Building feature-based machine learning regression to quantify urban material stocks: A Hong Kong study

Liang YUAN[1], Weisheng LU[1]*, Fan XUE[1], and Maosu LI[2]

[1] Department of Real Estate and Construction, Faculty of Architecture, the University of Hong Kong.

* Corresponding email: wilsonlu@hku.hk

[2] Department of Urban Planning and Design, Faculty of Architecture, the University of Hong Kong.

This is the post-print version of the paper:

Yuan, L., Lu, W., Xue, F., & Li, M. (2023). Building feature-based machine learning regression to quantify urban material stocks: A Hong Kong study. *Journal of Industrial Ecology*, *to appear*. Doi: 10.1111/jiec.13348

The final version is available here: https://doi.org/10.1111/jiec.13348

Abstract

Urban material stock (UMS) represents an elegant thinking by perceiving cities as a repository of construction materials that can be reused in the future, rather than a burdensome generator of construction and demolition waste. Many studies have attempted to quantify UMS but they often fall short in accuracy, primarily owing to the lack of proper quantification methods or good data available at a micro level. This research aims to develop a simple but satisfactory model for UMS quantification by focusing on individual buildings. Generally, it is a 'bottom-up' approach that uses building features to proximate the material stocks of individual buildings. The research benefits from a set of valuable, 'post-mortem' ground truth data related to 71 buildings that have been demolished in Hong Kong. By comparing a series of machine learning-based models, a multiple linear regression model with six building features, namely building type, building year, height, perimeter, total floor area, and total floor number, is found to yield a satisfactory estimate of building material stocks with a mean absolute percentage error of 9.1%, root-mean-square error of 474.13, and R-square of 0.93. The major contribution of this research is to predict a building's material stock based on several easy-to-obtain building features. The methodology of machine learning regression is novel. The model provides a useful reference for quantifying UMS in other regions. Future explorations are recommended to calibrate the model when data in these regions is available.

Keywords

Urban material stocks; Sustainable development; Circular economy; Construction materials; Machine learning regression; Industrial ecology

1. Introduction

An urban material stock (UMS) is the accumulation of construction materials (e.g., bricks, concrete, timber, or steel) used in buildings and infrastructure within the urban area (Manelius et al., 2019; Tanikawa & Hashimoto, 2009). UMS plays a crucial but underappreciated role in shaping the use of material and energy resources (Krausmann et al., 2017). Research interest in UMS has grown in recent years. The UMS changes can be used as a proxy for understanding

urban metabolism (Schandl et al., 2020; Tanikawa & Hashimoto, 2009), and as an indicator for calculating embodied greenhouse gas emissions. UMS is considered a precursor for predicting future material demand and supply, and (Gassner et al., 2020; Huang et al., 2013). In this perception, urban material is a future anthropogenic resource deposit temporarily stocked in cities that can be reused or recycled in the future (Mesta et al., 2019; Nasir et al., 2021), rather than passively treating cities as material and energy consumers, or construction and demolition (C&D) waste generators (Marcellus-Zamora et al., 2016). The UMS is thus considered a proactive approach aligning with the global pursuit of sustainable development and a circular economy (Gontia et al., 2020; Haas et al., 2015).

Given the significant role of UMS, researchers have proposed various methods to measure it. For instance, Müller (2006) developed a demand-driven approach to quantify the concrete stock of the Netherlands' dwelling buildings. Tanikawa and Hashimoto (2009) proposed a spatial material stock analysis approach based on 4-dimensional geographic information system (4D-GIS) data. Fishman et al. (2014) proposed a novel method to quantify national material stock based on historical material flow data. Several recent studies have also tried to use a building component inventory-based approach to estimate material stock (de Tudela et al., 2020; Heeren & Hellweg, 2019; Heisel et al., 2022). Notably, studies have emerged to review and categorize UMS quantification methods. For example, Augiseau and Barles (2017) identify two main methodological approaches: bottom-up stock analysis and top-down retrospective stock analysis using a flow-driven model. Wiedenhofer et al. (2019) suggested a new classification: stock-driven vs. inflow-driven. More reviews and categorizations on UMS quantification methods can refer to Lanau et al. (2019) and Nasir et al. (2021).

Whilst existing methods have made important contributions to UMS quantification, they also suffer from shortcomings, mainly related to estimation accuracy. For both bottom-up and top-down approaches, the crux seems to lie in the accurate estimate of building material stock (BMS) or infrastructure material stock (IMS) at an individual level. By summing up the BMS and IMS in the urban area, a UMS can be derived. Nevertheless, the data to verify the BMS or IMS is often not available unless a building or an infrastructure is demolished, and its embodied materials are segregated and weighed. Such 'post-mortem' analyses are simply unrealistic. It is thus not surprising to see previous studies giving a possible error range for the estimate based on experience (Guo et al., 2019), or indirectly calibrating the accuracy level against studies in other regions (Mastrucci et al., 2017; Nasiri et al., 2021). The research on UMS quantification will see a breakthrough if a method is found to estimate the BMS or IMS with reasonable accuracy based on visible, easy-to-obtain building or infrastructure features.

This research aims to develop a novel UMS quantification approach based on easy-to-obtain building features (e.g., explicit building geometries, floor areas, and so on). As mentioned earlier, the repository of construction materials in the urban built environment comprises two

sectors: buildings and infrastructure. As the first step, this research focuses on buildings. The research benefits from a valuable data set related to waste generation and features of 71 buildings demolished in Hong Kong. The remainder of the paper is organized as follows. Subsequent to this introductory section, Section 2 reviews studies on UMS quantification. Section 3 introduces the research methodology and Section 4 describes model development and validation. Section 5 discusses the research contributions and shortcomings, and conclusions are drawn in Section 6.

2. Urban material stock quantification

A wealth of research has been conducted to quantify UMS. Following the widespread categorizations as adopted in the literature (see Table 1), we classified previous approaches into two types, namely top-down approaches and bottom-up approaches.

Table 1. Approaches for quantifying urban material stock

Type and basic principle	Variable	Variable gauging solution	Representative study
Top-down approaches: $MS = I - O$	Material inflow (I, unit: tonne)	Directly extracting from material flow statistic reports	Fishman et al. (2014)
		Multiplying construction activity amount (construction floor area or construction expense) by material input rate (material weight per floor area or expense)	Hashimoto et al. (2009))
	Material outflow (O, unit: tonne)	Multiplying material inflows by assumed stock survival probability distributions (e.g., normal distribution)	Bergsdal et al. (2007)
		Summing up C&D waste estimated based on waste generation rates	Wang et al. (2019)
Bottom-up approaches: $MS = \sum_{i=1}^{N} GFA \times MI$	Number of cohorts (N, unit: cohort)	Summing all building clusters divided by one or multiple criteria (e.g., usage type, building year, structure type).	Bergsdal et al. (2007); Ergun and Gorgolewski (2015); Mao et al. (2020); Miatto et al. (2019)
	Gross floor area (GFA, unit: m^2)	Multiplying population by floor area per capita (exclusive for residential buildings)	Müller (2006)
		Directly extracting from governmental building area statistic reports	Ortlepp et al. (2016)
		Extracting data from digital databases (e.g., GIS database, land-use database, Google Maps database, aerial photographs, earth-observation raster data) and then transforming them to the gross floor area	Haberl et al. (2021); Marcellus-Zamora et al. (2016); Tanikawa and Hashimoto (2009); Wang et al. (2019)
		Multiplying night-time light radiance by transformation coefficients	Peled and Fishman (2021)

	Material intensity (*MI*, unit: *tonne/m²*)	Directly extracted from public data (e.g., statistics reports, building specifications, design codes; or construction handbooks)	Hashimoto et al. (2007); Mesta et al. (2019); Tanikawa et al. (2015)
		Interviewing experts	Bergsdal et al. (2007)
		Calculating based on building documents (e.g., drawings and bill of quantities) and sample measuring	Ergun and Gorgolewski (2015)

2.1 Top-down approaches

Top-down approaches estimate construction material stock by calculating the difference between material inflows and outflows; that is, construction materials input and output from in-use buildings, respectively. There are two ways to estimate material inflows: 1) directly extracting them from material inflow statistics, and then, 2) multiplying the quantity of construction activities by a material input rate. Material outflows can be obtained by: 1) multiplying material inflows by an assumed stock survival probability distribution; or 2) adding up the construction waste quantity from new C&D activities. These approaches are mainly used to derive the material stock of countries or regions, and uncover the evolution of construction material stocks over a period by combining temporal material flows.

