Author's manuscript

Big data analytics to identify illegal construction waste dumping: A Hong Kong study

Weisheng Lu[1]

Abstract

Illegal dumping, referring to the intentional and criminal abandonment of waste in unauthorized areas, has long plagued governments and environmental agencies worldwide. Despite the tremendous resources spent to combat it, the surreptitious nature of illegal dumping indicates the extreme difficulty in its identification. In 2006, the Construction Waste Disposal Charging Scheme (CWDCS) was implemented, regulating that all construction waste must be disposed of at government waste facilities if not otherwise properly reused or recycled. While the CWDCS has significantly improved construction waste management in Hong Kong, it has also triggered illegal dumping problems. Inspired by the success of big data in combating urban crime, this paper aims to identify illegal dumping cases by mining a publicly available data set containing more than 9 million waste disposal records from 2011 to 2017. Using behavioral indicators and up-to-date big data analytics, possible drivers for illegal dumping (e.g., long queuing times) were identified. The analytical results also produced a list of 546 waste hauling trucks suspected of involvement in illegal dumping. This paper contributes to the understanding of illegal dumping behavior and joins the global research community in exploring the value of big data, particularly for combating urban crime. It also presents a three-step big data-enabled urban crime identification methodology comprising 'Behavior characterization', 'Big data analytical model development', and 'Model training, calibration, and evaluation'.

[1] Associate Professor, Corresponding author, Department of Real Estate and Construction, The University of Hong Kong, Hong Kong, Email: wilsonlu@hku.hk, Telephone: +852 3917 7981;

Keywords: Construction waste management; Illegal dumping; Criminal behavior analysis; Big data analytics; Hong Kong

1. Introduction

Illegal dumping, sometimes called fly-tipping, refers to the intentional and illegal abandonment of waste in unauthorized public or private areas, usually to avoid tipping fees and save on transport time and cost, or simply for the sake of convenience (Webb et al., 2006). It is generally treated as a criminal offence across jurisdictions. The UK Department for Environment, Food & Rural Affairs (Defra), for example, deals with illegal disposal of waste under Section 33 of the Environmental Protection Act 1990. Defra (2017) reported that local authorities in England dealt with 936 thousand fly-tipping incidents in 2015/16, a 4.0% increase over 2014/15. In the U.S., dumping waste in unauthorized areas is illegal under the federally enforceable Protection of the Environment Operations Act 1997 (USEPA, 1998). Illegal dumping has become a global issue and is frequently reported in Australia (Meldrum-Hanna et al., 2017), Italy (Massari and Monzini, 2004), Spain (Sáez et al., 2014), Israel (Seror et al., 2014), Mainland China (Jin et al., 2017), and Hong Kong (Audit Commission, 2016), and is a particular problem in countries with rapid gross domestic product (GDP) growth (Nunes et al., 2009).

Illegal dumping is not only a nuisance in its own right but can also lead to many other problems (Esa et al., 2017). It is a human health concern and can damage the environment in a variety of ways (Romeo et al., 2003). Fly-tipped waste causes habitat destruction, wildlife deaths (Webb et al., 2006), and is a major source of soil and underground water pollution (Shenkar et al., 2011). It also causes aesthetic damage to the natural landscape. When illegal waste dumping is discovered, local governments often dispatch an abatement crew to clean it up as quickly as possible because the contained oil, solvents, fuel, rusted metal, and batteries can cause severe environmental damage. Such clean-up comes at great expense. According to Defra (2017), local authorities in England spent around £49.8 million cleaning up fly-tipped waste in 2015/16 alone. Romeo et al. (2003) report that the City of San Antonio in the U.S. spends hundreds of millions of dollars annually mitigating the environmental consequences of illegal waste dumping. In Hong Kong, Lin (2016) reported that around one hectare of wetland and mangrove forest had been affected by illegal dumping committed by two individuals, with a repair cost estimated by the Environment Protection Department (EPD) at HK$6 million.

Governments and environmental agencies have committed extensive resources to combat illegal dumping (Gálvez-Martos et al., 2018). For example, to overcome patchy data collection in order to better understand the scale of the problem, the UK government launched Flycapture® in 2004 (later replaced by WasteDataFlow®), requiring all local authorities and the Environment Agency to submit monthly returns on the number, size, waste types and location types of fly-tips (Webb et al., 2006). Israel has explored vehicle impoundment policy and evaluated its effect on illegal dumping of construction waste (Seror et al., 2014). In Hong Kong, a fly-tipping spotting system (similar to Flycapture®) has been implemented to encourage public reporting of illegal dumping activities. Researchers have also explored various policy and technological recommendations for addressing illegal dumping problems. Examples include enhancing prosecution and enforcement (Yuan et al., 2011), increasing surveillance and ambush (Navarro et al., 2016), adopting new construction method (Li et al., 2014), and using Global Positioning System and satellite images to catch illegal dumping activities (Persechino et al., 2010). However, the effectiveness of these approaches is questionable. Illegal dumping activities are committed stealthily and are thus difficult to catch (Scherer, 1995).

Big data is increasingly advocated as a powerful instrument for detection and deterrence of contemporary urban issues such as crime, corruption, and fraud. Reports published by the World Economic Forum (WEF) (2015), Transparency International (2017), Ernst & Young (2014), and Unisys (2012) advocate for the power of big data and analytics in reducing corruption and fraud. Since urban crimes are generally conducted in a stealthy way, evidence of them may be deeply buried in a dataset if captured at all. The problem of identifying such activities is extremely difficult to crack. However, offenders may have left unintentional clues or exhibited hidden patterns, identifable when the dataset is sufficiently large and with the use of proper analytics. Williams et al. (2017) reviewed studies making use of 'naturally occurring' socially relevant data (e.g., on Twitter or Facebook) to complement and augment conventional curated data to address the classic problem of crime pattern estimation. By combing through datasets on government bidding processes, contracting firms' financial disclosures, the beneficial ownership of contracting firms, public officials' tax and family records, and complaints to authorities about bribery by competing contractors, Fazekas et al. (2013) tried to uncover patterns of fraud and bribery in public procurement. There have been several stories on the success of big data, based on which an exploration of how big data analytics can be employed to identify illegal dumping as a contemporary urban issue promises to be intriguing as well as meaningful.

The primary aim of this research is to develop a big data-driven methodological approach that can be used to identify suspected cases of illegal dumping. It is contextualized in Hong Kong, which has long been suffering from the problems caused by illegal dumping, and focuses on construction waste, which constitutes a prodigious proportion of total municipal solid waste. The rest of this paper is structured as follows. Subsequent to this introductory section is a literature review covering big data and analytics for urban crime identification. The big data of illegal dumping in Hong Kong is introduced in the Section 3. The research methods are described in Section 4. These methods are devised to achieve three specific research objectives: (1) To develop a set of indicators for suspected dumping activities using mixed methods research; (2) To develop an analytical model by applying these indicators and big data analytics; and (3) To train, calibrate, and evaluate the analytical model by trying out different data analytics. Section 5 reports the data analyses and findings and Section 6 is an in-depth discussion including both methodological contributions and policy implications of this research. Conclusions are drawn in the final section.

