detritus
Multidisciplinary
Journal for Waste
Resources & Residues

CISA

BIG DATA IN CONSTRUCTION WASTE MANAGEMENT: PROSPECTS AND CHALLENGES

Weisheng Lu [1,*], Chris Webster [2], Yi Peng [3], Xi Chen [1] and Ke Chen [1]

[1] Department of Real Estate and Construction, Faculty of Architecture, The University of Hong Kong, Pokfulam, Hong Kong
[2] Faculty of Architecture, The University of Hong Kong, Pokfulam, Hong Kong
[3] School of Public Administration, Zhejiang University of Finance & Economics, Hangzhou, PR China

Article Info:
Received:
6 July 2018
Revised:
12 November 2018
Accepted:
14 November 2018
Available online:
22 November 2018

Keywords:
Big data
Construction waste management
Big data analytics
Hong Kong

ABSTRACT

'Big data' has been rapidly sprawling in various research disciplines such as biology, ecology, medical science, business, finance, and public governance but rarely in construction waste management (CWM). The CWM community around the world generally relies on 'small data' collected via active solicitation such as sampling and ethnographic methods. This small data is intrinsically limited by its inability to account for the totality of CWM and research findings generated from the small data cannot be accepted with a high level of confidence. With the growing interests in big data, it can be reasonably expected that the waste management community will augment efforts to develop big data and its analytics. However, the efforts are currently constrained by the limited knowledge to do so. This research aims to provide a synoptic overview of the prospects and challenges of big data in CWM. It adopts an inductive, qualitative case study method whereby the empirical data is collected using an ethnographic–action-meta-analysis research approach and triangulated with data from literature, ongoing debate, and other sources. The paper offers some insights on big data acquisition, storage, analytics, implementation, and challenges. Although having a focus on waste management in the construction sector, the insights generated from this study can be of value to general waste management research, which suffers from the same problems of erratic and poor quality data as CWM.

1. INTRODUCTION

A consensus has yet to be reached on what is meant by 'big data'. According to Padhy (2013), big data is a collection of data sets so large and complicated that it becomes difficult to process using traditional data management tools. Likewise, Schönberger and Cukier (2013) proposed big data as "things one can do at a large scale that cannot be done at a smaller one, to create a new form of value". Researchers tend to adopt Gartner's three defining characteristics of big data, namely, volume, variety, and velocity, or the three 'Vs' (McAfee et al., 2012). Volume is the quantities of data in the forms of records, transactions, tables, or files; velocity can be expressed in batch, near time, real time and streams; and variety can be structured, unstructured, semi-structured and a combination thereof (Russom, 2011; Zaslavsky et al., 2013). Data is relentlessly generated from such sources as web logs, sensor networks, unstructured social networking, and streamed video and audio. Analytics have been developed to analyze big data in order to uncover hidden patterns, unknown correlations and other useful information that will guide better business predictions and decision-making (Shen et al., 2016); in effect, value is advocated as the fourth 'V'.

Notwithstanding the disagreement on terminology, big data has rapidly become the new frontier across a wide variety of fields, including biology, medical science, ecological science, business, urban planning, public governance, innovation, competition, and productivity. "Government agencies use big data to generate statistics, to help them understand local and global patterns and trends, in order to improve their services" (Shen et al., 2016). Using its ability to harness information in novel ways to create insights and services, big data can become a crucial source of innovation (Schönberger and Cukier, 2013). Through analyzing big data, researchers aim at identifying some 'latent knowledge' (Agrawal, 2006) or 'actionable information' (World Economic Forum, 2012), which can be utilized for future decision-making.

However, the euphoria of big data is yet to be seen in the waste management research community. This is particularly held in construction waste management (CWM), where research is suffering from notoriously erratic 'small' data. One explanation for this is the temporary nature of

* Corresponding author:
Weisheng Lu
email: wilsonlu@hku.hk

https://doi.org/10.31025/2611-4135/2018.13737

construction projects (Senaratne and Rasagopalasingam, 2017), whereby once a project is completed it ceases to generate construction waste and the window of opportunity to collect the data closes. The data collection methods adopted by previous CWM studies involved sampling and ethnographic methods during construction processes, such as: direct observation (Poon et al., 2001); questionnaire survey (McGregor et al., 1993); sorting and weighing the waste materials on-site (Bossink and Brouwers, 1996; Kazaz et al., 2018); collecting data through consultation with construction employees (Treloar et al., 2003); tape measurement (Skoyles, 1976); and truck load records (Poon et al., 2004). These data collection approaches are widely perceived as costly, non-value added, and disruptive to the ongoing construction process. Hence, in practice construction companies are not obliged to record and report the characteristics of the waste generated (Fatta et al., 2003; Lu et al., 2017). Most studies have a relatively small sample size or sampled relatively small sites due to the difficulties of covering the whole population. As such, their data has long been limited by its inability to comprehensively represent the totality of waste generated throughout the construction process.

Nevertheless, with the vogue of big data in other disciplines, researchers have started to explore its applications to CWM. For example, Lu et al. (2015) revisited waste generation rates (WGRs) as performance indicators of CWM using big data, which allowed them to say with greater confidence that there is a notable CWM performance disparity between the public and private sectors (Lu et al., 2016a); Chen and Lu (2017) identified factors influencing demolition waste generation in Hong Kong through big data analytics; Bilal et al. (2016a) proposed a conceptual framework of big data architecture for construction waste analytics. However, the general sentiment is that understanding of big data in CWM is still rather superficial. There is a plethora of bestsellers and online articles eloquently promoting the use of big data, but impartial, skeptical insights preferred by researchers are rare. A succession of questions remains unanswered, such as 'What are the potentials of big data for CWM?'; 'Is there a definite size over which a dataset can be called big data?'; 'Will it be financially viable to purposely develop big data for CWM?'; and 'What are the main challenges of big data in CWM?'.

This paper explores the prospects and challenges of big data in CWM, with a view to facilitating pursuit of the research agenda related to big data and its analytics in the realm of CWM and beyond. The remaining sections of the paper report: the research methods used in the study, organized in the form of a CWM case study; a description of big data as a basis as used in our case study; a presentation of the results and findings; an in-depth discussion; and, conclusions. Although a particular big data set of CWM in Hong Kong is described, it is suggested that the analysis yields generalizable insights that are independent of this data set and the CWM setting per se.

2. RESEARCH METHODS

Since only a limited number of studies had been reported with this focus at this point in time, this paper adopts a mixed-methods approach with an inductive, qualitative case study (Yin, 1989) at the kernel of the research methodology. Unlike the stereotype that it may have the problems of generalization, case study approach is widely used in management research and considered useful to promote scientific development through deepening understanding of the context and relevant experiences. Over the past five years, the authors have endeavored to investigate CWM performance by taking real actions in acquiring and analyzing CWM big data in the specific context of Hong Kong. Several papers about CWM performance have been published. During the research, it is noticed that some of the insights of big data analytics can be drawn from the action research and contribute to the wider knowledge body of big data in a more general setting. Therefore, other members of the research team, with a strong humanity and sociology background, took an ethnographic approach to observe the "action researchers", e.g. how they collect the data, interacting with practitioners or other researchers. They observed and analyzed from a distance. They conducted meta-analyses of the published papers by "hovering" from the specific research findings on CWM performance but induced some propositions of the prospects and challenges of big data in CWM as a setting. They triangulated the propositions against new literature, ongoing debate, and finally form the insights that are generalizable to general waste management realm or beyond.