2.2 Bottom-up approaches

Bottom-up approaches mean to sum up the total material bank building by building. It involves three variables: 1) number of building cohorts, 2) gross floor area (GFA) of each building cohort, and 3) typical material intensity of each building cohort. Since no two buildings in a building cohort have the same material stock, ideally this would be quantified building by building. However, the cost to do so would be prohibitively high. In a trade-off between accuracy and cost, previous studies chose to classify buildings into multiple cohorts according to certain criteria such as building age, usage type, and structure type (Kleemann et al., 2017; Mesta et al., 2019).

Researchers have proposed various methods to calculate GFA. Müller (2006) proposed to multiply the population by floor area per capita to estimate the GFA of residential buildings. Increasingly, governments start to issue building area statistical reports, from which researchers can directly extract GFA data for their uses (Han & Xiang, 2013; Ortlepp et al., 2016). Urban digital databases (e.g., GIS, building footprint databases) also enable researchers to aggregate GFA via simple data extraction and transformation (Haberl et al., 2021; Marcellus-Zamora et al., 2016). Peled and Fishman (2021) innovatively used night-time light radiance data to calculate GFA in Europe.

Material intensity, or 'building material composition', describes the construction material weight per GFA. It represents the typical material composition of each building cohort. One method used to obtain material intensities is to estimate based on public data, including

building specifications (Tanikawa & Hashimoto, 2009), design codes (Han & Xiang, 2013), and construction handbooks (Gao et al., 2020). When these public data sets are unavailable or incomplete, researchers tended to interview local experts (Bergsdal et al., 2007; Mesta et al., 2019). The third method is to calculate material intensities based on building documents gathered, such as drawings (Kleemann et al., 2017; Nasiri et al., 2021) and bill of quantities (Mao et al., 2020), with site visits to collect supplementary data (Surahman et al., 2017). Since cost and time limits, researchers usually sampled one or several representative buildings from different cohorts, and then applied the abovementioned methods to derive representative material intensities.

2.3 Strengths and weaknesses of existing approaches

Top-down approaches are convenient and efficient. However, they involve simplified building lifespan assumptions (Fishman et al., 2014), and estimated coefficients, e.g., material input rate (Hashimoto et al., 2009), or waste generation rate (Wang et al., 2019), which may lead to large deviations in the final estimate. Also, they rely heavily on the availability of statistical data about material inflows or construction activities. Therefore, they are mainly applicable for national or regional material stock quantification.

By comparison, bottom-up approaches can quantify material stocks at a micro spatial scale. However, it also has two weaknesses. The typical material intensity of each building cohort is usually derived based on one or several sampled building representatives. Sampling representative buildings inevitably leads to bias in the quantification result due to building heterogeneity (Mollaei et al., 2021; Stephan & Athanassiadis, 2017). Furthermore, no research explains why a building is representative (Brøgger & Wittchen, 2018). Inappropriate selection of representative buildings will also cause inaccurate UMS quantification. Therefore, in short, the crux is the quantification of BMS. Another weakness of the bottom-up approaches is that little research has derived material intensities by using real-life material stock data. This may be due to the difficulty of weighing all construction materials in a building.

3. Methodology

3.1 The rationale

The rationale of our BMS quantification approach is intuitive: there should be a correlation between building material stock and building features. Generally, for the geometrical features (i.e., building height, floor area, perimeter, and storeys), any increase will naturally result in a rise in BMS. Some semantic features also impact BMS. For instance, Mao et al. (2020) and Kleemann et al. (2017) demonstrated that building age is related to BMS, as different building materials and codes may have been adopted in different eras. Mastrucci et al. (2017) and Lanau and Liu (2020) showed that BMS also depends on building usage (e.g., residential or commercial). Usually, these visible building features are easy to obtain, e.g., by referring to drawings, building approvals, or simple surveys. If a reliable correlation between a building's

features and its embodied material stock can be ascertained, the BMS can be accurately estimated rather than waiting for the building to be demolished and its materials to be segregated and weighed.

3.2 Research methods

Figure 1 illustrates our proposed BMS quantification approach. At the core of the approach is machine learning (ML) regression. According to Ij (2018), traditional statistical regression mainly infers and interprets the relationships between the dependent variable (e.g., the BMS) and one or more independent variables (e.g., building features). In the end, a human interpretable model like $y = a_i x_i + b$ or non-linear relationships will be derived. Unlike traditional statistical regression, ML regression aims to construct a generalizable model that can accurately predict the unobserved output (Ij, 2018). It is largely a data-driven approach by choosing a regression algorithm and harnessing the power of data. In other words, it may lead to a satisfactory predicting model that can be applied for specific objectives such as BMS estimation, although the model itself may not be human-interpretable. Given the primary aim of the research, ML regression is chosen.

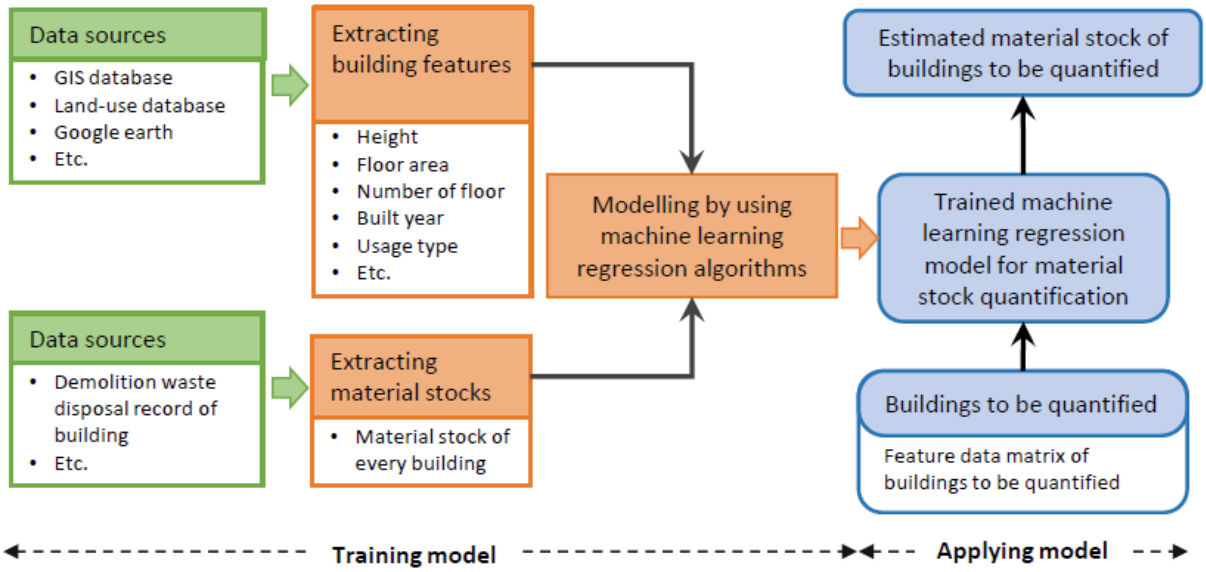


Figure 1. Framework of the proposed UMS quantification approach

3.3 Data collection

Two data sets from the same batch of buildings are needed: construction material stocks (by weight) and building features. For the former, in 2005 the Hong Kong Environmental Protection Department (HKEPD) established a Construction Waste Disposal Charge Scheme to manage C&D waste. According to the Scheme, any building demolition project with a contract sum of more than HK$ 1 million must apply for a sole disposal account from the HKEPD. This account records the features of the demolition project (e.g., location, client, and

contract sum) and the details (e.g., weight-in, weight-out) of every truckload of demolition waste material disposed of at government facilities.