2. Big data analytics to tackle contemporary urban issues

According to Padhy (2013), big data can be characterized as a collection of datasets so large and complex that it is difficult to process using traditional data management tools. Mayer-Schönberger and Cukier (2011) describe big data techniques as ‘things one can do at a large scale that cannot be done at a smaller one, to create a new form of value’. Many researchers accepted Gartner’s three defining characteriztics of big data, namely, volume, variety and velocity, or the ‘three Vs’ (McAfee et al., 2012). *Volume* is the quantity of data in the form of records, transactions, tables or files; *velocity* can be expressed in batches, near time, real time and streams; and *variety* can be structured, unstructured, semi-structured or a combination thereof (Chen et al., 2014). Big data analytics can uncover hidden patterns, unknown correlations, and other useful information to guide business predictions and decision-making (Shen et al., 2016); in effect, *value* is advocated as the fourth ‘V’. By analyzing big data, ‘latent knowledge’ (Agrawal et al., 2006) or ‘actionable information’ (WEF, 2012) can be identified.

Big data success stories abound in a wide range of areas, including science, business, public governance, innovation, competition, and productivity (Sagiroglu and Sinanc, 2013). It is also increasingly being advocated as an effective means of tackling contemporary urban issues such as terrorism, crime, corruption, fraud, and financial non-compliance. Access to big data is a prerequisite for combating urban crime. As Vona (2017) suggests, ‘even the world’s best auditor using the world’s best audit program cannot detect fraud unless their sample includes a fraudulent transaction’. Baesens et al. (2015) estimate that fewer than 0.5% of credit card transactions are typically fraudulent. The problem of identifying fraudulent activities is thus commonly referred to as a needle-in-a-haystack problem. However, when the dataset is sufficiently large, clues unintentionally left or hidden patterns exhibited by offenders become identifable.

Another prerequisite for combating urban crime is proper data analytics. Pramanik et al. (2017) reviewed five big data techniques that can be used to extract hidden network structures among criminals: link analysis, intelligent agents, text mining, neural networks, and machine learning. Clearly, neither urban crime problems nor analytical methods are new. It is the expontential growth of data in the digital era that provides both new opportunities and challenges. Fazekas et al. (2013) and Fazekas and Tóth (2014) describe a methodology for identifying corruption in public procurement. They first collected a massive amount of data relating to public procurement. In parallel, they identified a series of indicators that could predict suspected corruption cases (e.g., ‘exceptionally short bidding periods’ or ‘bids repeatedly won by the same company’) and incorporated them into a corruption risk index model. Finally, using inferential statistical analysis, they identified corrupt behavior based on deviations from ordinary patterns.

A review of previous studies seems to suggest that there is no one-size-fits-all big data-enabled solution to urban criminal issues. A good starting point, however, is to characterize the criminal activities in question, e.g., illegal dumping, and then identify anomalous behavior and 'red flags'. In a big data-driven methodology comprising 'Behavior characterization', 'Big data analytical model development' and 'Model calibration', these three steps in combination can indicate, at the very least, highly suspected activities. In the context of public procurement, for example, Fazekas and Tóth (2014) characterized the behavior by proposing more than 30 indicators of high corruption risk. Based on the characteristics and the indicators, the next step is to develop the big data analytical model. Data analytical methods ranging from 'simple' regression analysis to complex techniques such as support vector machines, artificial neural networks, association rules, case-based reasoning, and *K*-means clustering are widely applied in urban crime detection (Fawcett and Provost, 1997). Finally, the big data analytical model needs to be trained, calibrated, and evaluated using known cases, e.g., crime convictions, before it can be applied to the big data set to identify other suspected cases and for further follow-up actions.

3. The big data of illegal dumping in Hong Kong

In Hong Kong, the adverse environmental impacts of construction waste resulting from creation of its impressive built environment are a grave concern. As in other states and territories, construction waste in Hong Kong is classified into inert and non-inert components. EPD (2017) statistics show that the total solid waste deposited in Hong Kong landfills in 2015 amounted to 15,102 tons per day (tpd), of which 4,200 tpd, or 27.8%, was from construction activities. Thus, construction generates around one-quarter of the total solid waste finding its way into landfills. Owing to its significant adverse impacts, construction waste is heavily regulated in Hong Kong, and a series of statutory and non-statutory policies, including regulations, codes, and schemes have been introduced over the past few decades (Lu and Tam, 2013). In particular, the Construction Waste Disposal Charging Scheme (CWDCS), which mandates that all construction waste, if not otherwise reused or recycled, must be disposed of at government waste facilities (e.g., landfills, offsite sorting facilities [OSFs] or public fill banks) was implemented in 2006. According to this scheme, the main contractor is charged HK$200 for every ton of non-inert waste it dumps in landfills; HK$175 per ton for mixed inert and non-inert waste accepted by OSFs; and HK$71 per ton for inert waste accepted by public fills (raised from HK$125, HK$100, and HK$27 respectively in April 2017). As a policy system, the CWDCS together with its enforcement measures has been praised for its efficiency in construction waste minimization (Lu et al., 2015).

At the same time, the CWDCS has incentivized illegal dumping. Illegal disposal of one load of construction waste immediately saves contractors between HK$405 to HK$3,750 in tipping fees, depending on the volume and type of waste. This does not include savings in transport costs (normally HK$800-1,500 per trip) and waiting time at government facilities. In response to a Legislative Council (LegCo) query, the Environment, Transport and Works Bureau (2006) reported that 508 complaints of construction waste illegal dumping were received between 20 January 2006 (the CWDCS implementation date) and 31 May 2006, a significant rise from the 101 received in the same period in 2005. After that, fly-tipping reports have continuously become epidemic. Hong Kong’s Audit Commission (2016) recently found that public reports of illegally dumped construction materials increased a phenomenal 328% in 2015, rising from 1,517 to 6,499. In that year, 6,300 tons of illegally dumped construction materials were cleared by government departments. Without quick abatement, such waste can cause severe environmental damage. For example, environmentalists have warned that wetland fauna and mangroves are particularly vulnerable to illegal dumping (Lau, 2016).

The structure of the big data is illustrated in Fig. 1, which comprises:

- the EPD Facility database containing all government construction waste management (CWM) facilities, including landfills, OSFs and public fills (See Fig. 1_1)
- the EPD Project database containing all projects that have dumped waste in the above facilities. A total of 27,536 construction projects, along with information on site address, client, project type and other details, are recorded (see Fig. 1_2).
- the EPD Waste Disposal database (see Fig. 1_3), which records every truckload of construction waste received at CWM facilities. A total of 9,338,243 disposal records were generated from all construction projects carried out during the eight-year period from 2011 to 2017, with around 3,500 records being added every day. The unique account number links projects and waste disposal records.
- the EPD Vehicle database containing 9,863 vehicles involved in construction waste transport (see Fig. 1_4), which can be linked to data from the Transport Department.