2.1 The data set

To effectively manage construction waste in Hong Kong, a Construction Waste Disposal Charging Scheme (CWDCS) was enacted in 2006 based on the polluter pays principle (Lu and Tam, 2013). According to the Scheme, a contractor should pay HK$125 per ton for non-inert construction waste that is accepted by landfills; HK$100 per ton of mixed inert and non-inert waste material received by off-site sorting facilities; and HK$27 per ton of inert construction waste material ending up in public fill reception facilities (Hong Kong Environment Protection Department - HKEPD, 2014). Under this Scheme, contractors must send their construction waste to the government-run facilities if not otherwise reduced, reused, or recycled. Every truckload of construction waste ending up in any of the facilities leaves a record with the HKEPD. Waste disposal facilities record information on every load of construction and demolition (C&D) waste received from every construction/demolition site. This practice leads to a database of more than one million transaction records in a year, which is considered a full coverage of the waste generated from all construction sites in Hong Kong. The Scheme also requires all contractors involved in C&D activities to open a billing account with the information of the activities also recorded by the HKEPD. These records form the account information database, which includes account number, construction name, category, site, and contract sum of all C&D projects in Hong Kong. A third database is information about the disposal facilities, which includes facility name, received waste type, and facility address, and a fourth database is the information of all the vehicles, including their license plate number and the permitted gross weight they can carry. The links between the four databases are shown in Figure 1.

The three defining characteristics of big data, i.e. volume, velocity, and variety seem evident in the data set. The data is of considerable volume. The main database contains more than 6 million well-structured waste disposal records, recording almost every truck load of C&D waste generated from construction sites and disposal at the designated CWM facilities over the past six years. Although the total physical volume only slightly exceeds 700 megabytes, we argue this is 'big data', given it is a well-structured data set that may contain much more meaningful information than the same volume of messy, raw data. The data has significant velocity. The data in the main database is incoming as a rate of about 4,000 records per day. In addition to the rich data fields exhibited above, the variety of the data set is still expanding, e.g. by collecting more details on new, renovation, or demolition projects from the government Buildings Department (HKBD) and linking green building information publicly available in the Hong Kong Green Building Council (HKGBC) and other potential databases to the main databases in the future (See Figure 1). Therefore, volume, and velocity and variety are all significantly high and dynamic in this data set.

2.2 Obtaining and analyzing the big data

To obtain the data, the research team approached the HKEPD through its general inquiry service, followed by emails clarifying the specific data requested and what it will be used for. After the initial request had been granted, the HKEPD advised the research team that it would be more convenient to obtain further data from its themed website where data is updated every fortnight. For security reason, the data was stored in the cloud data service of The University of Hong Kong (HKU) and mapped on the hard disk drive of two computers for further use. Use of the data is governed by general research ethics and HKU's policy on the management of research data and records.

Over a period of five years the research team conducted a series of studies to analyze the acquired big data using various statistical analyses and data mining, with results published in journals or shared at international conferences. This study used these research experiences as a case study to extract the general prospects and challenges of big data in CWM. In addition, preliminary findings from the case study were triangulated with the literature of big data in other fields. The following sections present the critical reflections from this study.

3. ANALYSES AND DISCUSSIONS

3.1 "In God we trust; all others please bring data"

The famous quotation is widely attributed to Edwards Deming to reflect his fundamental principle of using data to back up any decisions in production or business. It further reinforces the truism about the importance of data to scientific research, where quantification is believed to generate high forms of knowledge in social sciences (Shelton, 2017). Unlike other pollutants such as dust and noise, C&D waste is easy to see, as well as relatively easy to measure (Formoso et al., 2002), albeit not so easy to sample systematically. No other method is more reliable than directly measuring construction waste generation in order to obtain primary data. However, contractors are usually not mandated to record waste generation data on site and generally perceive doing so as disruptive and as not adding value to the ongoing construction process. Contractors therefore record waste generation sporadically, if at all, and so it is not feasible to expect them to be the source of such data.

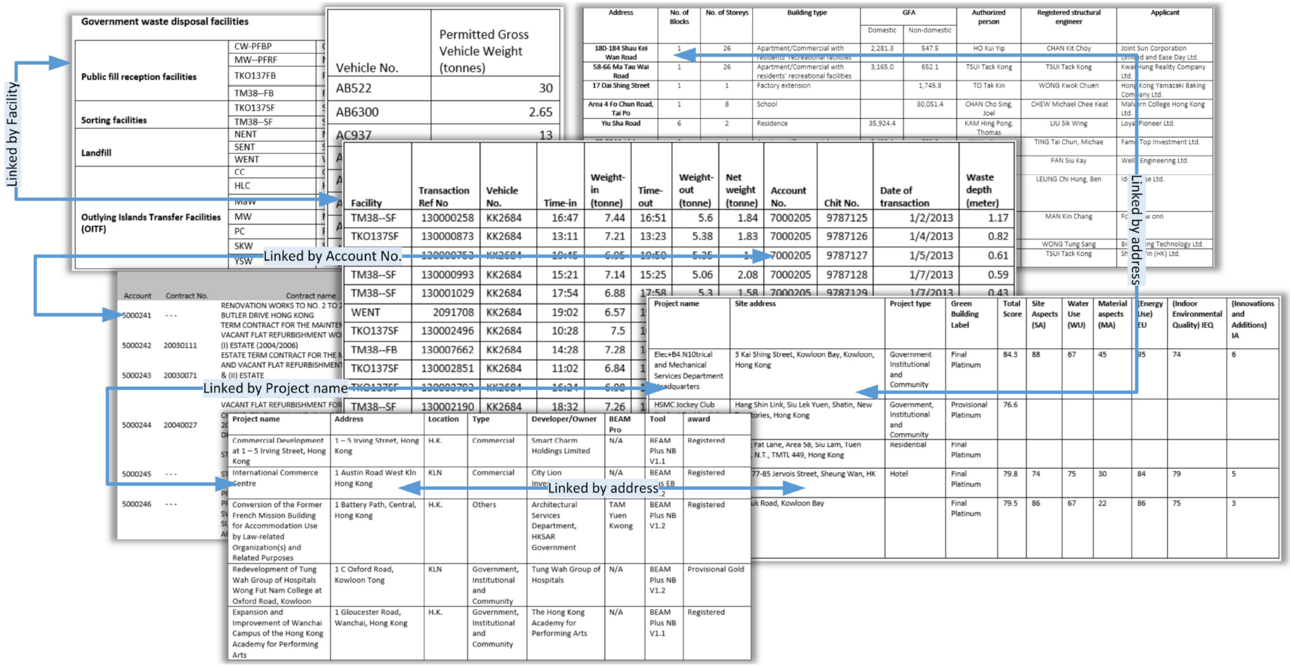

FIGURE 1: Links between the databases relating to construction waste management in Hong Kong.

