estimation output in this study is not numerical, as shown in Figure 5 (Part III). Thus, the confusion matrix method as shown in Equation (7) is adopted to calculate accuracy:

$$Accuracy = \frac{AP + RN}{AP + RP + RN + AN} \times 100\% \quad (7)$$

where *AP* is the number of accepted positive ground truth data. *RN* is the number of rejected negative ground truth data. *RP* is the number of rejected positive ground truth data. *AN* is the number of accepted negative ground truth data. Acceptance means the model-estimated inert waste composition is in the > 50% by weight range, while rejection means it is not. *AP* and *RN* are correct estimation results. *RP* and *AN* are incorrect estimation results.

The ROC curve is a graphical plot that illustrates the diagnostic ability of a classification model at all classification thresholds (Fawcett, 2006). This curve presents two parameters: 1) the true positive rate, which is equivalent to the accepted positive rate (APR) in this study; and 2) the

false positive rate, which is equivalent to the accepted negative rate (ANR) in this study. APR reflects the model's sensitivity (or detectable rate) while ANR reflects the model's fall-out rate (Metz, 1978). Equation (8) illustrates their calculation methods. Additionally, the index Area Under Curve (AUC) derived from the ROC curve is usually used to quantitatively evaluate the model's diagnostic ability. The AUC value is equivalent to the probability that a randomly chosen positive sample is ranked higher than a randomly chosen negative sample (Fawcett, 2006).

$$APR = \frac{AP}{AP + RP}$$
$$ANR = \frac{AN}{AN + RN} \tag{8}$$

Model speed refers to the time consumption of conducting one complete computation for waste composition estimation. It mainly depends on the model's complexity, the computation power of used computers, and the weight and volume of construction waste loads. This study uses the consumed time for completing the computation of one construction waste load (namely, second per load) to represent model speed.

5.3 Conducting experiments and analyzing results

The BD-P model is coded into a computer program using MATLAB. The constructed 604 ground truth data are fed into the program one by one for experimental composition estimation. During the procedure for exploring the possible inert and non-inert composition ratio (Figure 5, Part I), the decimal places of the randomly selected inert composition weight and volume are set as three, usually the maximum for practical weight and volume measuring when using tonnes and cubic meters as units. The input composition range CR (Figure 5, Part III) is $IWP >$50% according to the classification of ground truth data. It means that the BD-P model will accept or reject a construction waste load if it estimates the inert composition weight range as more or less than 50%, respectively.

Figure 9 presents the experimental results of using the BD-P model to estimate the 604 waste loads. Figure 9 (a) shows that 252 of 294 positive data are accepted by the BD-P model and 293 of 310 negative data are rejected. These are correct estimations. However, the model incorrectly rejected 39 construction waste loads which actually meet the composition range CR, and incorrectly accepted 20 loads which actually do not meet the composition range CR. To sum up, of the 604 construction waste loads, the BD-P model correctly estimated 545 loads and incorrectly estimated 59. The model's estimation accuracy is 90.2%. Figure 9 (b) is the ROC curve of the BD-P model. An AUC of 0.8775 demonstrates the good diagnostic ability of the model. Figure 9 (c) indicates the median of the BD-P model's estimation speed is 52 seconds per construction waste load.

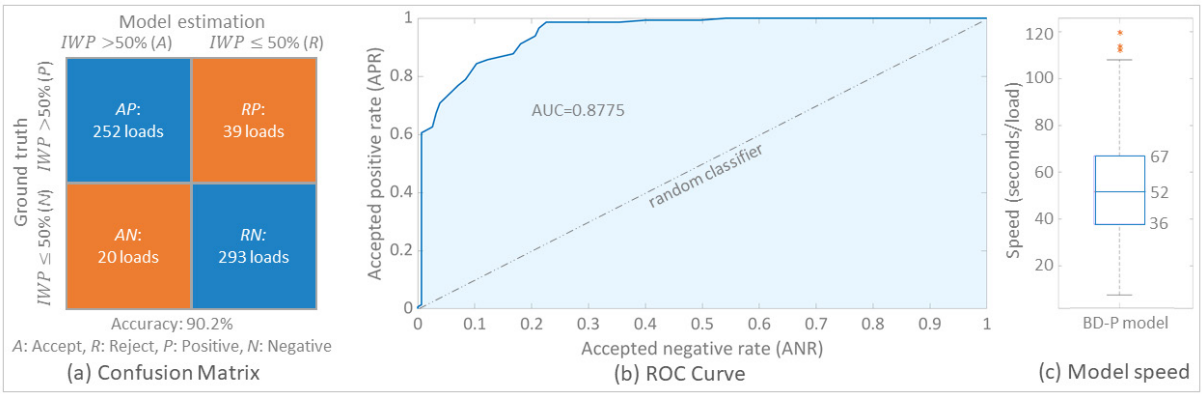

Figure 9. The experimental results of using the BD-P model to estimate the composition of the 604 waste loads