According to the three Vs (i.e., volume, velocity, and variety), this CWM dataset qualifies as big data. By mining it, it is anticipated that cases of illegal dumping can be identified. It can also facilitate understanding of the magnitude of the problem in order to develop countermeasures.

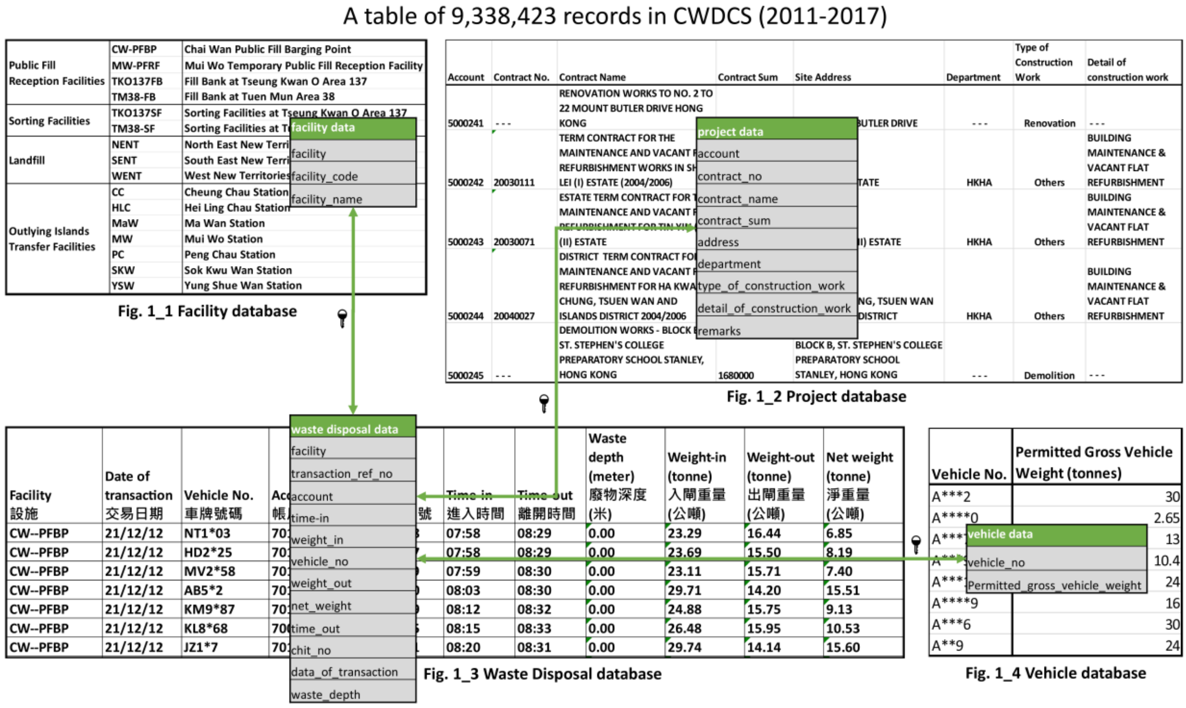


Fig. 1 The big data structure and example records

4. Methods

Following the three steps of behavior characterization, big data analytical model development, and model calibration, this research develops a big data-driven methodology for illegal dumping identification. Firstly, a set of red-flag indicators for predicting illegal dumping activities are developed. Next, an analytical model is developed by applying the indicators and searching for proper data analytics. Finally, the model is trained, calibrated, and evaluated before application to the big data set to generate high-confidence identification of illegal dumping cases.

4.1 Developing a set of red-flag indicators

To develop red-flag indicators, illegal dumping behavior is characterized by adopting a mixed method approach. Since 2013, the research team has conducted a series of research projects with construction clients (both public and private), main contractors, government departments (e.g., the EPD and Construction Industry Council), LegCo members, waste haulers, unions, environmentalists, and other informants to try to understand the motivations for illegal dumping and offenders' behavior.

Waste haulers are the focal point as they are direct illegal dumping offenders. Their vehicles must be registered with the EPD (i.e., in the EPD Vehicle database) before they can provide construction waste hauling services. Haulers charge a flat per-trip rate regardless of what they are transporting. While it would seem that they have no incentive to commit illegal dumping, which benefits only their clients via tipping-fee savings, waste haulers may be more likely to do so if they are associated with a main contractor rather than operating as freelancers. Distance from construction site to landfill site also matters. A longer distance means higher transport costs which could induce illegal dumping. A list of indicators for predicting illegal dumping activities is presented in Table 1. It must be pointed out that this list is very tentative: it is unknown whether some of the indicators are useful and whether there is available data for them. In addition, it is not an exhaustive list. There may be other indicators that have not been identified, including those that could be discovered by big data analytics.

Table 1. List of indicators for predicting illegal dumping activities

ID	**Name**	**Unit**	**Source and calculation**
I_1	Time spent in a facility	Minute	Difference between 'departure time' and 'entering time'
I_2	Dumping weight	Ton	Difference between 'departure weight' and 'entering weight'
I_3	Rest/absent days between two working periods	Day	The number of absent days from last dumping record
I_4	The number of clients served per day	1	The counts of project accounts/clients associated with the same hauler per day
I_5	Loading ratio	%	Dumping weight/maximum capacity
I_6	Dumping depth	m	Excessive depth of waste defined in the "waste disposal" database
I_7	Dumping weight by facility type	Ton	Dumping weight according to type of facility (e.g., landfill, OSF, public fill)
I_8	Percentage of dumping weight by facility type	%	Percentage of dumping weight according to type of facility
I_9	Dumping count by facility type	1	Dumping transaction counts by different types of facility per day

4.2 Developing an analytical model

The second step is to develop the core algorithms, which are encapsulated and figuratively referred to as the Illegal Dumping Filter (IDF) in this study (see Fig. 2). In developing the IDF, a well-structured data table containing all indicators and their computed values from the big data is created. However, it is unclear how the indicators will interact with one another (e.g., linearly or as a network). It would constitute too much arbitrariness if weights were attached to them by the researchers or even the informants, so this is conducted using data analytics: a general term referring to the process of automatically or semi-automatically examining datasets to discover the information (e.g., hidden patterns or anomalies) they contain (Witten et al., 2011). Data analysts have long used tools such as rule-based reasoning, pattern recognition, anomaly detection, social networks, and nodal analysis to detect financial non-compliance. Since there is no *prior* knowledge on which analytical methods will be most suitable for illegal dumping identification, one needs to try different models and examine their results. Here, a satisfactory result will be the IDF being able to identify offending waste haulers (i.e., by their plate numbers).