An alternative method of data collection is for researchers themselves to measure tangible waste generation. Research assistants can be dispatched to construction sites to collect primary data by recording actual waste generated, e.g. measuring cement waste, counting bricks ordered, left, and wasted, or checking the materials used against orders. Given the fact that a construction project will last for a relatively long period of time ranging from a few months to years, it is impractical for these research assistants to station on a site to record all the waste generation data relating to a project. Instead, sampling is often adopted to make onsite inspection more tractable, e.g. measuring the waste generation from a typical section or a floor. Aside from the limited data collected, the big issue with sampling is whether the data are sufficiently comprehensive and representative, i.e. the issue of sampling frame and method. Because each construction project is unique, in principle, there is no a priori method or theory for constructing a sampling frame and each sampling effort is thus necessarily ad hoc.

The situation has changed little since Lu et al. (2011b) described a sampling-based data collection approach: "When a trade had finished, the site manager cordoned off an area of the construction site to facilitate the on-site measuring exercises. The area (usually a room plus a section of common walkway) was selected as being representative of a typical floor so that the WGRs derived from that area could be applied to the whole floor." There are two points of weakness with such approaches. First, if too few sampling sites are selected, the estimate cannot be treated as accurate, even if it is an appropriately chosen site. Second, if several or many sites are chosen, then if they are not sufficiently representative of the whole construction process, the estimate would be systematically biased. Neither the degree of accuracy nor the reliability of the measure and the derived WGRs are too variant to be generalized to other projects. Katz and Baum (2011) and Lu et al. (2016b) noted that most previous studies on CWM, even though using objective methods such as weighing waste onsite, had a relatively small sample size or sampled relatively small sites due to the difficulties involved in conducting a full coverage survey, whatever 'full coverage' might mean. These studies are thus limited in their ability to account for the totality of waste generation throughout the construction process, and as a consequence, their results cannot be accepted with a high level of confidence.

In view of the ongoing emphasis on the importance of data to CWM research, it can be expected that researchers and construction companies themselves, will intensify efforts to collect more reliable and representative data (Bilal et al., 2016a). With better sensing and recording technology, CWM systems are expected to emerge that no longer rely on sampling. This is analogous to other big data domains in which routinely sensed and stored data are replacing occasionally collected data. For example, occasional surveys of shoppers at supermarkets to obtain customer profiles has been replaced by data collected at electronic points of sale. We are moving to an era in which the researcher's task is not so much to sample from the real world but to sample from a database that is a complex and voluminous model of the real world. A CWM big data source will provide something approximating full coverage rather than a sample of the population of interest. With continuing advances in data acquisition technologies and the lowering of data processing costs, collecting big data is becoming ever more feasible. In the future, it might be required that "all others please bring big data" to the CWM community. Kitchin and Lauriault (2015) echoed by contending that big data in the future will become as common as small data is in today's research. It is therefore expected that big data regarding CWM in other regions and countries would emerge, although they currently have no such structured databases of CWM big data as the one reported in Section 2 of this paper.

3.2 Size does matter

There are two basic premises behind the exhortation of investing in big data. First, the large volume of big data can alleviate the potential bias inherent in small data and provide a fuller picture so as to have a closer claim of objective truth (Bilal et al., 2016a). Second, by analyzing big data it is possible to discover hidden patterns, unknown correlations and other useful actionable information that will help with devising more informed CWM approaches. With its characteristics of volume, velocity, and variety, analyses of big data can lead to actionable information that would not be possible to discover with small data. Our mining of CWM-related big data in Hong Kong illustrates these points.

WGR is widely accepted as a CWM performance indicator, which is calculated by dividing waste generation in volume (m3) or quantity (tons) by per m2 of gross floor area (Poon et al., 2004) or by per million US$'s worth of construction work (Lu et al., 2015). The lower the WGR, the better the CWM performance. Without readily available secondary data relating to waste generation (e.g. volume or quantity of waste), researchers have to use sample and ethnographic methods to collect the data from the project to calculate WGR, as described above in Soibelman (2016) and Lu et al. (2011b). However, in this study, every truckload of C&D waste generated from all the construction sites over the past six years was recorded by the HKEPD. The WGRs (ton/mHK$) of all Hong Kong's 4,062 sites are plotted in Figure 2. Every dot in the figure represents a project with its WGR calculated by summing the truckloads of waste and then dividing that figure by the contract sum of the project.

Using a sampling method, researchers would only be able to collect a limited portion of the dozens or hundreds of loads of waste as depicted in Figure 2, with the burden to justify whether the sample represents the holistic waste generation pattern of the project. This leads to a situation similar to the ancient parable of "blind men and an elephant" – the researcher is at the risk of probing into only a small scope of actual waste generation from all the construction projects. This leads to uncertainty over whether the data collected is sufficiently comprehensive and representative.

Size definitely does matter in this case because big data can portray a fuller picture of C&D waste generation, and the calculated WGRs converge to a range. This is supported by the law of large numbers: the average of the results obtained from a large number of trials tend to become convergent to a certain value as more trials are performed (Sen and Singer, 1993; Shen et al., 2011). The advantage of big data over small data allows more in-depth analyses of WGRs. Given the abundant data covering the waste generation from all the projects, it might be considered rigorous to average the WGRs and derive a mean to represent the general CWM performance. However, after having calculated and plotted the frequency distribution WGRs of all the projects (see Figure 3), it was found that the distribution is far from a normal one but rather a heavily skewed distribution. The median of the group of WGR, 15 t/mHK$ is much lower than that of the mean of 76 t/mHK$ (see Table 1). Using the mean to represent the general CWM performance is thus very misleading, which however is a common problem in existing CWM research with small data. Without big data covering the whole population, this insight would have been difficult to discover.

Bigger data size also allows some hidden patterns, unknown correlations and other useful information to be discovered (Zhou et al., 2016). For example, by analyzing one day's waste disposal records randomly selected from the 6 years' pool, it is discovered that a considerable number (734 out of 4780) of waste haulers tend to overload than their permitted load weight (See the red dots in Figure 4). Transporting the waste is charged by trips and it is often costly, sometimes costlier than the waste disposal charge itself. Tracing individual lorries may reveal the ones that are consistently involved in this overloading as an unsafe behavior so that they can be more closely monitored or possibly be subjected to legal action. Meanwhile, as shown by green dots in Figure 4, often lorries are underloaded (the

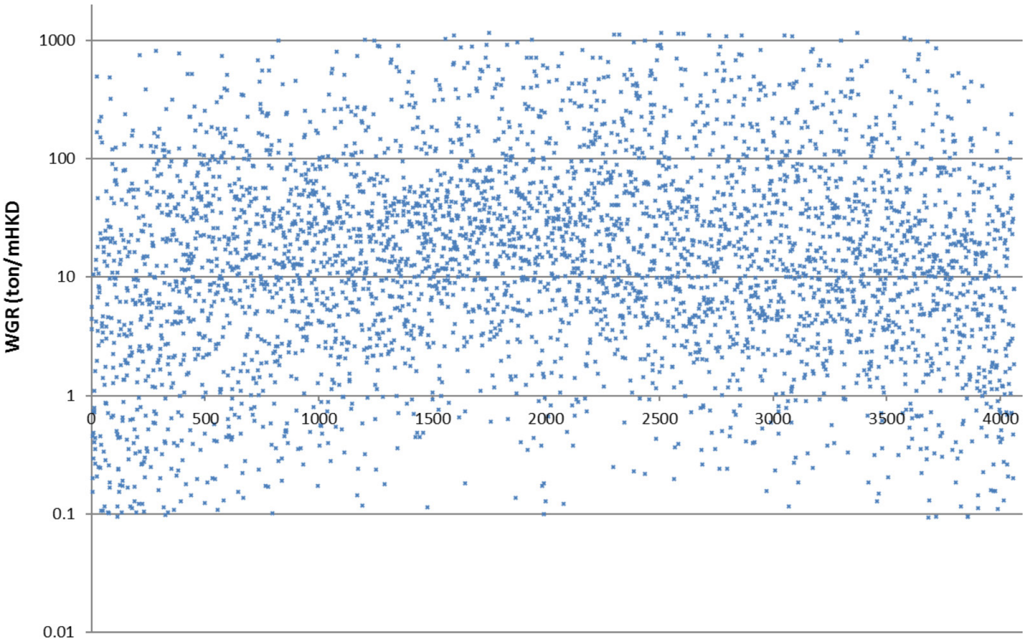

FIGURE 2: WGRs of the individual projects (Sample size=4,062).