Furthermore, Figure 10 visualizes the estimated results of the 604 construction waste loads by using the BD-P model, where the input composition range CR is IWP >50%. Of particular interest is that there has an obvious division line in the estimated results. It implies a linear boundary. Generally, using a linear function would have a significantly smaller computation burden than using the BD-P model. Therefore, the research goes on to investigate the feasibility of using the linear function of this division line to substitute the BD-P model for composition estimation. Based on the output results by the BD-P model, we first use the Perceptron Learning algorithm (Stephen, 1990) to derive the optimal linear decision boundary function, as Equation (9):

$$\begin{cases} W - 0.4152V - 0.2944 > 0 & Yes; \\ W - 0.4152V - 0.2944 < 0, & No. \end{cases} \tag{9}$$

where W is the total weight of a construction waste load. V is its total volume. *Yes* means its estimated inert waste composition weight exceeds 50%. *No* means it does not. Then, we use the function to re-estimate the 604 constructed ground truth data. The experimental result indicates an estimation accuracy of 88.8%, slightly lower than the BD-P model's 90.2% but still acceptable. However, the computation speed of the linear function is about 130 times faster than that of the BD-P model, changing from an average of 52 seconds to 0.4 seconds per case. Therefore, it is recommendable to use the linear decision boundary function informed by the BD-P model to estimate construction waste composition.

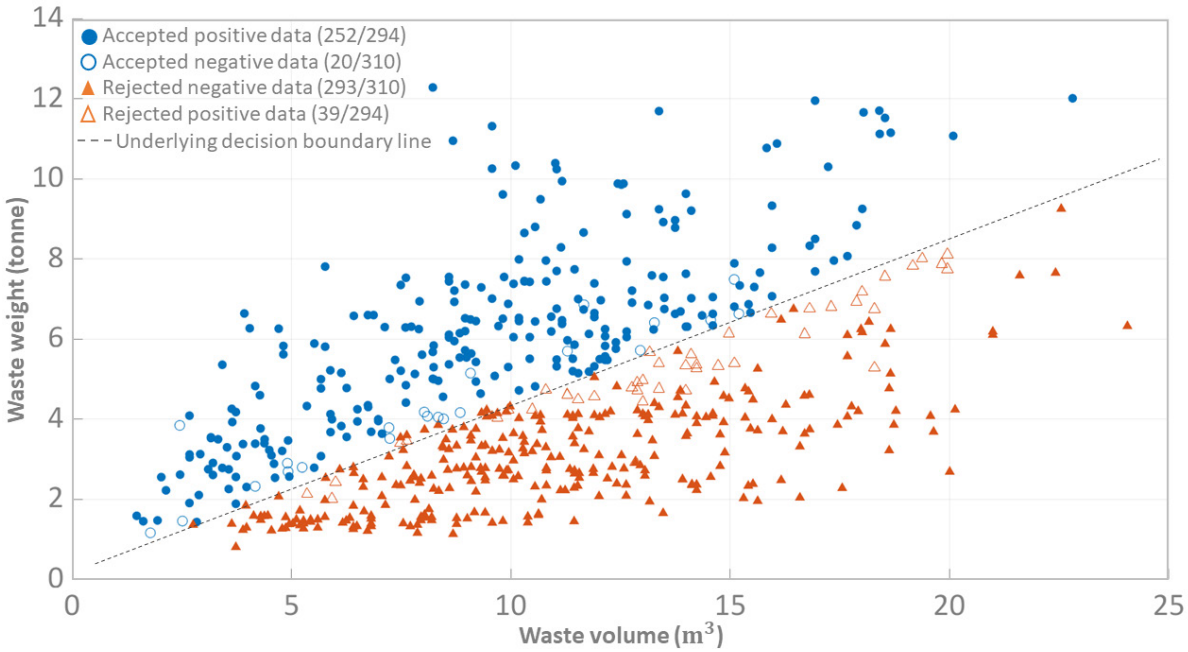

Figure 10. Composition estimation model results (accuracy: 90.2%)