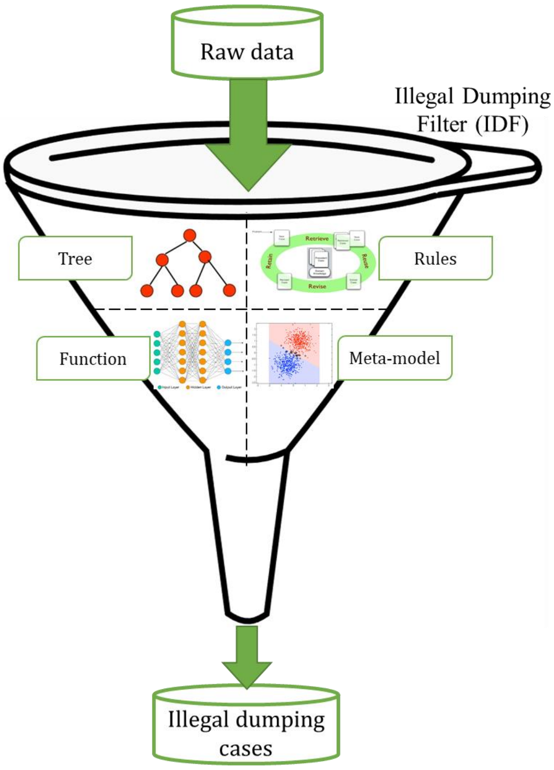


Fig. 2 An illustration of the Illegal Dumping Filter (IDF) in this study

4.3 Training, evaluating, and calibrating the big data analytical model

The third step is to train, evaluate, and calibrate the model before it can be applied to the big data set to identify illegal dumping cases. The sample, mainly comprising cases of illegal dumping convictions, will be used as the experimental/target group while a comparable sample will be used as the control group. It is critical that effectiveness of models is gaugeable. In math language, the effectiveness of the models can be gauged by precision rate, recall rate, and F_1-measure, which is the weighted average of Precision and Recall and is considered more accurate than if they are used individually. The research team needs to adjust the variable settings in the given software platform until a satisfactory result is reached.

5. Data analyses and findings

The IDF was first trained on a sampled data with a binary value, i.e., True and False, for the target label 'committed illegal dumping or not'. A target group included six trucks engaged in illegal dumping based on local news and video clips recorded by environmental activists. The control group was six non-offending trucks of a similar model and loading capacity. The two groups accounted for 36,678 dumping records between January 2011 and December 2017; very big data that might help identify hidden illegal dumping patterns or anomalies. The data of the two groups was selected into an independent table in MySQL (Version 5.7). Table 2 shows an excerpt of the training sample of yearly statistics of waste dumping behaviors based on Table 1. The first column in Table 2 indicates the target group ('True') or the control group ('False'). For the indicators I_1, I_2, I_3, I_4, I_5, and I_6, seven statistics, i.e., the minimum, 5% percentile, average, maximum, 95% percentile, sum, and standard deviation, were calculated using MySQL functions, e.g., *avg*() and *max*(), for each indicator. For the indicators I_7, I_8, and I_9, four-yearly statistics by facility types, i.e., the transaction counts for land fill, public fill, sorting, and islands, were computed. The final training sample of the IDF, as shown in Table 2, was a 'monster' data table consisting of 55 columns and 57 rows, with personal or privacy data anonymized in comma separated vector (CSV) format.

Table 2. An excerpt of the training sample of yearly statistics of waste dumping behaviors

Label	The 54 yearly statistics of behavioral indicators																	
Illegal	I_1 (7)							I_2 (7)	I_3 (7)	I_4 (7)	I_5 (7)	I_6 (7)	I_7 (4)				I_8 (4)	I_9 (4)
(1)*	I_1^{min}	$I_1^{5\%}$	I_1^{avg}	$I_1^{95\%}$	I_1^{max}	I_1^{Σ}	I_1^{σ}	…	…	…	…	…	I_7^{PF}	I_7^{LF}	I_7^{SF}	I_7^{OI}	…	…
True	4	4	7.42	13	28	3,250	3.08	…	…	…	…	…	5,319.63	26.62	0	0	…	…
True	3	5	7.85	13	51	2,111	3.94	…	…	…	…	…	3,369.65	0	0	0	…	…
True	4	5	8.84	15	60	6,328	4.75	…	…	…	…	…	11,429.38	0	0	0	…	…
True	2	4	12.40	37	57	5,494	9.91	…	…	…	…	…	7,291.96	0	0	0	…	…
True	3	4	7.48	17	45	6,489	4.98	…	…	…	…	…	13,795.46	128.95	0	0	…	…
False	2	4	13.04	32	82	5,348	9.78	…	…	…	…	…	6,527.19	0	0	0	…	…
False	2	4	14.29	31	107	20,875	9.32	…	…	…	…	…	22,844.93	9.92	0	0	…	…
False	2	3	10.09	24	54	12,062	7.29	…	…	…	…	…	18,697.39	56.4	31.28	0	…	…
*: The number of data columns is shown in parentheses																		

The next step was to identify the behavioral drivers of illegal dumping by trying linear models. A straightforward and easy-to-understand metric of the driving factors is Pearson's linear correlation coefficient. The correlations between the 54 indicators and the label in Table 2 were first tested, using IBM SPSS (version 24.0). Table 3 lists the eleven indicators showing statistical significance at the level 0.01 (two-tailed). The eleven indicators are statistics of three types of indicators, i.e., the duration in facility (I_1), the number of daily clients served (I_4), and waste depth (I_6). In other words, the three indicators were more related with the drivers of illegal dumping. Three statistics of I_4, i.e., I_4^{avg}, $I_4^{95\%}$, and I_4^{σ} had moderate negative correlations, while all the rest had weak negative correlations. To sum up, a truck with illegal dumping behaviors usually had fewer daily clients, less time spent at the government facilities, and less waste depth in the government's waste records.

Table 3. List of indicators correlated with illegal dumping (significant at the level 0.01)

	I_1^{avg}	$I_1^{95\%}$	I_1^{max}	I_1^{σ}	I_4^{avg}	$I_4^{95\%}$	I_4^{max}	I_4^{Σ}	I_4^{σ}	$I_6^{95\%}$	I_6^{σ}
Pearson's Correlation	-0.464	-0.417	-0.355	-0.417	-0.633	-0.550	-0.359	-0.407	-0.581	-0.346	-0.350
Significance (2-tailed)	0.000	0.001	0.007	0.001	0.000	0.000	0.006	0.002	0.000	0.008	0.008

The training data was further processed using Weka (version 3.9), which is an open source data mining software program (Frank et al., 2009). Data mining methods can discover nonlinear models of correlations, which approximates the illegal dumping behaviors better than the linear correlations in Table 3. Fig. 3 (a) shows a rule about illegal dumping concluded by JRip, which is a Java version of the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) method (Cohen, 1995). The rule in Fig. 3 (a) says a yearly record in Table 2 involves illegal dumping actions if and only if all three of the following conditions are met:

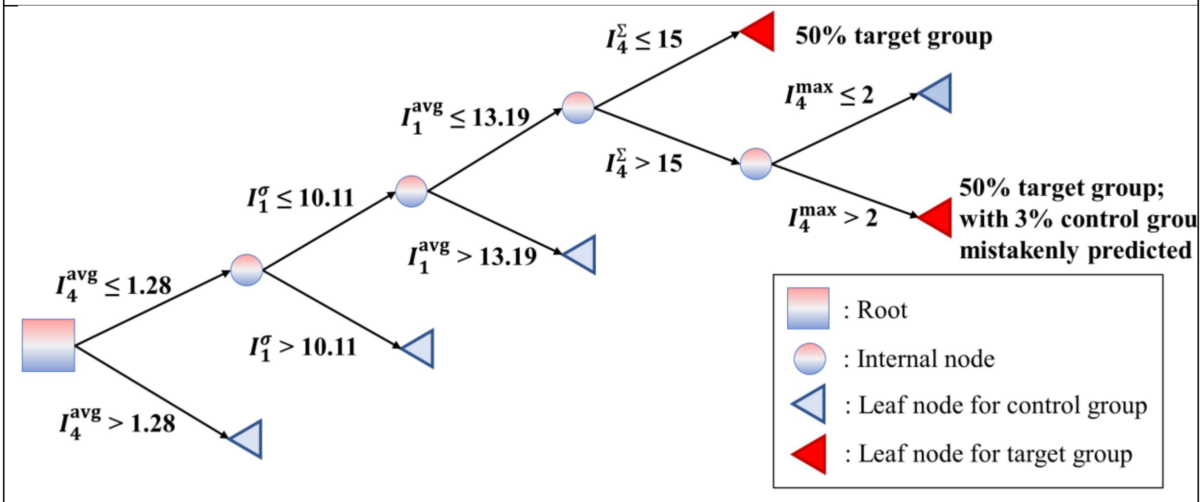
1. The average number of daily clients (I_4^{avg}) is no more than 1.28,
2. The average duration in facilities (I_1^{avg}) is no more than 12.75 minutes, and
3. The maximum duration in facilities (I_1^{max}) is no more than 165 minutes.

Fig. 3 (b) shows a decision tree concluded by another well-known data mining method, J48, a Java version of the C4.5 (see Quinlan, 1993). Decision trees reflect human decision-making and are easy to interpret (James et al., 2013). A decision process starts from the left-most square 'root' node, then follows the spitting paths ('burst' nodes) by matching conditions until a final decision on 'leaf' nodes is reached (Quinlan, 1986; Dey, 2002). In the decision tree, a yearly record involves illegal dumping actions if and only if all four of the following conditions are met:

1. The average number of daily clients (I_4^{avg}) is no more than 1.28 (the same as the first condition in Fig. 3 (a),
2. The standard deviation of the duration in facilities (I_1^{σ}) is no more than 10.11 minutes,
3. The average duration in facilities (I_1^{avg}) is no more than 13.19 minutes, and
4. The overall number of yearly clients (I_4^{Σ}) is no more than 15, or the maximum number of daily clients ($I_4^{\max}$) is more than 2.

$$Illegal = \begin{cases} \text{True} & if\ I_4^{\text{avg}} \le 1.28 \text{ and } I_1^{\text{avg}} \le 12.75 \text{ and } I_1^{\max} \le 165 \\ \text{False} & otherwise \end{cases}$$

(a) The rule concluded by JRip



(b) The decision tree developed by using J48

Fig. 3 A rule and a decision tree discovered for IDF using Weka

There were similar behavior analytical results in the linear Pearson's correlations model analyses and in the nonlinear models (i.e., the rules and the decision tree). Firstly, illegal dumping records had few regular clients, e.g., $I_4^{\text{avg}} \leq 1.28$, in all the results. This could be attributed to the fact that small businesses, normally registered as a one-man/truck company, are more prone to commit illegal dumping. They have weaker ties with clients (e.g., a main contractor) and so do not show loyalty or responsibility. The convicted illegal dumping cases in Hong Kong echo this analysis. Another indicator is average time spent in waste facilities, e.g., $I_1^{\text{avg}} \leq 12.75$ or 13.19 minutes. Since the trucks were in the same model, no matter from the target ('illegal') or control ('normal') groups, their time spent should not differ significantly. However, Table 3 shows a significant difference. One possible reason is that trucks in the target group deliberately avoid a long wait time in rush hours or on busy days. Fig. 4 shows a curve of the average waiting time of all the records and of the target group, with both curves increasing slightly over time. In Hong Kong, these waste haulers are often freelance businesses charging by trip. Within the fast-paced construction industry, small waste-hauling businesses are more likely to risk illegal dumping to save time and maximize profits.

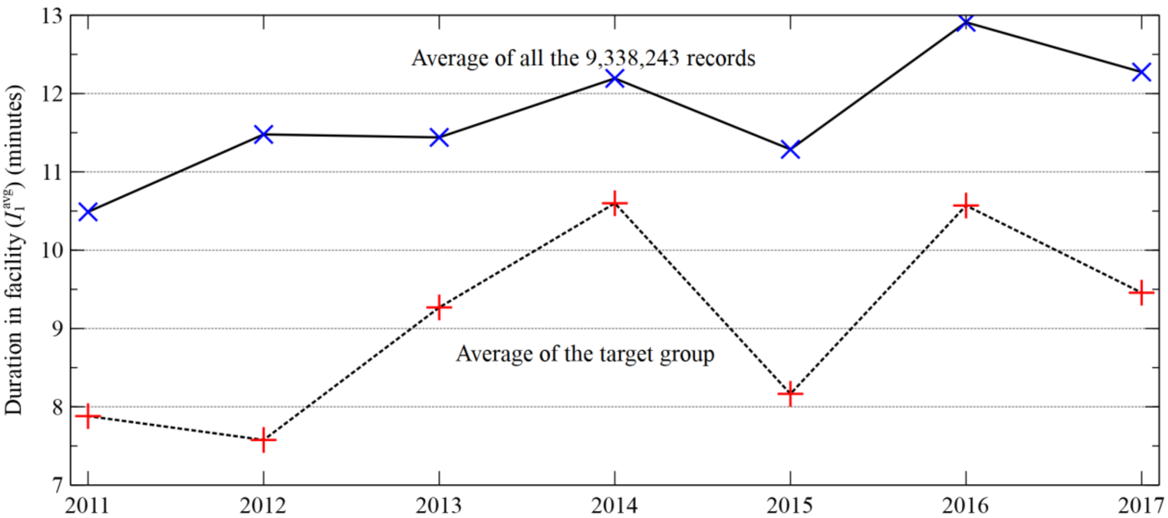


Fig. 4 Slow increments of the average time spent in waste facilities

In summary, two major behavioral drivers were identified: (a) small freelance business and (b) long queuing time. As shown in Fig. 5, long queuing time has long been a problem in Hong Kong due to the outdated service capacity of the government's waste facilities. For example, there are only three landfill sites in Hong Kong, and each has only one entrance and one exit gate. With more gates, unnecessary queuing time could be considerably reduced and at least one driving factor of illegal dumping alleviated.