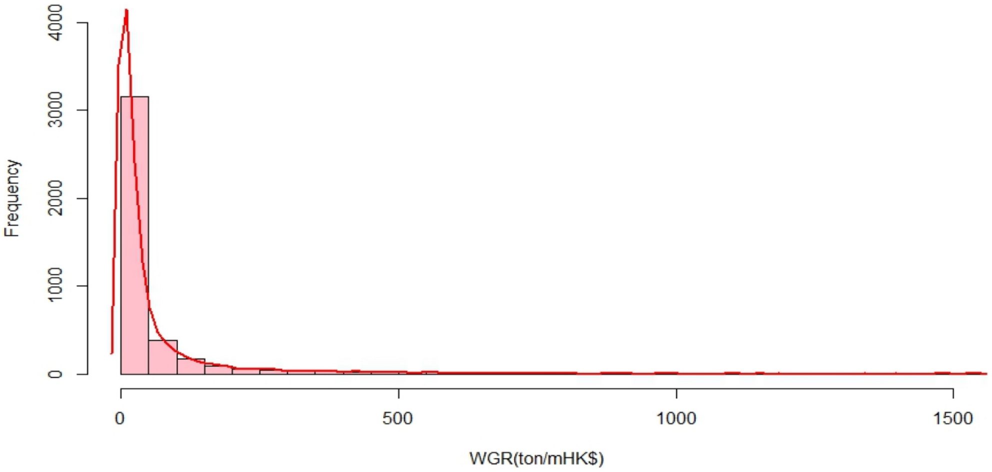

FIGURE 3: Frequency distribution of WGRs of all projects (Sample size=4,062).

TABLE 1: Means, standard deviations (SD), and medians of WGRs of the projects.

Projects	Sample size	Mean (t/mHK$)	SD	Median (t/mHK$)	Range (t/mHK$)
Overall	4062	76	192	15	0.13~1793.33

lower the point, the more underloaded a lorry is), which is more likely than due to poor fleet management. Likewise, by further analyzing the WGRs of the individual projects, it is found that a handful of companies achieved consistently low WGRs, such as Company A in Figure 5. Perhaps these companies are truly good at managing C&D waste, in which case their experiences should be disseminated to the whole industry. On the other hand, the WGRs of Company B are consistently high suggesting that a review of the company's poor performance might be advisable. This kind of useful actionable information can only be revealed with big data.

3.3 How big is big data? The relativeness of big data

There is a misconception among the CWM community that to be considered big data, a dataset should be in

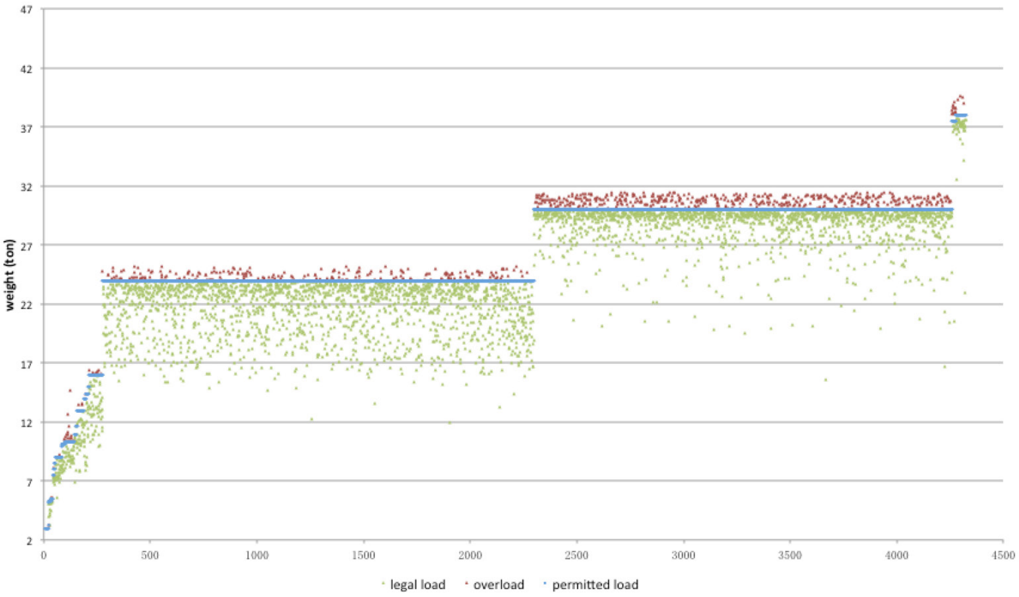

FIGURE 4: Pattern of overloaded or underloaded of waste haulers in one day (Sample size=4,780 truck loads).

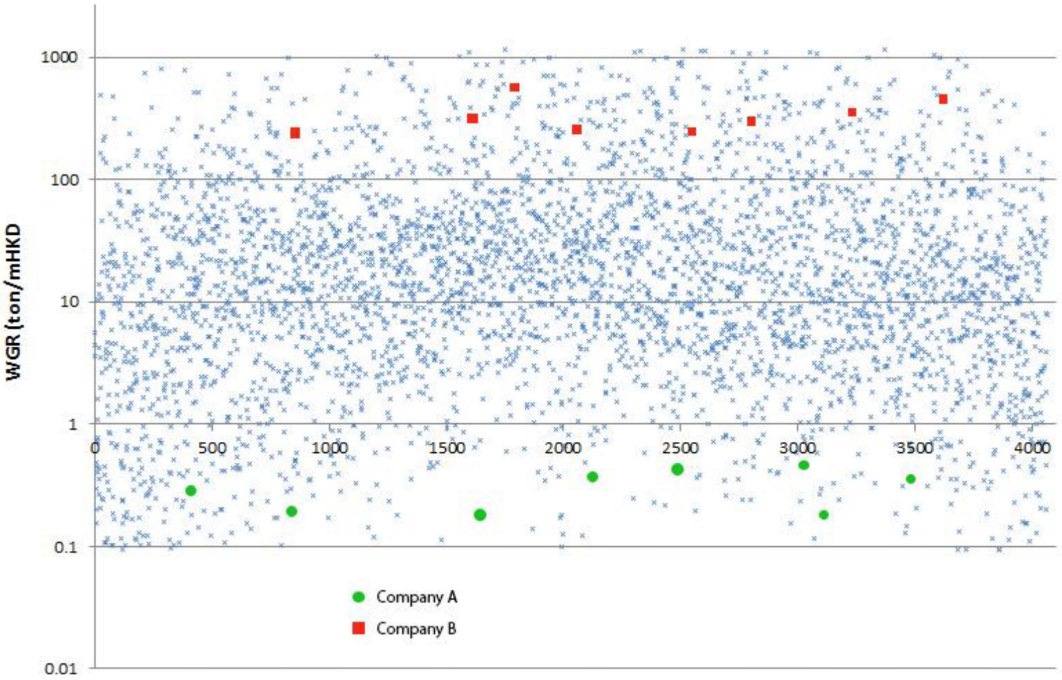

FIGURE 5: Unusual construction waste management performance using big data analytics.

terabytes or petabytes. This misconception is evident in debates in international conferences, comments from journal paper reviewers, and reviewers of research grant applications. These commentators and reviewers at times simply judge that some datasets are not big data on the basis of the data's size in electronic format (i.e. megabyte, gigabyte, or the like). Press (2013) asked, "Is there a definite size over which data becomes big data? How big is big data?"