Figure 2 illustrates the C&D waste disposal and recording system in Hong Kong. Because the Scheme adopts a closed-loop management approach, any piece of demolished waste, unless properly reused or recycled, must be disposed of at government facilities and use the account opened with the HKEPD (Yuan et al., 2013). We thus consider that the total waste materials derived from an account accurately represent the material stock of that building, with a caveat that a few materials might have been reused or recycled elsewhere and hence not been properly recorded in the system. We obtained the accounts of 433 demolition projects implemented between 2011 and 2020 with a contract sum larger than HK$1 million. We then derived the weight of waste materials for individual buildings.

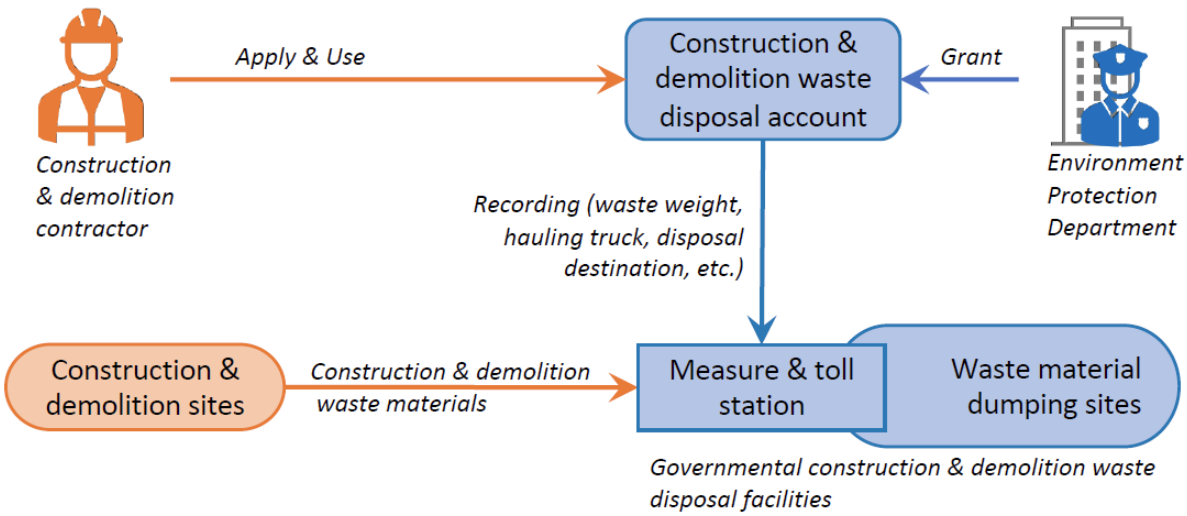
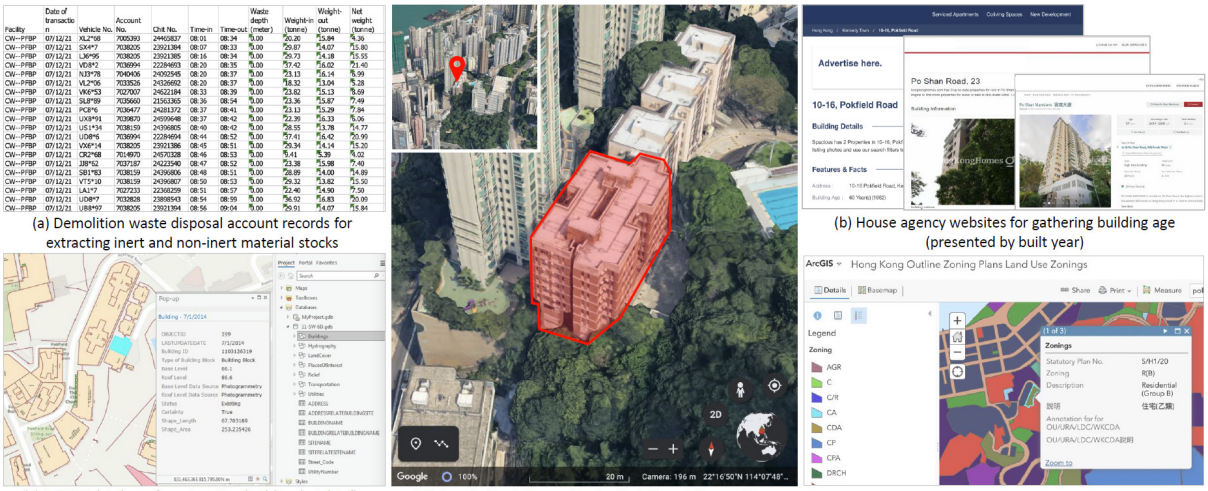


Figure 2. The construction and demolition waste disposal and recording system in Hong Kong

Then, we acquired the building feature data. Based on the building location recorded in the accounts, we first used Google Earth 3D views to count total floor number (above ground) of the building. Building height, floor area, and perimeter were then derived by using iB1000 (HKLD, 2022a), which is a 1:1,000 digital topographic map issued by the Lands Department providing address, hydrography, land coverage, place of interest, relief, transportation, and utility information of all buildings in Hong Kong. For those buildings that had been wiped from the iB1000 database, their heights, floor areas, and perimeters were obtained from Google Earth and Google Maps. To derive building usage types (i.e., residential or non-residential), we referred to the Planning Department's land utilization type database (HKPD, 2021). Building ages were gathered via housing agencies' websites.

Figure 3 shows the data collection process. We finally collected a data set of 71 buildings. Figure 4(a) visualizes the structure of collected data. The data set contains the building stocks

and six easy-to-obtain building features: (1) building year, (2) total floor number, (3) building height, (4) total floor area, (5) building perimeter, and (6) building usage type. Given Hong Kong's hilly terrain and steep slopes, it is a common practice to build a podium and then construct multiple high-rise towers on it. Therefore, this study further divides a building into two parts: building blocks and podium. When a building has no podium, the value of the podium variable is set to zero.



(a) Demolition waste disposal account records for extracting inert and non-inert material stocks

(b) House agency websites for gathering building age (presented by built year)

(c) iB1000 database for extracting building height, floor area, and perimeter

(d) Google Earth for confirming building while acquiring building height and floor number