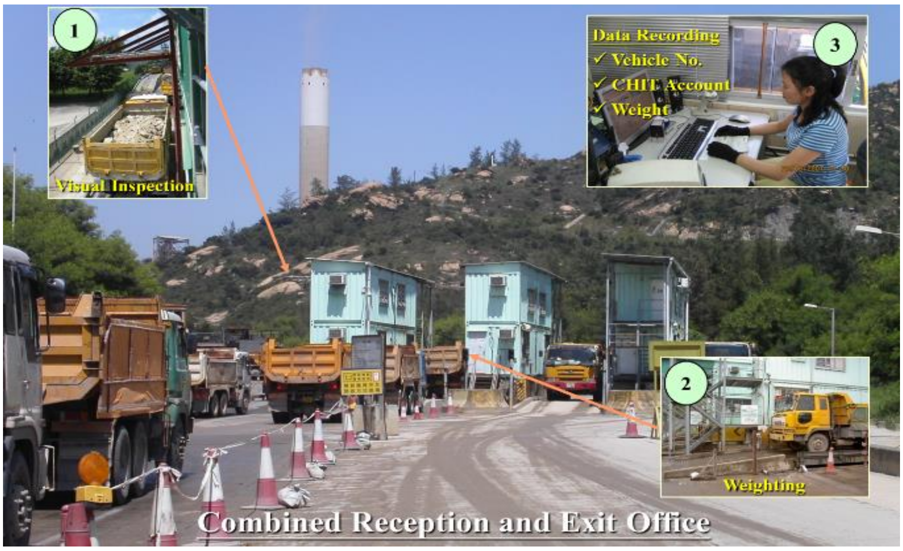


Fig. 5 Queuing at a waste facility in Hong Kong [Source: CEDD]

The IDF can also classify suspected illegal dumping records by applying the concluded reasoning models, e.g., the rule and the decision tree in Fig. 3. First, the models for IDF were selected using 10-folf cross-validation experiments, which are well-established for model selection (Fushiki, 2011). Over 30 classification methods of four types were tested, including: (1) tree, (2) rule, (3) function, and (4) meta-model. Table 4 lists the best method selected for each type and the performance metrics including precision, recall, and F_1-measure. The best method for tree models was J48 with 0.843 precision, 0.842 recall, and 0.842 F_1-measure; JRip was the selected method for rule models, yet with a slight lower-level performance. Both decision trees and rules can be interpreted by humans, as shown in Fig. 3. The selected method for the function model was Radial Basis Function (RBF) classifier (Frank, 2014), which returned a high-level performance of 0.862 precision, 0.860 recall, and 0.860 F_1-measure. Random Committee (Lira et al., 2007), a meta-model method that employs random trees as a low-level method for evolutionary tuning, returned the same performance as J48. The results of the latter two methods were not interpretable directly. Visibility of the classification models, as shown in Fig. 3, is important for domain experts to understand and verify the IDF model. Therefore, J48 can be used as the method for training the IDF model and classifying all the yearly truck records beside the target and control groups.

Table 4. The IDF model selection with 10-fold cross-validation

IDF's reasoning model		The results of 10-fold cross-validation experiments (higher is better)			
Human readable?	Type of method	The best method for the type	Precision	Recall	F_1-measure
Yes	Tree	J48	0.843	0.842	0.842
	Rule	JRip	0.811	0.807	0.807
No	Function	RBF classifier	0.862	0.860	0.860
	Meta-model	Random committee	0.843	0.842	0.842

The IDF model was applied using the selected J48 method to filter the suspected illegal dumping actions from the database, with a view to understanding the overall magnitude of the illegal dumping problem. The target dataset was a CSV format table comprising 10,924 rows of yearly statistics of 3,189 waste trucks, calculated from the about 10 million records (1.4GB file size) introduced in Section 3 using MySQL statistical functions. The prediction results of the IDF indicated that 546 trucks, about 17%, had suspected illegal dumping actions, as shown in Appendix B. Table 5 shows an excerpt of the suspected trucks, with a check mark indicating possible illegal dumping actions in a year.

Table 5. An excerpt of the most suspected trucks with detected illegal dumping actions

Truck plate No.	Suspected illegal dumping actions predicted by IDF using J48							
	2011	2012	2013	2014	2015	2016	2017	Suspicion score (%)

A***2	✓	✓	✓	✓	✓	N.A.	N.A.	100
B***	N.A.	N.A.	N.A.	✓	✓	✓	✓	100
B***3	N.A.	✓	✓	✓	✓	✓	✓	100
B***30	N.A.	N.A.	N.A.	✓	✓	✓	✓	100
B***0	✓	✓	✓	✓	N.A.	N.A.	✓	100
B***62	✓	✓	✓	✓	✓	N.A.	N.A.	100
B***	N.A.	N.A.	N.A.	✓	✓	✓	✓	100
B***1	✓	✓	✓	✓	✓	✓	✓	100
B***96	✓	✓	✓	✓	✓	N.A.	N.A.	100

N.A. indicates no available data

6. Discussion

6.1 The trilogy of big data analytics for illegal dumping identification

Too often, the media play up big data's power to tackle crime, corruption, and fraud, adding little to knowledge on how to actually apply big data to solve these contemporary urban issues. Based on previous studies, this paper formalises the methodology of using big data analytics for urban crime identification as a 'trilogy' of 'Identifying indicators/monitors of anomalies', 'Developing a big data analytical model', and 'Model training, calibration, and evaluation'. This paper enriches the trilogy through a vivid case study.

The first step in using big data analytics to identify urban crimes is to characterize crimimal behavior and develop a set of indicators to guage the behavior. These indicators are heavily dependent on specific criminal scenarios. In this study, an understanding of illegal dumpers' economic motivations and particular behavior patterns was first developed. Some red-flag indicators stemmed from our own knowledge, literature review, and desktop studies, while others were contributed by experienced individuals including LegCo councillors, reporters, criminologists, and environmental activists.

With the indicators of anomalies, the next step is to develop big data analytical models. For indicators to be used for modelling, they must be readily measurable using the big data; if not, they must be dropped from the indicator set. It is expected that a single identified anomaly may not imply a crime, but an accumulation of anomalies from multiple indicators increases the confidence with which a suspected crime can be identified. With the increase of the red-flag indicators, certainly, the required data should be bigger. It is often the case that there is no *prior* knowledge on the 'weights' of the indicators (i.e., linear relationship), or how the indicators interact with each other (i.e., non-linear relationship) in determining a suspected crime. One needs proper big data analytical tools. In addition to the decision tree adopted in this study, many other analytics such as case-based reasoning, artificial neural network, decision-tree, graphical/statistical outlier detection, and clustering, have been raised by researchers (e.g., Baesens, 2015; Vona, 2017).

