

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

2 Weisheng Lu¹

3
4 **Abstract**

5 Illegal dumping, referring to the intentional and criminal abandonment of waste in
6 unauthorized areas, has long plagued governments and environmental agencies worldwide.
7 Despite the tremendous resources spent to combat it, the surreptitious nature of illegal
8 dumping indicates the extreme difficulty in its identification. In 2006, the Construction Waste
9 Disposal Charging Scheme (CWDCS) was implemented, regulating that all construction waste
10 must be disposed of at government waste facilities if not otherwise properly reused or recycled.
11 While the CWDCS has significantly improved construction waste management in Hong Kong,
12 it has also triggered illegal dumping problems. Inspired by the success of big data in combating
13 urban crime, this paper aims to identify illegal dumping cases by mining a publicly available
14 data set containing more than 9 million waste disposal records from 2011 to 2017. Using
15 behavioral indicators and up-to-date big data analytics, possible drivers for illegal dumping
16 (e.g., long queuing times) were identified. The analytical results also produced a list of 546
17 waste hauling trucks suspected of involvement in illegal dumping. This paper contributes to
18 the understanding of illegal dumping behavior and joins the global research community in
19 exploring the value of big data, particularly for combating urban crime. It also presents a three-
20 step big data-enabled urban crime identification methodology comprising 'Behavior
21 characterization', 'Big data analytical model development', and 'Model training, calibration,
22 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;

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

25 **1. Introduction**

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

40

41 Illegal dumping is not only a nuisance in its own right but can also lead to many other problems
42 (Esa et al., 2017). It is a human health concern and can damage the environment in a variety
43 of ways (Romeo et al., 2003). Fly-tipped waste causes habitat destruction, wildlife deaths
44 (Webb et al., 2006), and is a major source of soil and underground water pollution (Shenkar
45 et al., 2011). It also causes aesthetic damage to the natural landscape. When illegal waste
46 dumping is discovered, local governments often dispatch an abatement crew to clean it up as
47 quickly as possible because the contained oil, solvents, fuel, rusted metal, and batteries can

48 cause severe environmental damage. Such clean-up comes at great expense. According to
49 Defra (2017), local authorities in England spent around £49.8 million cleaning up fly-tipped
50 waste in 2015/16 alone. Romeo et al. (2003) report that the City of San Antonio in the U.S.
51 spends hundreds of millions of dollars annually mitigating the environmental consequences of
52 illegal waste dumping. In Hong Kong, Lin (2016) reported that around one hectare of wetland
53 and mangrove forest had been affected by illegal dumping committed by two individuals, with
54 a repair cost estimated by the Environment Protection Department (EPD) at HK\$6 million.

55

56 Governments and environmental agencies have committed extensive resources to combat
57 illegal dumping (Gálvez-Martos et al., 2018). For example, to overcome patchy data collection
58 in order to better understand the scale of the problem, the UK government launched
59 Flycapture® in 2004 (later replaced by WasteDataFlow®), requiring all local authorities and
60 the Environment Agency to submit monthly returns on the number, size, waste types and
61 location types of fly-tips (Webb et al., 2006). Israel has explored vehicle impoundment policy
62 and evaluated its effect on illegal dumping of construction waste (Seror et al., 2014). In Hong
63 Kong, a fly-tipping spotting system (similar to Flycapture®) has been implemented to
64 encourage public reporting of illegal dumping activities. Researchers have also explored
65 various policy and technological recommendations for addressing illegal dumping problems.
66 Examples include enhancing prosecution and enforcement (Yuan et al., 2011), increasing
67 surveillance and ambush (Navarro et al., 2016), adopting new construction method (Li et al.,
68 2014), and using Global Positioning System and satellite images to catch illegal dumping
69 activities (Persechino et al., 2010). However, the effectiveness of these approaches is
70 questionable. Illegal dumping activities are committed stealthily and are thus difficult to catch
71 (Scherer, 1995).

72

73 Big data is increasingly advocated as a powerful instrument for detection and deterrence of
74 contemporary urban issues such as crime, corruption, and fraud. Reports published by the
75 World Economic Forum (WEF) (2015), Transparency International (2017), Ernst & Young
76 (2014), and Unisys (2012) advocate for the power of big data and analytics in reducing
77 corruption and fraud. Since urban crimes are generally conducted in a stealthy way, evidence
78 of them may be deeply buried in a dataset if captured at all. The problem of identifying such
79 activities is extremely difficult to crack. However, offenders may have left unintentional clues
80 or exhibited hidden patterns, identifiable when the dataset is sufficiently large and with the use
81 of proper analytics. Williams et al. (2017) reviewed studies making use of ‘naturally occurring’
82 socially relevant data (e.g., on Twitter or Facebook) to complement and augment conventional
83 curated data to address the classic problem of crime pattern estimation. By combing through
84 datasets on government bidding processes, contracting firms’ financial disclosures, the
85 beneficial ownership of contracting firms, public officials’ tax and family records, and
86 complaints to authorities about bribery by competing contractors, Fazekas et al. (2013) tried
87 to uncover patterns of fraud and bribery in public procurement. There have been several stories
88 on the success of big data, based on which an exploration of how big data analytics can be
89 employed to identify illegal dumping as a contemporary urban issue promises to be intriguing
90 as well as meaningful.