6. Discussion

The strengths of the BD-P model are multifold. Firstly, it is easy to use. One can simply measure the weight and volume of a load of waste to calculate its bulk density, just like we measure one's body weight and height to calculate the BMI. Secondly, the BD-P model is easy to adjust for different criteria. The criterion for inert component waste proportion adopted in the experiments of this study is set at 50%, but it can be adjusted to other real-life criteria without necessarily changing the underlying BD-P modelling method. Thirdly, the accuracy of the BD-P model is at an acceptable level: 90.2% using the original BD-P model and 88.8% using the linear model generated from the BD-P model. The accuracies are slightly lower than the currently highest accuracy of construction waste composition estimation. Davis et al. (2021) achieved an accuracy of 94% by using a deep learning-enabled photogrammetry approach. Nevertheless, the BD-P model only requires two easy-to-obtain inputs: the weight and volume of construction waste loads. This saves much manual operation required by previous approaches, e.g., spreading waste to an even and thin depth for photogrammetry-based approaches, or pre-sorting construction waste into different composition categories for statistical sampling approaches. Meanwhile, the BD-P model can complete the computation at a speed of around 52s per construction waste load. Taking the time of measuring the waste weight and volume by using currently popular technologies and equipment (i.e., around one minute) into consideration, the BD-P model is expected to estimate the composition of a truckload of construction waste in two minutes.

The accuracy and speed achieved by the BD-P model make it highly applicable for real-life applications. In February 2021, we conducted a workshop with governmental officials, site inspectors, and system operators who are managing the CWM facilities in Hong Kong. The

BD-P model was generally accepted by the experts and stakeholders, who support us with more ground truth data to fine-tune the model towards 100% accuracy. This research has practical value to not only CWM facility managers but also construction contractors and waste hauliers. These latter stakeholders also face the difficulty of reasonably estimating the composition of their waste loads and deciding proper disposal destinations.

The biggest novelty of this research is the harmonization of traditional probability theories and modern big data analytics. There is a popular debate that big data is able to paint a fuller picture of a subject matter, therefore, undermine the necessity of traditional theories built upon probability. This research proves that big data is indeed able to inform a fuller spectrum of something but goes further to show that big data analytics and probability theories are allies instead of adversaries. The research also provides a compelling case of using probability distribution obtained from big data for predictive analysis. Traditional statistical approaches tend to find an existing theoretical probability distribution model (e.g., normal, lognormal, or Poisson distribution model) to fit the given data, and then use the fitted model for predictive analyses. However, in the real world, it is not often feasible to fit a theoretical distribution model. Instead, this study uses the raw probability distribution for predictive analysis without fitting the underlying theoretical probability model. The key is that the saturated big data set reveals the total spectrum of the probability distributions.

Nevertheless, the BD-P model also has several shortcomings. Firstly, it can only classify the construction waste into inert and non-inert portions. It is unable to identify specific construction materials (e.g., concrete, wood, vegetation, etc.) in a load of construction waste. Further research can select a more detailed composition classification or add extra information (e.g., images) to improve the estimation. Secondly, the BD-P model is not simply transferable to other economies with different construction techniques and waste composition characteristics. Necessary adjustments are needed in line with different data sets and admissibility criteria. However, as mentioned above, the underlying modelling is the same.

7. Conclusions

With an aim to solve a real-life conundrum as faced by many construction waste management facilities, this research developed a big data-probability (BD-P) model to quickly and accurately estimate the composition of construction waste for these facilities. The yardstick of the BD-P model is the convergent interval and stable probability distribution of construction waste bulk density, which are derived from big data analytics using 4.27 million truckloads of construction waste. Just two easy-to-obtain inputs, waste weight and waste volume, are required when applying the BD-P model to estimate construction waste composition. Experimental tests were conducted to validate the BD-P model. A ground truth data set consisting of 604 construction waste loads was constructed in-house and fed into the BD-P model to test its estimation

performance. Results revealed a model accuracy of 90.2%, Area Under Curve (AUC) of 0.8775, and speed of around 52 seconds per construction waste load.

Furthermore, to optimize the efficiency of composition estimation, this study proposes using the fitted linear function of the decision boundary generated by the BD-P model to substitute the BD-P model for composition estimation. An additional experiment was conducted to validate the proposal. Experimental results showed an accuracy of 88.8%, slightly lower than the BD-P model's 90.2% but still acceptable. However, the model speed is about 130 times faster than that of the BD-P model, changing from around 52s to 0.4s per construction waste load, proving the feasibility of the proposal. Therefore, both the BD-P model and the linear model have great promise for improving the operations of CWM facilities through rapid and accurate waste composition estimation.