It is argued that big data is a relative concept. In particular, big data is time relative. A dataset that appears to be massive today will almost surely appear small in the near future (MIT Technology Review, 2013). Current big data may be considered small data in the future due to the rapid development of technology, particularly in the area of cloud-based data storage and retrieval. Here, the CWM data compared to the 1990s is 'big', and the data in the 1990s was considered big data compared to the 1970s. Big data is also user relative. A dataset treated as big data by one entity may be considered 'small' by another depending on its intended use. For example, the CWM dataset in Hong Kong may be treated as big data by researchers interested in construction management, urban planning, or transportation, since it can provide many useful insights. However, it may not be considered as big for the purpose of either estimating total construction waste generation in China. Some researchers (e.g. Sivarajah et al., 2017) thus highlighted the intricacy of a dataset as a significant factor in determining whether it is big. Big data does not necessarily always mean better data (Taylor and Schroeder, 2015). Leek (2014) pointed out that in general the bigger the sample size, the better, but that meaningful sample size and raw data size are not always tightly correlated (Akter and Wamba, 2016). Large datasets from Internet sources are often unreliable, prone to outages and losses, and errors and gaps are magnified when multiple datasets are used together (Boyd and Crawford, 2012). There is also a lot of noises in this CWM dataset that must be excluded (Lu et al., 2015). Data cleansing helps detect and correct incomplete, incorrect, inaccurate or irrelevant parts of the raw data and allows users to perceive a dataset's true size, value, and relevance to a particular research inquiry.

It can be concluded that although size does matter, there is no definitive size at which a dataset can be called big data and it most definitely does not mean that a dataset must be in the volume of a terabyte or petabyte. The point is to examine the dataset's ability to account for the totality of the subject under investigation and determine whether it allows values to be created that could not be arrived at with data on a smaller scale. Conference discussants and paper reviewers should not take the automatic position that the data smaller than terabytes or petabytes is definitely not big data. One more useful definition of big data suggested by this research so far, is that big data can be perceived as a data model of a system, its dynamics in particular, that can be analyzed in totality or by systematic or random sampling to identify and interrogate trends. The 'big', in this definition, refers not to absolute size of data stored but to the degree of coverage of the data model vis-a-vis the system being represented.

3.4 Big data analytics: applied statistics vs data mining

Another misconception found in the big data literature is that big data analytics is equated to 'pattern finding algorithms', 'unattended machine learning', 'deep learning', 'artificial intelligence', NoSQL database, Hadoop, and other fascinating methods (Bilal et al., 2016b). Traditional applied statistics, as Leek (2014) argued, has been largely left out in the discussion. Traditional statistical analyses can be used for multiple purposes, e.g. describing the nature of the data, exploring the relation of the data, creating a model, proving or disapproving the validity of a hypothesis, and so on (Moses, 1986). Arguably, one of the purposes of sophisticate data mining techniques is to search and structure a large and unwieldy data base in such a way that makes possible the use of traditional statistical methods designed to formally describe data in ways that are scientifically well understood.

Data mining is useful for automatically discovering valuable information from a large collection of data and transforming it into organized knowledge (Han et al., 2012). Rather than simply locating, identifying, understanding and citing data, data mining serves as a computational process where patterns in large datasets can be discovered (Clifton, 2010). Other approaches such as pattern finding algorithms and unattended machine learning are also useful, although they have been over stated by the media to such an extent that it gives the illusion that they are the only approaches appropriate for exploiting the value of big data. It is the experience of the research team, when mining the CWM-related big data in Hong Kong, that purely relying on machine intelligence is ineffective at best. Predictions and human intervention can save a large amount of computational time and increase the effectiveness of data mining. As a formalized procedure it is advisable to plot big data, using 'data visualization' methods, before engaging in any data mining techniques (Kostelnick, 2007). The intention is to observe potential patterns and provide a direction for the subsequent data mining. This is essential a human intervention process.

Both traditional statistics and data mining are indispensable means of harnessing the power of big data, although there are challenges with both techniques. A typical challenge is to select statistical indicators to interpret the results (Ekbia et al., 2015). Traditionally, in small data analysis, p-value is commonly used in the context of null hypothesis testing to indicate the statistical significance of evidence. If the p-value is less than or equal to the chosen significance level, either 5% or 1%, the test suggests that the observed data is inconsistent with the null hypothesis, so the null hypothesis must be rejected. Although this method is disputed (Goodman, 1999; Wasserstein and Lazar, 2016), it is commonly used as a license for making a claim of a scientific finding or implied truth in numerous fields including CWM (Wetzels et al., 2011). However, in very large samples, p-values go quickly to zero, and solely relying on it can lead the researcher to claim support for results of no practical significance (Lin et al., 2013). In this context, data interpretation and discussion become more sophisticated and care is need in making claims.

Researchers should also keep a wary eye on the data saturation, which is under explored in big data analytics. Data saturation is generally used to refer to the process of gathering and analyzing data to the point at which no new insights are added (Wray et al., 2007). In analyzing waste haulers' transportation behavior, one day's data was randomly selected and plotted (Figure 4). When the analyses were gradually extended to more days, it is noticed that the patterns are largely stable without new insights added (see Figure 6), i.e. it has reached a point of data saturation. With really big data the computational power, energy and time savings can be very high. Leek (2014) suggested that in big data analytics, it is best to define a metric for success up front and stop wasting resources when the data is saturated. It is an analogous issue to the question of parsimonious sample size in small-data research. For example, a political polling researcher will rarely sample more than 2000 voters under normal expectations of the distribution of votes, since the reduction of the standard error of the estimate beyond that number is tiny compared with the cost of surveying additional people. Data saturation challenges the orthodox view that the bigger the data, the better.

3.5 "The gold mine" to be protected or to be shared?