(e) Land utilization database for obtaining building usage type

Figure 3. An illustration of the data collection processes

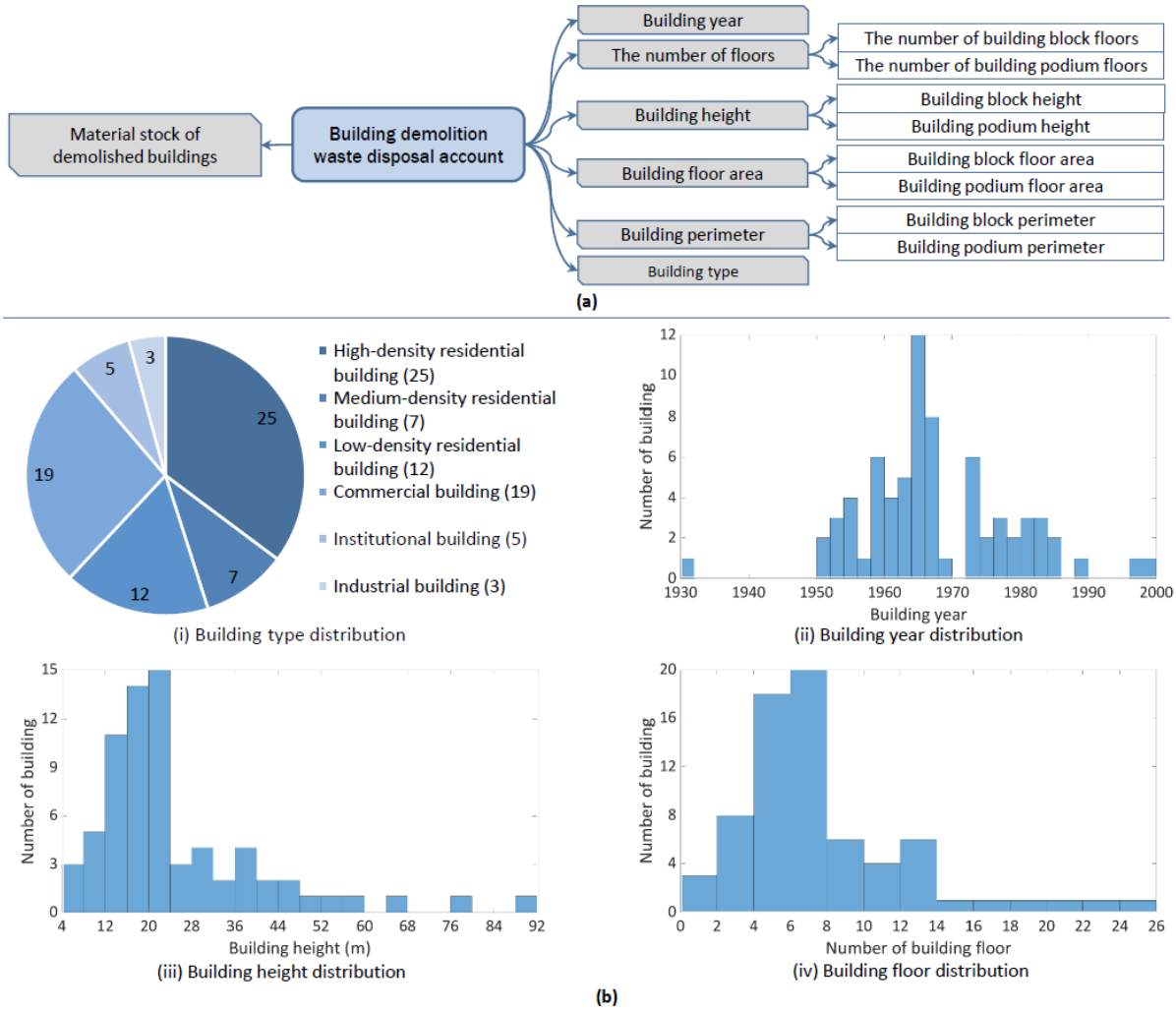


Figure 4. (a) The structure of collected data; (b) The data distribution under four common building categorization features (underlying data for Figure 4b are available in Table S2 of Supporting Information S1)

235

At a glimpse, the 71 buildings consist of 44 residential and 27 non-residential buildings. Their building years range from 1930 to 1999 with most between 1950 and 1980. 65 out of the 71 buildings have no podium. Building heights range from 4.0m to 88.8m and the total floor numbers range from 1 to 26. The building material stocks range from 299 to 9,967 tonnes with
240 most concentrating between 840 and 2,800 tonnes. It is noticed that the material stocks have no detailed composition information, e.g., inert or non-inert; concrete, copper, steel, etc.). Figure (4)b presents the data distribution in line with four building features.

4. Model development

245 ### *4.1 Selecting building features*

Building features are independent variables. Previous studies indicate that BMS is correlated with building geometrical features such as GFA (Nasiri et al., 2021), building height (Kleemann et al., 2017), and perimeter-area ratio (i.e., perimeter per unit floor area) (Stephan

& Athanassiadis, 2017); as well as semantic features such as building type (e.g., residential, non-residential) (Gontia et al., 2019). We also referred to the building features used in other research domains, such as building energy consumption prediction (Fan et al., 2014; Seyedzadeh et al., 2018) and C&D waste estimation (Lu et al., 2021; Maués et al., 2020), particularly on demolition waste generation (Chen & Lu, 2017; Lu et al., 2016). By combining the previous literature with the data presented in Figure 4, we sourced 33 building features that might be relevant to BMS quantification (see Table 2). In addition, we were aware of the presence of other potential features, such as roof type, façade material, and so on. However, this research does not consider them due to the data availability limitation.

Table 2. The 33 building features relevant to BMS quantification

Type	Name	Symbol	Data collection or calculation methods
Original variables	Building block height	H_{block}	Collecting from the first sources
	Building block floor area	FA_{block}	
	The number of building block floors	FN_{block}	
	Building block perimeter	P_{block}	
	Building podium height	H_{podium}	
	Building podium floor area	FA_{podium}	
	The number of building podium floors	FN_{podium}	
	Building podium perimeter	P_{podium}	
	Building year	Y_{built}	
	Building type (residential vs. non-residential)	UT	
Composite variables	Building block floor height	FH_{block}	$H_{block} \div FN_{block}$
	Building block façade area	FCA_{block}	$P_{block} \times H_{block}$
	Gross volume of building block	GV_{block}	$FA_{block} \times H_{block}$
	Gross floor area of building block	GFA_{block}	$FA_{block} \times FN_{block}$
	Height-floor area ratio of building block	H/FA_{block}	$H_{block} \div FA_{block}$
	Envelope-gross volume ratio of building block	E/GV_{block}	$(FCA_{block}+FA_{block}) \div GV_{block}$
	Perimeter-floor area ratio of building block	P/FA_{block}	$P_{block} \div FA_{block}$
	Building podium floor height	FH_{podium}	$H_{podium} \div FN_{podium}$
	Building podium façade area	FCA_{podium}	$P_{podium} \times H_{podium}$
	Gross volume of building podium	GV_{podium}	$FA_{podium} \times H_{podium}$
	Gross floor area of building podium	GFA_{podium}	$FA_{podium} \times FN_{podium}$
	Height-floor area ratio of building podium	H/FA_{podium}	$H_{podium} \div FA_{podium}$
	Envelope-gross volume ratio of building podium	E/GV_{podium}	$(FCA_{podium}+FA_{podium}) \div GV_{podium}$
	Perimeter-floor area ratio of building podium	P/FA_{podium}	$P_{podium} \div FA_{podium}$
	Total building height	H_{total}	$H_{block}+H_{podium}$
	Total number of floors	FN_{total}	$FN_{block}+FN_{podium}$

Total façade area	FCA_{total}	FCA_{block}+FCA_{podium}
Gross floor area	GFA_{total}	GFA_{block}+GV_{podium}
Gross building volume	GV_{total}	GV_{block}+ GV_{podium}
Mean floor height	FH_{mean}	(FH_{block}+FH_{podium})÷2
Mean height-floor area ratio	H/FA_{mean}	(H/FA_{block}+ H/FA_{podium})÷2
Mean envelope-gross volume ratio	E/GV_{mean}	(E/GV_{block}+ E/GV_{podium})÷2
Mean perimeter-floor area ratio	P/FA_{mean}	(P/FA_{block}+ P/FA_{podium})÷2

Notes: (1) For the definition of features such as block, and podium; as well as the calculation method of building height, perimeter, and floor area, please refer to HKLD (2022b); (2) We have conducted a preliminary regression prediction experiment, which demonstrated that re-categorizing the original six building types into two categories is optimal.

Of the features, 10 are original variables derived from original data collection while the other 23 features are composite variables, which means they are made up of two or more variables or measures that are highly related to one another. Using composite variables is a common practice for organizing multiple highly correlated variables into more concise yet meaningful ones (Song et al., 2013). Among the 33 feature variables, some of them are cross-correlated (e.g., H_{block} and FN_{block}). In statistical regression, the cross-correlation may result in the multicollinearity problem, affecting the coefficient estimation of independent variables. For regression prediction performance, however, one does not need to worry about the specific role of independent variables and their multicollinearity. According to Kutner et al. (2004), multicollinearity would not change regression prediction performance.

It is not necessarily a case of the more variables, the better a model performs (Guyon & Elisseeff, 2006). It is thus critical to choose appropriate variables for an ML regression. There are three methods for variable selection. The *Filter* method uses variable ranking techniques for variable selection (Chandrashekar and Sahin (2014). A suitable ranking criterion (e.g., Pearson's *r*) is used to score variables and a threshold is used to remove variables below the threshold. Its drawbacks are that the selected variables might not be optimal, and a redundant subset might be obtained. The *Wrapper* method decides the best variable subset by comparing the real model performance when inputting different variable combinations. Its shortcoming is the huge computation cost to exhaust all feature combinations. Finally, the *Embedded* method also compares the real model performance, but it integrates feature ranking techniques to optimize the computation cost. In this study, we adopt the *Embedded* method to investigate the best combination of variables (i.e., building features). The general procedure of the *Embedded* method can refer to the work by Lal et al. (2006) and Chandrashekar and Sahin (2014).