The third step is model training, calibration, and evaluation to determine the optimal big data analytical model for urban crime identification. This is apparently a data-driven process. The true cases (e.g., the convicted illegal dumping cases in this study) are fed into the models to determine the weights of the indicators, or the way they interact. Model calibration is conducted during this process. More fraudulent or legitimate cases are fed into the calibrated model to validate it before it can be accepted to detect crimes in the future. There are some cases wherein anomalous behaviors are changing quickly, and the models should be adaptive enough to these changes (Fawcett and Provost, 1997).

6.2 Prospects and challenges of big data analytics for identifying illegal dumping

The predictions, as shown in Table 5, can only be used for filtering possible offenders. Similar to big data analytics in other urban crime identification cases (e.g., corruption in public procurement, or credit card fraud), they cannot be used for prosecution. Direct evidence must be obtained from other means. That does not mean the post-mortem analyses using big data are useless. Rather, they can be used as important information for follow-up interventions to combat illegal dumping, such as opening more gates at waste disposal facilities. Government departments have debated using GPS to track all waste hauling trucks but such a measure would be prohibitively expensive. However, the measure could be piloted in highly suspected vehicles as a means of deterrence.

Readers might have noticed that rather than needing a long list of indicators, just two can satisfactorily detect suspected illegal dumping in this study. It just so happens that these two indicators could be computed and utilized with the available big data. However, data may not be so readily accessible in other urban crime identification scenarios. Data analysts are therefore discussing possible strategies to use technical means (e.g., sensor networks, surveillance) to proactively collect big data.

The capture and use of big data have both benefits and risks. Ever since its advent, there have been ethical concerns over misuse of its power. Although the conceptual, regulatory, and institutional resources of research ethics have developed greatly over the past few decades and are familiar to researchers, there remain many unaddressed issues with respect to the big data phenomenon (Boyd and Crawford, 2012). Existing norms governing data and research ethics have difficulty accommodating the special features of big data. The ethics of its use are intimately tied to questions of ownership, access and intention, all of which are often disputed. Social media sites such as Facebook claim to own their big data and have exclusive access to it, even though it is actually contributed by users.

Informed consent, premised on the liberal tenets of individual autonomy, freedom of choice and rationality, is a cornerstone of personal data regulation and ethics (Cheung, 2016). However, researchers cannot possibly obtain consent from every waste hauler passively leaving data as a part of their operations. Traditional de-identification approaches (e.g., anonymization, pseudonymization, encryption, or data sharding) to protect privacy and confidentiality and allow analysis to proceed are problematic in big data, as even anonymized data can be re-identified and attributed to specific individuals (Ohm, 2009). De-identification is not always helpful as companies can be re-identified from records in other databases. 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.

7. Conclusions

Illegal dumping of construction waste has long plagued cities around the world, and its surreptitious nature has presented a major challenge to the identification of suspected cases. Utilizing more than nine million waste disposal records over the past eight years in Hong Kong and a decision tree as the major analytical tool, this research identified 546 waste hauling trucks suspected of involvement in illegal dumping. Through big data analytics, previously unknown characteristics of illegal dumpers were identified: for example, they are freelance, and less patient in queuing at government waste disposal facilities. These characteristics exist alongside known motivations such as saving time and cost, or simply convenience. Although the analytical results cannot be used as evidence to prosecute suspected offenders, they offer important decision-support information for follow-up interventions to combat illegal dumping.

This research also makes significant methodological contributions, particularly to the field of big data analytics for urban crime identification by formalizing the methodology as a trilogy. Specifically, this paper demonstrates that indicators of anomalies can be identified using prior knowledge, traditional research methods (e.g., interviews, observation), and big data analytics. The best method for tree models was J48 with 0.843 precision, 0.842 recall, and 0.842 F1-measure; a high-level performance returned. Even with big data analytics there is no one-size-fits-all solution to urban crime identification. This paper, however, enriches the field by providing a vivid case study which can serve as a useful reference for other big data-enabled urban crime identification scenarios such as corruption in public procurement and fraud detection.

Big data analytics has serious potential ethical ramifications and should be treated with caution. Its power is to discover hidden patterns, unknown correlations and other useful information. At the same time, it could lead to privacy infringement and other issues that still have no readily available theoretical explanation or practical solution.

Declarations of interest: none

References

Agrawal, A. (2006). Engaging the inventor: Exploring licensing strategies for university inventions and the role of latent knowledge. Strategic Management Journal, 27(1), 63-79.

Audit Commission (2016). Management of abandoned construction and demolition materials. Chapter 4 of the Director of Audit's Report No. 67, Audit Commission.

Baesens, B., Van Vlasselaer, V., and Verbeke, W. (2015). Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection. John Wiley & Sons.

Chen, M., Mao, S., and Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171-209.

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

Cohen, W. W. (1995). Fast effective rule induction. In Machine Learning Proceedings 1995 (pp. 115-123).

Department for Environment Food & Rural Affairs (Defra) (2017). Fly-tipping statistics for England, 2015/16, Defra, UK.

Dey, P. K. (2002). Project risk management: a combined analytic hierarchy process and decision tree approach. Cost Engineering, 44(3), 13-27.

Environmental Protection Department (EPD) (2017). Monitoring of solid waste in Hong Kong: Waste statistics for 2015. EPD.

Environment, Transport and Works Bureau (2006). LCQ 4: Illegal dumping of construction waste (Press Releases). Available at: https://goo.gl/4piUnG.

Ernst & Young (2014). Big data to play key role in fraud detection/prevention. Available at: https://www.ey.com/in/en/newsroom/news-releases/ey-press-release-big-data-to-play-key-role-in-fraud-detection-prevention.

Esa, M. R., Halog, A., and Rigamonti, L. (2017). Strategies for minimizing construction and demolition wastes in Malaysia. Resources, Conservation and Recycling, 120, 219-229.

Fawcett, T., and Provost, F. (1997). Adaptive fraud detection. Data Mining and Knowledge Discovery, 1(3), 291-316.

Fazekas, M., and Tóth, I. J. (2014). New ways to measure institutionalised grand corruption in public procurement. U4 Anti-Corruption Resource Centre, Bergen, Norway.

Fazekas, M., Tóth, I. J., and King, L. P. (2013). Corruption manual for beginners "corruption techniques" in public procurement with examples from Hungary. Working Paper CRCB-WP/2013:01. The Corruption Research Center Budapest, Hungary.

Frank, E. (2014). Fully supervised training of Gaussian radial basis function networks in WEKA. Technical report, University of Waikato, New Zealand. Available at: https://researchcommons.waikato.ac.nz/bitstream/handle/10289/8683/uow-cs-wp-2014-04.pdf.

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., and Trigg, L. (2009). Weka-a machine learning workbench for data mining. In Data mining and knowledge discovery handbook (pp. 1269-1277). Springer, Boston, MA.