The major novelty of this research is not only to showcase the power of big data to reach a more comprehensive truth of a subject, but also to illustrate that combining traditional probability theories and big data analytics can catalyse many powerful applications that cannot be done before. This study makes a methodological contribution in the context of data science. The developed BD-P model provides a method case of harnessing the unfitted probability distribution derived from big data analytics for predictive analysis. With big enough data to plot the probability distribution of a target event, it is not necessary to deduce the underlying theoretical probability model for an event's probability. Instead, the unfitted probability distribution can be used for predictive analysis directly. Moreover, future studies are recommended to strengthen the estimation model so that it can accurately estimate the quantities of specific construction waste materials (e.g., concrete, bricks, rocks, and timber).

Acknowledgement

This research is jointly supported by the Strategic Public Policy Research (SPPR) (Project No.: S2018.A8.010.18S) Funding Schemes and the Environmental Conservation Fund (ECF) (Project No.: ECS Project 111/2019) of the Hong Kong SAR Government.

References

Allegri, T. H. (1986). The Code of Federal Regulations—CFR Title 40. In *Handling and management of hazardous materials and waste* (pp. 98–170). Springer.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Anderson, C. (2008). *The end of theory: the data deluge makes the scientific method obsolete*. WIRED. https://www.wired.com/2008/06/pb-theory/

Asgari, A., Ghorbanian, T., Yousefi, N., Dadashzadeh, D., Khalili, F., Bagheri, A., Raei, M., & Mahvi, A. H. (2017). Quality and quantity of construction and demolition waste in Tehran. *Journal of Environmental Health Science and Engineering*, *15*(1), 14.

Aslam, M. S., Huang, B., & Cui, L. (2020). Review of construction and demolition waste management in China and USA. *Journal of Environmental Management*, *264*, 110445.

Australian Government. (2011). *Construction and demolition waste status report:*

Management of construction and demolition waste in Australia. https://bit.ly/39IJhLE
Califice, A., Michel, F., Dislaire, G., & Pirard, E. (2013). Influence of particle shape on size distribution measurements by 3D and 2D image analyses and laser diffraction. *Powder Technology*, *237*, 67–75.
Ceri, S. (2018). On the role of statistics in the era of big data: A computer science perspective. *Statistics & Probability Letters*, *136*, 68–72.
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).
Cochran, K., Townsend, T., Reinhart, D., & Heck, H. (2007). Estimation of regional building-related C&D debris generation and composition: Case study for Florida, US. *Waste Management*, *27*(7), 921–931.
Davis, P., Aziz, F., Newaz, M. T., Sher, W., & Simon, L. (2021). The classification of construction waste material using a deep convolutional neural network. *Automation in Construction*, *122*, 103481.
De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing*, *192*, 38–48.
Defra, U. (2020). *Landfill operators: Environmental permits*. Environment Agency. https://bit.ly/3sG0RZ6
EPA. (1997). *Construction Waste Management: A Guide for Municipalities*. https://bit.ly/2N9SWmE
EU. (2018). *EU construction and demolition waste protocol and guidelines: Internal market, industry, entrepreneurship and SMEs*. European Commission, Policies, Information and Services. http://m9q.net/mmfun
Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review*, *1*(2), 293–314.
Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.
Fu, X., Li, G., Zhang, X., & Qiao, Z. (2018). Failure probability estimation of the gas supply using a data-driven model in an integrated energy system. *Applied Energy*, *232*, 704–714.
HKAC. (2016). *Management of abandoned construction and demolition materials*. http://m9q.net/mmfgh
HKEPD. (2008). *Government waste disposal facilities for construction waste and charge level*. http://www.epd.gov.hk/epd/misc/cdm/scheme.htm#j
HKEPD. (2019). *Management of abandoned construction and management of abandoned construction and demolition materials*. https://www.aud.gov.hk/pdf_e/e67ch04sum.pdf
Hoang, N. H., Ishigaki, T., Kubota, R., Tong, T. K., Nguyen, T. T., Nguyen, H. G., Yamada, M., & Kawamoto, K. (2020). Waste generation, composition, and handling in building-related construction and demolition in Hanoi, Vietnam. *Waste Management*, *117*, 32–41.
Kelley, D. (1994). *Introduction to probability*. Macmillan Publishing Company, London.
Krig, S. (2016). Ground truth data, content, metrics, and analysis. In *Computer Vision Metrics* (pp. 247–271). Springer.
LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, *52*(2), 21–32.
Li, J., Chen, Q., & Liu, B. (2017). Classification and disease probability prediction via machine learning programming based on multi-GPU cluster MapReduce system. *The Journal of Supercomputing*, *73*(5), 1782–1809.
Lu, W., Chen, X., Ho, D. C. W., & Wang, H. (2016). Analysis of the construction waste