4.2 Choosing machine learning regression algorithms

To determine an appropriate ML regression algorithm, the types of output variables must be considered. For a categorical variable with different logistic values, logistic regression

algorithms are recommended (Nasteski, 2017). When the variables are continuous, there are six types of ML regression algorithms: multiple linear regression, support vector machine regression, tree regression, Gaussian process regression, tree regression ensemble, and regression neural network algorithms (Fahrmeir et al., 2021). In this study, the dependent variable (i.e., material stock of buildings) is continuous. However, it is unknown which one of the six algorithms is better. Therefore, we compared these regression algorithms to ascertain the one with the best predicting performance.

The comparison requires one or more model evaluation criteria. Five criteria, namely mean squared error (MSE), root-mean-squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and R-square, are usually used to evaluate the predicting performance of regression models (Emmert-Streib & Dehmer, 2019). In this study, the ML regression model is expected to estimate the BMS with high accuracy. Thus, MAPE, which indicates the regression model accuracy (accuracy=100% − MAPE), is the top choice. RMSE and R-square are also selected for evaluating the accuracy of the estimate.

4.3 Data preparation

Prior to ML regression model training and validation, it is necessary to conduct data normalization and division. Data normalization, which means scaling data to a defined interval (Luor, 2015), is needed due to the big magnitude order difference of different feature data. For example, the GFA ranges from 219.77 to 41,077.04 (unit: m^2) whereas the total floor number only ranges from 1 to 26 (unit: floor). The range difference has negative impacts on the performance of regression models that leverage the relative weight of features (Gal & Rubinfeld, 2019). Among the multiple applicable data normalization methods, this study used z-score scaling to normalize feature data. z-score scaling transforms a data sample into a new set with a mean value of 0 and a standard deviation of 1, by using Equation (1):

$$z\text{-}score\ value = \frac{X_i - \mu}{\sigma} \quad (1)$$

where X_i is an observation value of a feature. μ and σ are the mean and standard deviation of all observation values of the feature, respectively.

Data division aims to split all sample data into two datasets, one for model training and the other for model validation. When the collected data sample is big enough, the popular method is to randomly split 70~80% of the sample for model training, while using the rest 20~30% for model validation. When sample data is not abundant, like in this case, k-fold cross-validation is recommended. Cross-validation is a resampling technique that uses different portions of the data to train and test a model on different iterations (Browne, 2000). However, for a small training data sample, the probability of model overfitting would be high (Karystinos & Pados,

2000). Overfitting means a model performs well on the small data sample in model training but badly in predicting based on new, unseen input data (Mohri et al., 2018). Under this circumstance, a recommended solution is to synthesize the new data based on the original data to expand the training dataset for model training (Karystinos & Pados, 2000; Ying, 2019). Data synthesis is akin to transferring the training data insufficiency problem to the classical data imbalance problem, and then solving it using oversampling techniques (Sharma et al., 2021).

With 33 determined building features, the data size of the 71 buildings is actually rather small. This study thus used the data synthesis technique to prepare datasets for model training and validation. Following the comparative analysis conducted by Aborujilah et al. (2020), this study chose the *Bootstrap* algorithm to synthesize new data. The specific data preparation process is as follows:

1) Shuffling the 71 samples and randomly extracting 50 of them as model training data and the remaining 21 for model testing;
2) Based on the training sample of 50 buildings, using the *Bootstrap* algorithm to synthesize new data until the sample is expanded to 1000;
3) Training an ML regression model based on the 1000 samples;
4) Testing the trained model by using the remained 21 samples and recording model performance; and
5) Repeating steps 1–4 ten times and then calculating the average model performance. This is equivalent to 10-fold cross-validation.

4.4 Model training and validation

Based on the *Embedded* feature selection method, we first conducted experiments to find out the best feature combination for every determined ML regression algorithm. The experiments were implemented in the MATLAB platform. Specifically, we used the *F-test* to calculate the importance score of each feature and ranked them based on the score (MathWorks, 2022) (See Figure 5(a)). Round by round, the feature with the lowest importance score was removed while the rest features were used as the model input. With 33 building features and six types of regression algorithms, 198 rounds of experiments were conducted to obtain the best combination of building features for each type of regression algorithm.

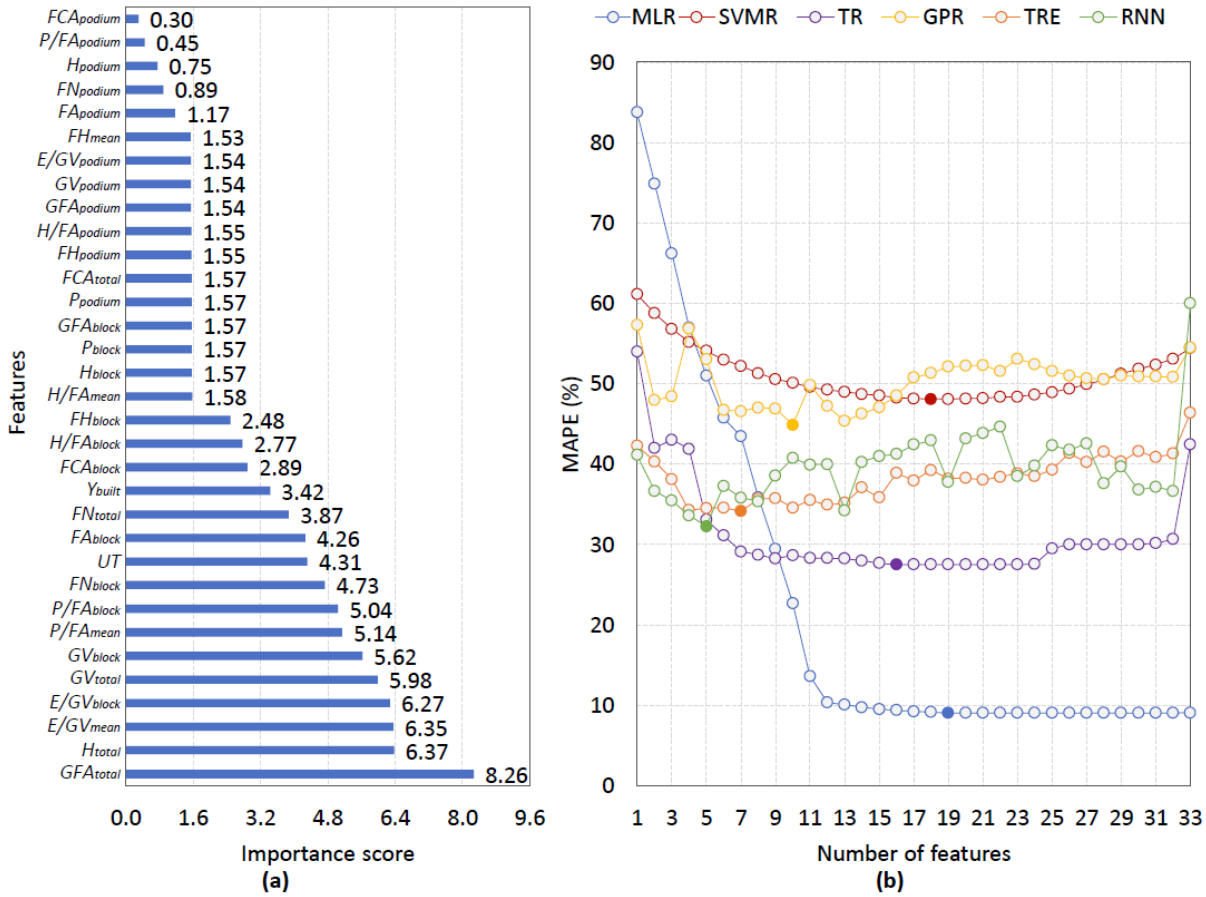


Figure 5. (a) Feature importance score ranking by using the F-test algorithm; (b) The change of MAPE when inputting different feature subsets and using different ML regression algorithms (MLR: multiple linear regression, SVMR: support vector machine regression, TR: tree regression, GPR: Gaussian process regression, TRE: tree regression ensemble, RNN: regression neural network; underlying data for Figure 5a&b are respectively available in Table S2&S3 of Supporting Information S2)

The feature selection experiment results are shown in Figure 5(b), which presents how the MAPE changes in line with the number of selected features. The feature number that corresponds to the lowest MAPE represents the optimal result. As shown in filled dots in Figure 5(b), the optimal building feature numbers are 19 for multiple linear regression algorithm, 18 for support vector machine regression algorithm, 16 for tree regression algorithm, 10 for Gaussian process regression algorithm, 7 for tree regression ensemble algorithm, and 5 for regression neural network algorithm. In addition, under a certain optimal feature number, the specific features, which constitute the best feature combination, were identified and then extracted from the full feature set (see Table 2) for further experiments.