Currently, many big data sets are left over unintentionally when businesses are done (Ekbia et al., 2015). For example, they are created as by-products of people travelling around, communicating using smart phones, or purchasing from supermarket or through e-commerce. Likewise, the Hong Kong CWM big data set is a by-product of measuring and monitoring CWM flows. The amassed data can be a corporate asset, the mining of which allows companies to make better business predictions and decisions. Big data is thus like a gold mine; researchers and data analysts gather around potentially rich sources like a 'gold rush'. Since it is incidentally created and describes natural business processes and captures revealed behavior, big data tends to be considered better than experimental data or simulation data as it potentially contains more ground truth with respect to social reality than traditional instruments (Hand, 2015). Big data portrays a fuller picture of a subject matter, which allows for a stronger claim to objective truth; as Anderson (2008) put it, "with enough data, the numbers speak for themselves". Researchers and data analysts are therefore abandoning carefully curated small data and are rushing to discover big data sources to exploit. It can therefore be predicted that data owners will become more protective of their big data and reluctant to share their gold mine with others. Facebook accumulates big data from its users but only a few individuals have free access to it. Some companies restrict access to their data entirely, others sell access for a fee, and others offer small datasets to university-based researchers (Boyd and Crawford, 2012). The big data on CWM in Hong Kong was granted to the research team for free as the request was made at a time when big data was not as highly sought after as it is today.

The open data movement around the world may offset the effects of this trend to a certain extent by calling for big data to be openly available. Open data is the idea that some data should be freely accessible to everyone to use and republish as they wish, without restrictions from copyright, patents, licenses or other mechanisms of control (Auer et al., 2007) exerted by both public and private organizations. The movement argues that these restrictions are at odds with the communal good and hence data should be made available without restriction. For some public organizations, such as the United Nations, the World Bank, statistics bureaus, or government agencies, it is their obli-

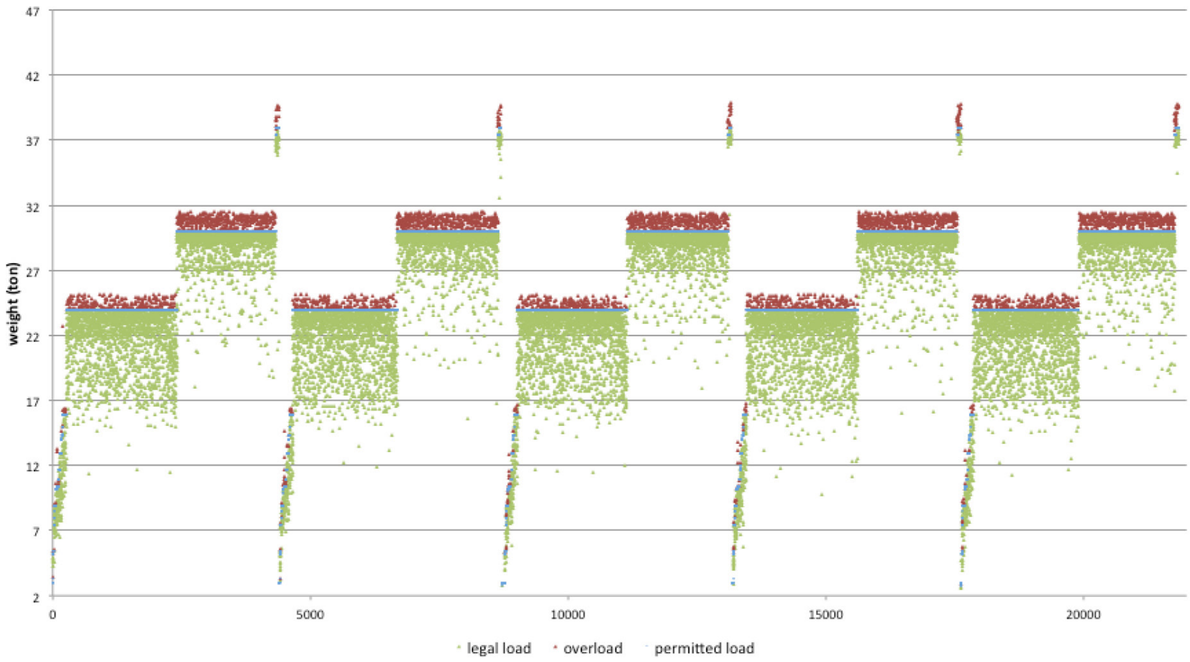

FIGURE 6: Pattern of overloaded or underloaded of waste haulers in five consecutive days (Sample size=20,841 track loads).

gation to make their data available to the public. These can all be sources for researchers to form big data. To enrich the Hong Kong CWM big data set, publicly available data from the HKBD and HKGBC were accessed and linked to the CWM process data. Enormous efforts went into data searching, slicing, stitching, and cleansing; a process similar to putting jigsaw pieces together. Manual interventions, string matching algorithms, and address geocoding were all employed to link the various data sets together.

3.6 Proactive big data strategies, "developing a mine to mine?"

Currently, many big data sets are largely left not intendedly, but discovered a 'gold mine' by mining which many meaningful, previously uncovered findings can be revealed (Terranova, 2000). Many institutions have adopted proactive strategies to develop big data. For example: statistical organizations such as Eurostat, United Nations Economic Commission for Europe, have formulated their big data roadmaps (Kitchin, 2015); Chicago has launched the 'Array of Things', a connected network of sensors that will be deployed throughout the city to collect data on environmental factors such as air quality, noise, and climate, which can then be used to discover hidden problems and develop targeted policies to improve city life (Thornton, 2015); and Barcelona deployed responsive technologies across urban systems including public transit, parking, street lighting, and waste management, which are intended to yield significant cost savings, improve the quality of life for residents, and provide better urban governance (Adler, 2016). The foregoing is essentially "developing a mine to mine", which is possible with increasing accessibility to ubiquitous and affordable sensing and communication technologies. An example relating to CWM in Hong Kong is probing into the behavior of waste haulers. The CWM big data captured for this study did not include the time when haulers left a site or their behavior en route to waste disposal facilities. Now, in addition to Radio Frequency Identification (RFID) technologies (Lu et al., 2011a; Flanagan et al., 2014), smarter technologies have been developed and embedded in the lorries to track their geographical positions (Niu et al., 2016; 2017). Such proactive big data strategies are similar to finding the missing pieces of a jigsaw puzzle.

Adopting proactive big data strategies raises many data platform design issues. For example, where to install the data capturing and communication devices such as sensing, RFID, laser scanning, webcam, and wireless. CWM researchers have suggested that new technologies should be unintrusive to ongoing construction processes otherwise their use is doomed to failure (Niu et al., 2017). Although data capturing and communication technologies are more accessible than ever, they still incur significant costs that need to be optimized against the value of data collected. In addition to the capital cost of installation, there is the ongoing cost of maintaining and renewing the data infrastructure. The lifecycle cost can be formidable, particularly where the system monitors performance over a large city. Furthermore, proactive data collection relies on a network of devices where the malfunction of one device could potentially cause the whole system to fail. Data infrastructures therefore need to be designed with a high degree of resilience.

3.7 Last but not the least: a little touch of big data ethics

The capture and use of big data has both benefits and risks. Ever since the advent of big data, there has been concern over the ethical ramifications of data analysts misusing its power, e.g. Facebook's data privacy scandal in March 2018. Although the conceptual, regulatory, and institutional resources of research ethics have developed greatly over the past few decades and have become familiar to researchers, there are always many unaddressed issues with respect to the ethical implications of the big data phenomenon (Boyd and Crawford, 2012). Existing norms governing data and research ethics have difficulties accommodating the special features of big data. The ethics of using big data are intimately tied to questions of ownership, access and intention, all of which are often disputed. Social media such as Facebook claim to own their big data and have exclusive access to it, even though the data itself is actually contributed by their users. It is also problematic for researchers to justify their actions as ethical simply because the data are accessible, let alone respond to the accusation that "limited access to big data creates new digital divides" (Boyd and Crawford, 2012).