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. Statistics and Computing, 21(2), 137-146.

Gálvez-Martos, J. L., Styles, D., Schoenberger, H., and Zeschmar-Lahl, B. (2018). Construction and demolition waste best management practice in Europe. Resources, Conservation and Recycling, 136, 166-178.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. New York: springer.

Jin, R., Li, B., Zhou, T., Wanatowski, D., and Piroozfar, P. (2017). An empirical study of perceptions towards construction and demolition waste recycling and reuse in China. Resources, Conservation and Recycling, 126, 86-98.

Lau, J. (2016). Illegal development at Hong Kong wetlands threatens bird life, activists say. South China Morning Post, 29 Mar. 2016.

Li, Z., Shen, G. Q., and Alshawi, M. (2014). Measuring the impact of prefabrication on construction waste reduction: an empirical study in China. Resources, Conservation and Recycling, 91, 27-39.

Lin, G. (2016). Two fined HK$15k each for illegally dumping construction waste in protected Hong Kong wetland. Available at: https://www.hongkongfp.com/2016/08/04/two-fined-hk15k-illegally-dumping-construction-waste-protected-hong-kong-wetland/.

Lira, M. M., de Aquino, R. R., Ferreira, A. A., Carvalho, M. A., Neto, O. N., and Santos, G. S. (2007). Combining multiple artificial neural networks using random committee to decide upon electrical disturbance classification. In 2017 International Joint Conference on Neural Networks. (pp. 2863-2868). IEEE.

Lu, W., Chen, X., Peng, Y., and Shen, L.Y. (2015). Benchmarking construction waste management performance using big data. Resources, Conservation and Recycling. 105(A), 49-58.

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.

Massari, M., and Monzini, P. (2004). Dirty businesses in Italy: a case-study of illegal trafficking in hazardous waste. Global Crime, 6(3-4), 285-304.

Mayer-Schönberger, V., and Cukier, K. (2011). Big data: A revolution that will transform how we live, work, and think. John Murray: London.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., and Barton, D. (2012). Big data: the management revolution. Harvard Business Review, 90(10), 60-68.

Meldrum-Hanna, C., Richards, D., and Davies, A. (2017). Organised network shifting waste to ‘dumping capital of Australia’ to avoid tariffs. Australian Broadcasting Corporation (ABC) News. Available at: https://goo.gl/2TyNTv.

Navarro, J., Grémillet, D., Afán, I., Ramírez, F., Bouten, W., and Forero, M. G. (2016). Feathered detectives: real-time GPS tracking of scavenging gulls pinpoints illegal waste dumping. PloS one, 11(7), e0159974.

Nunes, K. R. A., Mahler, C. F., and Valle, R. A. (2009). Reverse logistics in the Brazilian construction industry. Journal of Environmental Management, 90(12), 3717-3720.

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-MapReduce in cloud systems. International Journal of Cloud Computing and Services Science, 2(1), 16.

Persechino, G., Schiano, P., Lega, M., Napoli, R. M. A., Ferrara, C., and Kosmatka, J. (2010). Aerospace-based support systems and interoperability: The solution to fight illegal dumping. WIT Transactions on Ecology and the Environment, 140, 203-214.

Pramanik, M. I., Lau, R. Y., Yue, W. T., Ye, Y., and Li, C. (2017). Big data analytics for security and criminal investigations. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(4), e1208.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.

Quinlan, J. R. (1993). C4. 5: programs for machine learning. San Mateo, Calif: Morgan Kaufmann.

Romeo, V., Brown, S., and Stuver, S. (2003). A GIS analysis of illegal dumping in the 78249 Zip Code of Bexar County. In 23rd Annual Esri International User Conference, San Diego, CA (pp. 7-11).

Sáez, P. V., del Río Merino, M., Porras-Amores, C., and González, A. S. A. (2014). Assessing the accumulation of construction waste generation during residential building construction works. Resources, Conservation and Recycling, 93, 67-74.

Sagiroglu, S., and Sinanc, D. (2013). Big data: A review. In 2013 International Conference on Collaboration Technologies and Systems (pp. 42-47). IEEE.

Scherer, R. (1995). On the prowl with the sanitation police. Christian Science Monitor, 87, 10-12.

Seror, N., Hareli, S., and Portnov, B. A. (2014). Evaluating the effect of vehicle impoundment policy on illegal construction and demolition waste dumping: Israel as a case study. Waste Management, 34(8), 1436-1445.

Shen, Y., Li, Y., Wu, L., Liu, S., and Wen, Q. (2016). Big data overview. In Enabling the new era of cloud computing: Data security, transfer, and management (pp. 156-184). IGI Global.

Shenkar, M., Chen, Y., and Goldstein, M. (2011). Construction and demolition waste leachiest-a study of their composition and interactions with the unsaturated sub-layer and testing methods of evaluation. Faculty of Agriculture, University of Jerusalem, Research Report Prepared for The Israeli Ministry of Environment Protection (4-402).

Transparency International (2017). Brazil: Open data just made investigating corruption easier. Available at: https://www.transparency.org/news/feature/brazil_open_data_just_made_investigating_corruption_easier.

Unisys (2012). Data analysis using big data tools for financial crime prevention. Available at: http://blogs.unisys.com/eurovoices/data-analysis-using-big-data-tools-for-financial-crime-prevention/.

U.S. Environmental Protection Agency (USEPA) (1998). Illegal dumping prevention guidebook. EPA 905-B-97-001.

Vona, L. W. (2017). Fraud data analytics methodology: The fraud scenario approach to uncovering fraud in core business systems. Wiley.

Webb, B., Marshall, B., Czarnomski, S., and Tilley, N. (2006). Fly-tipping: causes, incentives and solutions. Available at: http://www.tacklingflytipping.com/Documents/NFTPG-Files/Jill-Dando-report-flytipping-research-report.pdf.

Williams, M. L., Burnap, P., and Sloan, L. (2017). Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns. The British Journal of Criminology, 57(2), 320-340.

Witten, I. H., Frank, E., and Hall, M. A. (2011). Data mining: Practical machine learning tools and techniques. Morgan Kaufmann.

World Economic Forum (WEF) (2012). Big data, big impact: New possibilities for international development. Available at: https://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development.

World Economic Forum (WEF) (2015). How to use big data to combat fraud. Available at: https://www.weforum.org/agenda/2015/01/how-to-use-big-data-to-combat-fraud/.

Yuan, H. P., Shen, L. Y., Hao, J. J., and Lu, W. S. (2011). A model for cost–benefit analysis of construction and demolition waste management throughout the waste chain. Resources, Conservation and Recycling, 55(6), 604-612.

Appendix A. The training data in this paper

(see Supplementary Interactive Plot Data Appendix A.csv)

Appendix B. List of the suspected 546 trucks filtered by the proposed IDF

(see Supplementary Interactive Plot Data Appendix B.csv)