Based on the best feature combination of each regression algorithm, we conducted experiments to explore the regression model that has the best performance in estimating material stocks. Six different models were trained and tested. Table 3 summarizes the experimental results. When

measured by MAPE, the multiple linear regression model is the best because of showing the smallest MAPE (9.1%). This value means the model estimation accuracy reaches 90.9%. Meanwhile, the regression tree model is also acceptable, with a MAPE of 27.5% which is equivalent to an accuracy of 72.5%. If only considering RMSE, the regression tree ensemble model performs better than the multiple linear models. However, both of them are excellent. When focusing on the R-square metric, both the multiple linear regression model and the regression tree ensemble model perform well, respectively being 0.93 and 0.78. In this study, the best model is determined by considering the MAPE, RMSE, and R-square together. Therefore, the trained multiple linear regression model with 19 building features is recommended for estimating BMS.

Table 3. The performance of models using different ML regression algorithms

Regression model algorithm	Model's MAPE	Model's RMSE	Model's R-square
Multiple linear regression	**9.1%**	**474.13**	**0.93**
Support vector machine regression	48.1%	1325.70	0.21
Regression tree	27.5%	1,050.50	0.46
Gaussian process regression	44.8%	1,904.41	0.28
Regression tree ensemble	34.2%	448.25	0.78
Regression neural network	32.3%	1,302.44	0.33

Referring to existing regression prediction studies, a consensus is that both MAPE<10% and R-square>0.9 represent a good model performance. Thus, the multiple linear regression model developed by using the data of the 71 buildings is sufficient for BMS estimation. Besides, unlike traditional statistical regression models, the trained multiple linear regression model is an ML model that consists of 121 function terms. It is therefore difficult to express it in mathematical forms or interpret it by a human. We can treat the model as a 'black box' considering the primary concern is its estimation performance. Equation (2) outlines the ML regression model:

$$\widehat{MS} = \mathcal{F}(H_{total}, GFA_{total}, GV_{total}, P/FA_{block}, FN_{total}, Y_{built}, FA_{block}, P_{block}, FN_{block}, UT, P/FA_{mean}, FCA_{block}, FH_{block}, FH_{mean}, H/FA_{block}, H/FA_{mean}, GV_{block}, E/GV_{block}, E/GV_{mean}) \quad (2)$$

where $\widehat{MS}$ refers to the estimated material stock of buildings, and the building features are defined in Table 2.

4.5 Model application

When applying the model to estimate the BMS, one needs to firstly collect six types of building features, including building type, building year, height, perimeter, floor area, and total floor number. Generally, benefiting from various urban digitalization initiatives, acquiring these data is no longer difficult. Building height, floor area, perimeter, and total floor number can be

obtained from Google Maps, Google Earth, and other available map databases like the iB1000 used in this research. The building type can be derived from land-use databases issued by the Land Department. The building year can be gathered by browsing the website of real estate agencies, calculating the average age of surrounding buildings (Aksözen et al., 2017), or making inferences based on some features (Biljecki & Sindram, 2017).

With the six types of data and the formulas shown in Table 2, the value of the 19 input variables can be smoothly derived and then combined into a matrix, which will be further normalized by using the z-score algorithm (see Equation [1] above). The input variable sequence in the matrix refers to Equation (2). The MATLAB code for calling the model is *EMS = predict(MSmodel_multipleLinearRegression, matrix)*. Here EMS is 'estimated material stock'. It represents the output of material stock estimation by weight. *Predict* is a common MATLAB program language for calling trained models, and *MSmodel_multipleLinearRegression* is the name of the multiple linear regression model developed by this research. We have encoded and shared the BMS quantification model through a GitHub link (Yuan et al., 2022).

5. Discussion

The key to UMS quantification, whether bottom-up or top-down, lies in the accurate estimation of material stocks at the individual building or infrastructure level. The biggest challenge faced by current studies of this kind is the lack of ground-truth data to verify their results, as the data are only available if the building is demolished and its embodied materials are properly measured. Benefiting from a valuable dataset related to 71 demolished buildings in Hong Kong, this research developed a BMS estimation model by using simple and easy-to-obtain building features.

This research makes two contributions. Firstly, it contributes a novel approach to UMS quantification by dealing with it at the building level. With reliable BMS being quantified, it is possible to scale up the estimate to neighbourhood, city, region, and even national scales for applications such as modelling urban metabolism, urban mining, and national resource planning. However, information on overall material stocks is also desired in multiple application scenarios such as stock-enabled socio-economic metabolism analysis (Fishman et al., 2015; Guo et al., 2019), spatiotemporal stock dynamic analysis (Marcellus-Zamora et al., 2016; Mastrucci et al., 2017; Mollaei et al., 2021), and so on. The approach is also able to support microscopic material stock management, e.g., proper deconstruction for cross-project waste material sharing or cross-jurisdiction construction waste material trading. Secondly, the research makes a methodological contribution to the UMS quantification domain. Unlike previous top-down and bottom-up approaches, this research successfully demonstrated that BMS can be accurately estimated by using six easy-to-obtain building features and ML regression algorithms. This provides a novel UMS quantification perspective and, more importantly, contributes a methodology that can be used in other regions.

This study is not free from shortcomings. Firstly, it might be difficult for other interested researchers to obtain similar ground truth data in other cities. Nevertheless, this difficulty is increasingly relieved. More and more cities have started to implement construction waste disposal levy schemes. These schemes may help capture and record various data in a central place. The demolition data like Hong Kong's may just stay with these related government departments without being harnessed. Our research is thus inspiring for researchers to tap into these data sets as "buried assets". Moreover, demolition data sets can also be obtained from demolition companies. We noticed this from previous studies, such as Kleemann et al. (2016) and Sprecher et al. (2021). Secondly, this study does not measure the detailed material composition of the BMS. This is attributable to the fact that the demolition data recorded inert and non-inert portions only. Notably, several leading construction companies in town have started to delve into the detailed compositions for internal material management. There are also research teams trying to use computer vision or other machine learning means to obtain the detailed waste compositions (Chen et al., 2021; Dong et al., 2022). Furthermore, with the accurate overall stock estimate, the specific material stock can also be derived through the following steps: 1) grouping the overall stock according to building cohort classifications, 2) obtaining the typical material composition of each building cohort, and 3) multiplying each group of stocks by corresponding typical material composition to obtain stock by material categories. Lastly, the research did not cover urban infrastructure as another component of UMS. Future studies can expand to this important sector.

6. Conclusion

Previous studies have successfully introduced many urban material stock (UMS) quantification approaches. However, their common crux is to have a relatively accurate estimate of a building material stock (BMS). One cannot ascertain whether his/her estimate of a BMS is accurate before the building is demolished and its embodied materials are separated and weighed. This research aimed to develop a BMS quantification model based on some easy-to-obtain building features without necessarily demolishing and weighing the building. Using 71 demolished buildings in Hong Kong as a valuable sample and machine learning regression techniques, we discovered that six building features, namely (1) building type, (2) building year, (3) height, (4) perimeter, (5) total floor area, and (6) total floor number can satisfactorily predict the BMS of individual buildings. A trained multiple linear regression model has a performance of mean absolute percentage error (MAPE) of 9.1%, root-mean-square error (RMSE) of 474.13, and R-square of 0.93. By summing up the BMS, a reasonably accurate UMS estimate can be derived.