management performance in Hong Kong: The public and private sectors compared using big data. *Journal of Cleaner Production*, *112*, 521–531.
Lu, W., Chen, X., Peng, Y., & Shen, L. (2015). Benchmarking construction waste management performance using big data. *Resources, Conservation and Recycling*, *105*, 49–58.
Lu, W., Lee, W. M. W., Bao, Z., Chi, B., & Webster, C. (2020). Cross-jurisdictional construction waste material trading: Learning from the smart grid. *Journal of Cleaner Production*, *277*, 123352.
Lu, W., Yuan, L., & Xue, F. (2021). Investigating the bulk density of construction waste: A big data-driven approach. *Resources, Conservation and Recycling*, *169*, 105480.
Lyon, T. L., & Buckman, H. O. (1922). *The nature and properties of soils: A college text of edaphology*. Macmillan.
Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*(4), 283–298.
Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *236*(767), 333–380.
NSWEPA. (2014). *Waste Levy Guidelines*. http://m9q.net/mmfu5
NSWEPA. (2020). *Protection of the Environment Operations (Waste) Regulation 2014*. http://m9q.net/mmfuk
Paine, D. P., & Kiser, J. D. (2012). *Aerial photography and image interpretation*. John Wiley & Sons.
Park, K. Il, & Park. (2018). *Fundamentals of probability and stochastic processes with applications to communications*. Springer.
Quarteroni, A. (2018). The role of statistics in the era of big data: A computational scientist'perspective. *Statistics & Probability Letters*, *136*, 63–67.
Ross, S. M. (2020). *Introduction to probability and statistics for engineers and scientists*. Academic Press.
Scott, E. M. (2018). The role of Statistics in the era of big data: Crucial, critical and under-valued. *Statistics & Probability Letters*, *136*, 20–24.
Sharma, R. C., Hara, K., & Hirayama, H. (2017). A machine learning and cross-validation approach for the discrimination of vegetation physiognomic types using satellite based multispectral and multitemporal data. *Scientifica*, *2017*.
Stephen, I. (1990). Perceptron-based learning algorithms. *IEEE Transactions on Neural Networks*, *50*(2), 179.
Succi, S., & Coveney, P. V. (2019). Big data: the end of the scientific method? *Philosophical Transactions of the Royal Society A*, *377*(2142), 20180145.
Torrecilla, J. L., & Romo, J. (2018). Data learning from big data. *Statistics & Probability Letters*, *136*, 15–19.
Villoria Sáez, P., del Río Merino, M., & Porras-Amores, C. (2012). Estimation of construction and demolition waste volume generation in new residential buildings in Spain. *Waste Management & Research*, *30*(2), 137–146.
Vrancken, C., Longhurst, P. J., & Wagland, S. T. (2017). Critical review of real-time methods for solid waste characterisation: Informing material recovery and fuel production. *Waste Management*, *61*, 40–57.
Wagland, S. T., Veltre, F., & Longhurst, P. J. (2012). Development of an image-based analysis method to determine the physical composition of a mixed waste material. *Waste Management*, *32*(2), 245–248.
Waller, M. A., & Fawcett, S. E. (2013). *Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management*. *34*(2), 77–84.

Wu, Z., Ann, T. W., & Poon, C. S. (2019). An off-site snapshot methodology for estimating building construction waste composition-a case study of Hong Kong. *Environmental Impact Assessment Review*, *77*, 128–135.

Yuan, L., Guo, J., & Wang, Q. (2020). Automatic classification of common building materials from 3D terrestrial laser scan data. *Automation in Construction*, *110*, 103017.

Zhang, L., Wang, H., & Meng, Q. (2015). Big data-based estimation for ship safety distance distribution in port waters. *Transportation Research Record*, *2479*(1), 16–24.

Zuccolotto, P., Manisera, M., & Sandri, M. (2018). Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions. *International Journal of Sports Science & Coaching*, *13*(4), 569–589.