Consent, in particular informed consent, premised on the liberal tenets of individual autonomy, freedom of choice and rationality, has been the cornerstone of personal data regulation and ethics (Cheung, 2016). However, it becomes impossible to ask researchers to obtain consent from every waste hauler who left the data passively as a part of the business process. Traditional de-identification approaches (e.g. anonymization, pseudonymization, encryption, or data sharding) to protect privacy and confidentiality and allow analysis to proceed are now problematic in big data, as it is the power of big data analytics that even anonymized data can be re-identified and attributed to specific individuals (Ohm, 2009). For example, analyzing the CWM big data can tell which companies performed well in CWM and which did not. De-identification is not always helpful, one can re-identify the companies which left their records in other databases, e.g. the one in the HKBD. Researchers thus need to start thinking more clearly about accountability of big data analytics; identifying methods, predictions and inferences that can be considered ethical, and those that are not.

The big data revolution has seen its ramifications including a series of ethical issues as listed above, none of which is resolvable with an easy answer. Metcalf et al. (2016) suggested that to get a grasp of the ethics of big data requires theorizing big data as something more than a technological artifact. The law is a powerful element in big data ethics, but it is far from able to handle the many nuanced scenarios that arise; organizational principles, institutional statements of ethics, self-policing, and other forms of ethical guidance are also needed (King and Richards, 2014).

4. CONCLUSIONS

Big data has rapidly become a game changer in many research realms, including waste management. Using the big data in construction waste management (CWM) in Hong Kong as an inductive case study, this study provides a synoptic overview of the prospects and challenges of big data in CWM. It is argued that big data, in comparison with small data collected from sampling and ethnographic methods, can portray a fuller picture, so that research findings from the big data can be accepted with a higher level of confidence. It is also illustrated that big data analytics can reveal hidden patterns, unknown correlations and other useful information to better inform CWM decisions.

Given the advantages of big data and the increasing availability of routinely-collected data, it is likely that big data will be a standard requirement for CWM research in the near future. However, it is also expected that data owners will become more protective of their big data for reasons of profit, privacy, or security. Some of the main issues using publicly available data have been reviewed, which will probably have to be the source for CWM research in the future. Hong Kong's CWM data, similar to its counterparts in other areas such as social media, e-commerce, or retailing, is left over unintentionally. Given the value of big data, it is expected that many researchers will take proactive strategies to collect big data. This is particularly opportune nowadays as data acquisition and communication technologies are becoming increasingly accessible. The findings provide references for big data in CWM in other regions and countries through specifying the prospects and challenges regarding data collection, analysis and applications.

There are misconceptions that prevail in big data research, one of which is in relation to the definitive size over which a dataset can be called big data. This paper argues that there is no definitive size and that the criteria should be whether the data is able to account for the totality of a relevant subject and whether it allows values to be created. It is suggested that academic arguments and positions in respect what big data is and is not are less important than understanding what large data sets can and cannot do. While many researchers are eager to explore the value of big data, both data mining and traditional applied statistics face challenges in dealing with its volume, velocity, and variety, and there are some researchers who do not consider big data research as representing scientific inquiry. Also, bigger data does not necessarily mean better data and researchers are advised to have a comprehensive and impartial understanding of the phenomena before embarking upon research that involves it.

Although this study has provided many interesting insights, they are just the tip of an iceberg. There is a massive agenda of big data for CWM researchers. Further research along these various lines will drive a shift from a theory-driven to data-driven regime investigations and from searching for correlation and causality to correction. The finer-grained the sensing technology and procedures underlying the database, the nearer to real-time will be the correctional options. As CWM systems are better modeled and understood, researchers will be able to move from basic descriptive analysis to behavioral analysis of CWM. It is also recommended that future studies should take an in-depth look into proactive big data strategies and big data ethics, neither of which has been fully deliberated in this paper.

ACKNOWLEDGEMENTS

The authors are grateful to Dr Elisabete Silva, Senior Lecturer at Department of Land Economy, Cambridge University, to include us in the Cambridge Big Data initiative and join the intellectual debate. Appreciations also go to Hong Kong Environmental Protection Department for granting the big data and other endless supports.

REFERENCES

Adler, L. (2016). How smart city Barcelona brought the internet of things to life. https://datasmart.ash.harvard.edu/news/article/how-smart-city-barcelona-brought-the-internet-of-things-to-life-789 (accessed on 17 December 2017).

Agrawal, R., Grosky, W., and Fotouhi, F. (2006). Image retrieval using multimodal keywords. In Proceedings of the Eighth IEEE International Symposium on Multimedia, 817-822.

Akter, S., and Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. Electronic Markets, 26(2), 173-194.

Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. https://www.wired.com/2008/06/pb-theory/ (accessed on 17 December 2016).

Auer, B., Christian, S., Georgi, K., Jens, L., Richard, C., and Zachary, I. (2007). Dbpedia: A nucleus for a web of open data. In The Semantic Web, Springer Berlin Heidelberg.

Bilal, M., Oyedele, L. O., Akinade, O. O., Ajayi, S. O., Alaka, H. A., Owolabi, H. A. (2016a). Big data architecture for construction waste analytics (CWA): A conceptual framework. Journal of Building Engineering, 6, 144-156.

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., and Pasha, M. (2016b). Big Data in the construction industry: A review of present status, opportunities, and future trends. Advanced Engineering Informatics, 30(3), 500-521.

Bossink, B. A. G., and Brouwers, H. J. H. (1996). Construction waste: Quantification and source evaluation. Journal of Construction Engineering and Management, 122(1), 55-60.

Boyd, D., and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662-679.

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

Cheung, A. (2016). Making sense and non-sense of consent in the big data era. In Symposium on Big Data and Data Governance.

Clifton, C. (2010). Encyclopædia britannica: Definition of data mining. https://www.britannica.com/EBchecked/topic/1056150/data-mining (accessed on 17 December 2017).

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., and Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. Journal of the Association for Information Science and Technology, 66(8), 1523-1545.

Fatta, D., Papadopoulos, A., Avramikos, E., Sgourou, E., Moustakas, K., and Kourmoussis, F. (2003). Generation and management of construction and demolition waste in Greece—an existing challenge. Resources, Conservation and Recycling, 40(1), 81-91.

Flanagan, R., Jewell, C, Lu, W., and Pekericli, K. (2014). Auto-ID – Bridging the physical and the digital on construction projects. Chartered Institute of Building. ISBN 1853800191.

Formoso, T. C., Soibelman, M. L., Cesare, C. D., and Isatto, E. L. (2002). Material waste in building industry: Main causes and prevention. Journal of Construction Engineering and Management, 128(4), 316-325.

Goodman, S. N. (1999). Toward evidence-based medical statistics: The P value fallacy. Annals of Internal Medicine, 130(12), 995-1004.

Han, J., Kamber, M., and Pei, J. (2012). Data Mining: Concepts and Techniques. Elsevier.

Hand, D. J. (2015). Official statistics in the new data ecosystem. In the New Techniques and Technologies in Statistics Conference.

HKEPD (2014). Construction waste disposal charging scheme. https://www.epd.gov.hk/epd/misc/cdm/scheme.htm (accessed on 17 December 2016).