This research provides an innovative solution to UMS quantification, BMS in particular, by using several simple and visible building features. The research is data-driven and rigorous. However, the data-driven approach is like a 'black box' that is not readily accessible to us. Future research is recommended to verify the model in other urban areas when data is available.

Efforts should also be paid to make the BMS quantification model explainable to humans. In addition, it would provide more valuable decision-making information if the BMS covers detailed material compositions. Finally, similar research can be expanded to quantify IMS so as to derive a complete UMS. Further studies are encouraged to cover these limitations in the future.

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.: ECF Project 111/2019) of the Hong Kong SAR Government.

References

Aborujilah, A., Nassr, R. M., Al-Hadhrami, T., Husen, M. N., Ali, N. A., Othmani, A. A., & Hamdi, M. (2020). Comparative Study of SMOTE and Bootstrapping Performance Based on Predication Methods. International Conference of Reliable Information and Communication Technology, 72, 3-9, Springer. Cham.

Aksözen, M., Hassler, U., Rivallain, M., & Kohler, N. (2017). Mortality analysis of an urban building stock. *Building Research & Information*, *45*(3), 259-277.

Augiseau, V., & Barles, S. (2017). Studying construction materials flows and stock: A review. *Resources, conservation and recycling*, *123*, 153-164.

Bergsdal, H., Brattebø, H., Bohne, R. A., & Müller, D. B. (2007). Dynamic material flow analysis for Norway's dwelling stock. *Building Research & Information*, *35*(5), 557-570.

Biljecki, F., & Sindram, M. (2017). Estimating building age with 3D GIS. Proceedings of the 12th International 3D GeoInfo Conference 2017, IV-4/W5, 17-24, ISPRS. Germany.

Brøgger, M., & Wittchen, K. B. (2018). Estimating the energy-saving potential in national building stocks–A methodology review. *Renewable and Sustainable Energy Reviews*, *82*, 1489-1496.

Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology*, *44*(1), 108-132.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16-28.

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.

Chen, X., & Lu, W. (2017). Identifying factors influencing demolition waste generation in Hong Kong. *Journal of Cleaner Production*, *141*, 799-811.

de Tudela, A. R. P., Rose, C. M., & Stegemann, J. A. (2020). Quantification of material stocks in existing buildings using secondary data—A case study for timber in a London Borough. *Resources, Conservation & Recycling: X*, *5*, 100027.

Dong, Z., Chen, J., & Lu, W. (2022). Computer vision to recognize construction waste compositions: A novel boundary-aware transformer (BAT) model. *Journal of environmental management*, *305*, 114405.

Emmert-Streib, F., & Dehmer, M. (2019). Evaluation of regression models: Model assessment, model selection and generalization error. *Machine learning and knowledge extraction*, *1*(1), 521-551.

Ergun, D., & Gorgolewski, M. (2015). Inventorying Toronto's single detached housing stocks to examine the availability of clay brick for urban mining. *Waste management*, *45*, 180-185.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. D. (2021). Regression models. In *Regression* (pp. 23-84). Springer.

Fan, C., Xiao, F., & Wang, S. (2014). Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy*, *127*, 1-10.

Fishman, T., Schandl, H., & Tanikawa, H. (2015). The socio-economic drivers of material stock accumulation in Japan's prefectures. *Ecological Economics*, *113*, 76-84.

Fishman, T., Schandl, H., Tanikawa, H., Walker, P., & Krausmann, F. (2014). Accounting for the material stock of nations. *Journal of Industrial Ecology*, *18*(3), 407-420.

Gal, M. S., & Rubinfeld, D. L. (2019). Data standardization. *New York University Law Review*, *94*, 737.

Gao, X., Nakatani, J., Zhang, Q., Huang, B., Wang, T., & Moriguchi, Y. (2020). Dynamic material flow and stock analysis of residential buildings by integrating rural–urban land transition: A case of Shanghai. *Journal of Cleaner Production*, *253*, 119941.

Gassner, A., Lederer, J., & Fellner, J. (2020). Material stock development of the transport sector in the city of Vienna. *Journal of Industrial Ecology*, *24*(6), 1364-1378.

Gontia, P., Thuvander, L., Ebrahimi, B., Vinas, V., Rosado, L., & Wallbaum, H. (2019). Spatial analysis of urban material stock with clustering algorithms: A Northern European case study. *Journal of Industrial Ecology*, *23*(6), 1328-1343.

Gontia, P., Thuvander, L., & Wallbaum, H. (2020). Spatiotemporal characteristics of residential material stocks and flows in urban, commuter, and rural settlements. *Journal of Cleaner Production*, *251*, 119435.

Guo, J., Miatto, A., Shi, F., & Tanikawa, H. (2019). Spatially explicit material stock analysis of buildings in Eastern China metropoles. *Resources, conservation and recycling*, *146*, 45-54.

Guyon, I., & Elisseeff, A. (2006). An introduction to feature extraction. In I. Guyon, M. Nikravesh, S. Gunn, & L. A. Zadeh (Eds.), *Feature extraction* (pp. 1-25). Springer.

Haas, W., Krausmann, F., Wiedenhofer, D., & Heinz, M. (2015). How circular is the global economy?: An assessment of material flows, waste production, and recycling in the European Union and the world in 2005. *Journal of Industrial Ecology*, *19*(5), 765-777.

Haberl, H., Wiedenhofer, D., Schug, F., Frantz, D., Virág, D., Plutzar, C., Gruhler, K., Lederer, J., Schiller, G., & Fishman, T. (2021). High-resolution maps of material stocks in buildings and infrastructures in Austria and Germany. *Environmental science & technology*, *55*(5), 3368-3379.

Han, J., & Xiang, W.-N. (2013). Analysis of material stock accumulation in China's infrastructure and its regional disparity. *Sustainability Science*, *8*(4), 553-564.

Hashimoto, S., Tanikawa, H., & Moriguchi, Y. (2007). Where will large amounts of materials accumulated within the economy go?–A material flow analysis of construction minerals for Japan. *Waste management*, *27*(12), 1725-1738.

Hashimoto, S., Tanikawa, H., & Moriguchi, Y. (2009). Framework for estimating potential wastes and secondary resources accumulated within an economy–A case study of construction minerals in Japan. *Waste management*, *29*(11), 2859-2866.

Heeren, N., & Hellweg, S. (2019). Tracking construction material over space and time: Prospective and geo-referenced modeling of building stocks and construction material flows. *Journal of Industrial Ecology*, *23*(1), 253-267.

Heisel, F., McGranahan, J., Ferdinando, J., & Dogan, T. (2022). High-resolution combined building stock and building energy modeling to evaluate whole-life carbon emissions and saving potentials at the building and urban scale. *Resources, conservation and recycling*, *177*, 106000.

HKLD. (2022a). *1:1000 Digital Topographic Map*. Retrieved June 20 from https://www.hkmapservice.gov.hk/OneStopSystem/map-search?product=OSSCatB&series=iB1000

HKLD. (2022b). *Hong Kong Outline Zoning Plans Land Use Zonings*. Retrieved 20, July from https://www.arcgis.com/home/webmap/viewer.html?webmap=5375a88ec76143ea974d5fb64efbec0d

HKPD. (2021). *Land Utilization in Hong Kong*. Retrieved June 20 from https://www.arcgis.com/home/webmap/viewer.html?webmap=5375a88ec76143ea974d5fb64efbec0d

Huang, T., Shi, F., Tanikawa, H., Fei, J., & Han, J. (2013). Materials demand and environmental impact of buildings construction and demolition in China based on dynamic material flow analysis. *Resources, conservation and recycling*, *72*, 91-101.

Ij, H. (2018). Statistics versus machine learning. *Nat Methods*, *15*(4), 233.

Karystinos, G. N., & Pados, D. A. (2000). On overfitting, generalization, and randomly expanded training sets. *IEEE Transactions on Neural Networks*, *11*(5), 1050-1057.

Kleemann, F., Lederer, J., Aschenbrenner, P., Rechberger, H., & Fellner, J. (2016). A method for determining buildings' material composition prior to demolition. *Building Research & Information*, *44*(1), 51-62.