Katz, A., and Baum, H. (2011). A novel methodology to estimate the evolution of construction waste in construction sites. Waste Management, 31(2), 353-358.

Kazaz, A., Ulubeyli, S., and Arslan, A. (2018). Quantification of fresh ready-mix concrete waste: order and truck-mixer based planning coefficients. International Journal of Construction Management, 1-12.

King, J. H., and Richards, N. M. (2014). What's up with big data ethics? https://www.forbes.com/sites/oreillymedia/2014/03/28/whats-up-with-big-data-ethics/#4e94d3703591 (accessed on 17 December 2017).

Kitchin, R. (2015). Big data and official statistics: Opportunities, challenges and risks. The Programmable City Working Paper 9.

Kitchin, R., and Lauriault, T. (2015). Small data in the era of big data. GeoJournal, 80, 463-475.

Kostelnick, C. (2007). The visual rhetoric of data displays: The conundrum of clarity. IEEE Transactions on Professional Communication, 50(4), 280-294.

Leek, J. (2014). 10 things statistics taught us about big data analysis. Simplystats blog, May 22. https://simplystatistics.org/2014/05/22/10-things-statistics-taught-us-about-big-data-analysis/ (accessed on 17 December 2016).

Lin, M., Lucas Jr, H. C., and Shmueli, G. (2013). Research commentary-too big to fail: Large samples and the p-value problem. Information Systems Research, 24(4), 906-917.

Lu, W., Chen, X., Ho, D. C. W., and Wang, H. (2016a). 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., and Shen, L. (2015). Benchmarking construction waste management performance using big data. Resources, Conservation and Recycling, 105, 49-58.

Lu, W., Huang, G. Q., and Li, H. (2011a). Scenarios for applying RFID technology in construction project management. Automation in Construction, 20, 101-106.

Lu, W., and Tam, V. W. (2013). Construction waste management policies and their effectiveness in Hong Kong: A longitudinal review. Renewable and Sustainable Energy Reviews, 23, 214-223.

Lu, W., Peng, Y., Chen, X., Skitmore, M., and Zhang, X. (2016b). The s-curve for forecasting waste generation in construction projects. Waste Management, 56, 23-34.

Lu, W., Webster, C., Peng, Y., Chen, X., and Zhang, X. (2017). Estimating and calibrating the amount of building-related construction and demolition waste in urban China. International Journal of Construction Management, 17(1), 1-12.

Lu, W., Yuan, H., Li, J., Hao, J. J., Mi, X., and Ding, Z. (2011b). An empirical investigation of construction and demolition waste generation rates in Shenzhen city, South China. Waste Management, 31(4), 680-687.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., and Barton, D. (2012). Big data: The management revolution. Harvard Business Review, 90(10), 61-67.

McGregor, M., Washburn, H., and Palermini, D. (1993). Characterization of construction site waste. Final report presented to the METRO Solid Waste Department, Portland, Oregon.

Metcalf, J., Emily F. K., and Danah, B. (2017). Perspectives on big data, ethics, and society. Council for Big Data, Ethics, and Society. https://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/ (accessed on 17 December 2017).

MIT Technology Review (2013). The big data conundrum: How to define it? https://goo.gl/nQhGWP (accessed on 17 December 2016).

Moses, L. E. (1986). Think and explain with statistics. Addison-Wesley

Niu, Y., Lu, W., Chen, K., Huang, G. Q., and Anumba, C. (2016). Smart construction objects. Journal of Computing in Civil Engineering, 30(4), 04015070.

Niu, Y., Lu, W., Liu, D., Chen, K., Anumba, C., and Huang, G. Q. (2017). An SCO-enabled logistics and supply chain management system in construction. Journal of Construction Engineering and Management, 143(3), 04016103.

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA Law Review, 57, 1701.

Padhy, R. P. (2013). Big data processing with Hadoop-Map reduce in cloud systems. International Journal of Could Computing and Services Science, 2(1), 16-27.

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

Poon, C. S., Yu, T. W., and Ng, L. H. (2001). A guide for managing and minimizing building and demolition waste. Hong Kong Polytechnic University, Hong Kong.

Press, G. (2013). What's the big data? https://whatsthebigdata.com (accessed on 17 December 2017).

Russom, P. (2011). Big data analytics. TDWI Best Practices Report, Fourth Quarter.

Schönberger, V. M., and Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. John Murray: London.

Sen, P. K., and Singer, M. J. (1993). Large sample method in statistics. Chapman & Hall, New York, United States.

Senaratne, S., and Rasagopalasingam, V. (2017). The causes and effects of work stress in construction project managers: the case in Sri Lanka. International Journal of Construction Management, 17(1), 65-75.

Shelton, T. (2017). The urban geographical imagination in the age of Big Data. Big Data & Society, 4(1), 2053951716665129.

Shen, Y., Li, Y., Wu, L., Liu, S., and Wen, Q. (2016). Big data overview. In IRMA (ed.) Big Data: Concepts, Methodologies, Tools, and Applications. IGI Global.

Shen, L., Lu, W., Peng, Y., and Jiang, S. (2011). Critical Assessment indicators for measuring benefits of rural infrastructure investment in China. Journal of Infrastructure Systems, 17(4), 176-183.

Sivarajah, U., Kamai, M. M., Irani, Z., and Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70(1), 263-286.

Skoyles, E. R. (1976). Materials wastage – a misuse of resources. Building Research and Practice, 232-243.

Soibelman, L. (2016). Big data and its Impact in the Architecture, Engineering, and Construction Industry. A keynote speech presented on the International Conference on Advancement of Construction Management and Real Estate.

Taylor, L., and Schroeder, R. (2015). Is bigger better? The emergence of big data as a tool for international development policy. GeoJournal, 80(4), 503-518.

Terranova, T. (2000). Free labor: Producing culture for the digital economy. Social Text, 18(2), 33-58.

Thornton, S. (2015). The internet of things in Chicago: Collaborative action for smarter cities. https://datasmart.ash.harvard.edu/news/article/the-internet-of-things-in-chicago-collaborative-action-for-smarter-cities-6 (accessed on 17 December 2017).

Treloar, G. J., Gupta, H., Love, P. E. D., and Nguyen, B. (2003). An analysis of factors influencing waste minimization and use of recycled materials for the construction of residential buildings. Management of Environmental Quality, 14(1), 134-145.

Wasserstein, R. L., and Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. The American Statistician, 70(2), 129-133.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., and Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t Tests. Perspectives on Psychological Science, 6(3), 291-298.

World Economic Forum (2012). Big data, big impact: New possibilities for international development. WEF.

Wray, N., Markovic, M., and Manderson, L. (2007). Researcher saturation: the impact of data triangulation and intensive-research practices on the researcher and qualitative research process. Qualitative Health Research, 17(10), 1392-1402.

Yin, R. K. (1989). Case study research: Design and methods. Newbury Park, CA: Sage Publications.

Zaslavsky, A., Perera, C., and Georgakopoulos, D. (2013). Sensing as a service and big data. https://arxiv.org/ftp/arxiv/papers/1301/1301.0159.pdf (accessed on 17 December 2017).

Zhou, K., Fu, C., and Yang, S. (2016). Big data driven smart energy management: From big data to big insights. Renewable and Sustainable Energy Reviews, 56, 215-225.