Kleemann, F., Lederer, J., Rechberger, H., & Fellner, J. (2017). GIS-based analysis of Vienna's material stock in buildings. *Journal of Industrial Ecology*, *21*(2), 368-380.

Krausmann, F., Wiedenhofer, D., Lauk, C., Haas, W., Tanikawa, H., Fishman, T., Miatto, A., Schandl, H., & Haberl, H. (2017). Global socioeconomic material stocks rise 23-fold over the 20th century and require half of annual resource use. *Proceedings of the National Academy of Sciences*, *114*(8), 1880-1885.

Lal, T. N., Chapelle, O., Weston, J., & Elisseeff, A. (2006). Embedded methods. In *Feature extraction* (pp. 137-165). Springer.

Lanau, M., & Liu, G. (2020). Developing an urban resource cadaster for circular economy: A case of Odense, Denmark. *Environmental science & technology*, *54*(7), 4675-4685.

Lanau, M., Liu, G., Kral, U., Wiedenhofer, D., Keijzer, E., Yu, C., & Ehlert, C. (2019). Taking stock of built environment stock studies: Progress and prospects. *Environmental science & technology*, *53*(15), 8499-8515.

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.

Lu, W., Peng, Y., Chen, X., Skitmore, M., & Zhang, X. (2016). The S-curve for forecasting waste generation in construction projects. *Waste management*, *56*, 23-34.

Luor, D.-C. (2015). A comparative assessment of data standardization on support vector machine for classification problems. *Intelligent Data Analysis*, *19*(3), 529-546.

Manelius, A.-M., Nielsen, S., & Kauschen, J. S. (2019). City as Material Bank–Constructing with Reuse in Musicon, Roskilde. IOP Conference Series: Earth and Environmental Science, 225(1), 012020, IOP Publishing. Brussels.

Mao, R., Bao, Y., Huang, Z., Liu, Q., & Liu, G. (2020). High-resolution mapping of the urban built environment stocks in Beijing. *Environmental science & technology*, *54*(9), 5345-5355.

Marcellus-Zamora, K. A., Gallagher, P. M., Spatari, S., & Tanikawa, H. (2016). Estimating materials stocked by land-use type in historic urban buildings using spatio-temporal analytical tools. *Journal of Industrial Ecology*, *20*(5), 1025-1037.

Mastrucci, A., Marvuglia, A., Popovici, E., Leopold, U., & Benetto, E. (2017). Geospatial characterization of building material stocks for the life cycle assessment of end-of-life scenarios at the urban scale. *Resources, conservation and recycling*, *123*, 54-66.

MathWorks. (2022). *Feature Selection and Feature Transformation Using Regression Learner App*. Retrieved 20, July from https://reurl.cc/V13oD6

Maués, L. M. F., do Nascimento, B. d. M. O., Lu, W., & Xue, F. (2020). Estimating construction waste generation in residential buildings: A fuzzy set theory approach in the Brazilian Amazon. *Journal of Cleaner Production*, *265*, 121779.

Mesta, C., Kahhat, R., & Santa-Cruz, S. (2019). Geospatial characterization of material stock in the residential Sector of a Latin-American City. *Journal of Industrial Ecology*, *23*(1), 280-291.

Miatto, A., Schandl, H., Forlin, L., Ronzani, F., Borin, P., Giordano, A., & Tanikawa, H. (2019). A spatial analysis of material stock accumulation and demolition waste potential of buildings: A case study of Padua. *Resources, conservation and recycling*, *142*, 245-256.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Mollaei, A., Ibrahim, N., & Habib, K. (2021). Estimating the construction material stocks in two Canadian cities: A case study of Kitchener and Waterloo. *Journal of Cleaner Production*, *280*, 124501.

Müller, D. B. (2006). Stock dynamics for forecasting material flows—Case study for housing in The Netherlands. *Ecological Economics*, *59*(1), 142-156.

Nasir, U., Chang, R., & Omrany, H. (2021). Calculation Methods for Construction Material Stocks: A Systematic Review. *Applied Sciences*, *11*(14), 6612.

Nasiri, B., Piccardo, C., & Hughes, M. (2021). Estimating the material stock in wooden residential houses in Finland. *Waste management*, *135*, 318-326.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, *4*, 51-62.

Ortlepp, R., Gruhler, K., & Schiller, G. (2016). Material stocks in Germany's non-domestic buildings: a new quantification method. *Building Research & Information*, *44*(8), 840-862.

Peled, Y., & Fishman, T. (2021). Estimation and mapping of the material stocks of buildings of Europe: a novel nighttime lights-based approach. *Resources, conservation and recycling*, *169*, 105509.

Schandl, H., Marcos-Martinez, R., Baynes, T., Yu, Z., Miatto, A., & Tanikawa, H. (2020). A spatiotemporal urban metabolism model for the Canberra suburb of Braddon in Australia. *Journal of Cleaner Production*, *265*, 121770.

Seyedzadeh, S., Rahimian, F. P., Glesk, I., & Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: a review. *Visualization in Engineering*, *6*(1), 1-20.

Sharma, S., Gosain, A., & Jain, S. (2021). A Review of the Oversampling Techniques in Class Imbalance Problem. International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, 1387(1), Springer. Singapore.

Song, M.-K., Lin, F.-C., Ward, S. E., & Fine, J. P. (2013). Composite variables: when and how. *Nursing research*, *62*(1), 45.

Sprecher, B., Verhagen, T. J., Sauer, M. L., Baars, M., Heintz, J., & Fishman, T. (2021). Material intensity database for the Dutch building stock: Towards Big Data in material stock analysis. *Journal of Industrial Ecology*.

Stephan, A., & Athanassiadis, A. (2017). Quantifying and mapping embodied environmental requirements of urban building stocks. *Building and Environment*, *114*, 187-202.

Surahman, U., Higashi, O., & Kubota, T. (2017). Evaluation of current material stock and future demolition waste for urban residential buildings in Jakarta and Bandung, Indonesia: embodied energy and CO 2 emission analysis. *Journal of Material Cycles and Waste Management*, *19*(2), 657-675.

Tanikawa, H., Fishman, T., Okuoka, K., & Sugimoto, K. (2015). The weight of society over time and space: A comprehensive account of the construction material stock of Japan, 1945–2010. *Journal of Industrial Ecology*, *19*(5), 778-791.

Tanikawa, H., & Hashimoto, S. (2009). Urban stock over time: spatial material stock analysis using 4d-GIS. *Building Research & Information*, *37*(5-6), 483-502.

Wang, H., Chen, D., Duan, H., Yin, F., & Niu, Y. (2019). Characterizing urban building metabolism with a 4D-GIS model: A case study in China. *Journal of Cleaner Production*, *228*, 1446-1454.

Wiedenhofer, D., Fishman, T., Lauk, C., Haas, W., & Krausmann, F. (2019). Integrating material stock dynamics into economy-wide material flow accounting: concepts, modelling, and global application for 1900–2050. *Ecological Economics*, *156*, 121-133.

Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, *1168*(2), 022022.

Yuan, H., Lu, W., & Hao, J. J. (2013). The evolution of construction waste sorting on-site. *Renewable and Sustainable Energy Reviews*, *20*, 483-490.

Yuan, L., Lu, W., Xue, F., & Li, M. (2022). *Material stock estimation model*. Retrieved June 20 from https://github.com/Lanny-yuan/Material-stock-estimation-model

Conflict of Interest Statement

The authors declare no conflict of interest.

Data Availability Statement

The building material stock data is available on request from the corresponding author. The data is not publicly available due to privacy or ethical restrictions. The building feature-related data were derived from the following resources available in the public domain:

(1) iB1000 Digital Map [https://www.hkmapservice.gov.hk/OneStopSystem/map-search?product=OSSCatB&series=iB1000];

(2) Land utilization type database of Hong Long [https://www.pland.gov.hk/pland_en/info_serv/open_data/landu/];

(3) Google Earth [https://earth.google.com/]; and

(4) Housing agency websites such as SPACIOUS [https://www.spacious.hk/en/hong-kong].

Supporting information

Supporting information can be found in the online version of the article at the publisher's website.