91

92 The primary aim of this research is to develop a big data-driven methodological approach that
93 can be used to identify suspected cases of illegal dumping. It is contextualized in Hong Kong,
94 which has long been suffering from the problems caused by illegal dumping, and focuses on
95 construction waste, which constitutes a prodigious proportion of total municipal solid waste.
96 The rest of this paper is structured as follows. Subsequent to this introductory section is a
97 literature review covering big data and analytics for urban crime identification. The big data

98 of illegal dumping in Hong Kong is introduced in the Section 3. The research methods are
99 described in Section 4. These methods are devised to achieve three specific research objectives:
100 (1) To develop a set of indicators for suspected dumping activities using mixed methods
101 research; (2) To develop an analytical model by applying these indicators and big data
102 analytics; and (3) To train, calibrate, and evaluate the analytical model by trying out different
103 data analytics. Section 5 reports the data analyses and findings and Section 6 is an in-depth
104 discussion including both methodological contributions and policy implications of this
105 research. Conclusions are drawn in the final section.

106

107 **2. Big data analytics to tackle contemporary urban issues**

108 According to Padhy (2013), big data can be characterized as a collection of datasets so large
109 and complex that it is difficult to process using traditional data management tools. Mayer-
110 Schönberger and Cukier (2011) describe big data techniques as ‘things one can do at a large
111 scale that cannot be done at a smaller one, to create a new form of value’. Many researchers
112 accepted Gartner’s three defining characteristics of big data, namely, volume, variety and
113 velocity, or the ‘three Vs’ (McAfee et al., 2012). *Volume* is the quantity of data in the form of
114 records, transactions, tables or files; *velocity* can be expressed in batches, near time, real time
115 and streams; and *variety* can be structured, unstructured, semi-structured or a combination
116 thereof (Chen et al., 2014). Big data analytics can uncover hidden patterns, unknown
117 correlations, and other useful information to guide business predictions and decision-making
118 (Shen et al., 2016); in effect, *value* is advocated as the fourth ‘V’. By analyzing big data, ‘latent
119 knowledge’ (Agrawal et al., 2006) or ‘actionable information’ (WEF, 2012) can be identified.
120
121 Big data success stories abound in a wide range of areas, including science, business, public
122 governance, innovation, competition, and productivity (Sagiroglu and Sinanc, 2013). It is also

123 increasingly being advocated as an effective means of tackling contemporary urban issues
124 such as terrorism, crime, corruption, fraud, and financial non-compliance. Access to big data
125 is a prerequisite for combating urban crime. As Vona (2017) suggests, ‘even the world’s best
126 auditor using the world’s best audit program cannot detect fraud unless their sample includes
127 a fraudulent transaction’. Baesens et al. (2015) estimate that fewer than 0.5% of credit card
128 transactions are typically fraudulent. The problem of identifying fraudulent activities is thus
129 commonly referred to as a needle-in-a-haystack problem. However, when the dataset is
130 sufficiently large, clues unintentionally left or hidden patterns exhibited by offenders become
131 identifiable.

132

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

145

146 A review of previous studies seems to suggest that there is no one-size-fits-all big data-enabled
147 solution to urban criminal issues. A good starting point, however, is to characterize the

148 criminal activities in question, e.g., illegal dumping, and then identify anomalous behavior and
149 ‘red flags’. In a big data-driven methodology comprising ‘Behavior characterization’, ‘Big
150 data analytical model development’ and ‘Model calibration’, these three steps in combination
151 can indicate, at the very least, highly suspected activities. In the context of public procurement,
152 for example, Fazekas and Tóth (2014) characterized the behavior by proposing more than 30
153 indicators of high corruption risk. Based on the characteristics and the indicators, the next step
154 is to develop the big data analytical model. Data analytical methods ranging from ‘simple’
155 regression analysis to complex techniques such as support vector machines, artificial neural
156 networks, association rules, case-based reasoning, and *K*-means clustering are widely applied
157 in urban crime detection (Fawcett and Provost, 1997). Finally, the big data analytical model
158 needs to be trained, calibrated, and evaluated using known cases, e.g., crime convictions,
159 before it can be applied to the big data set to identify other suspected cases and for further
160 follow-up actions.

161

162 **3. The big data of illegal dumping in Hong Kong**

163 In Hong Kong, the adverse environmental impacts of construction waste resulting from
164 creation of its impressive built environment are a grave concern. As in other states and
165 territories, construction waste in Hong Kong is classified into inert and non-inert components.
166 EPD (2017) statistics show that the total solid waste deposited in Hong Kong landfills in 2015
167 amounted to 15,102 tons per day (tpd), of which 4,200 tpd, or 27.8%, was from construction
168 activities. Thus, construction generates around one-quarter of the total solid waste finding its
169 way into landfills. Owing to its significant adverse impacts, construction waste is heavily
170 regulated in Hong Kong, and a series of statutory and non-statutory policies, including
171 regulations, codes, and schemes have been introduced over the past few decades (Lu and Tam,
172 2013). In particular, the Construction Waste Disposal Charging Scheme (CWDCS), which

173 mandates that all construction waste, if not otherwise reused or recycled, must be disposed of
174 at government waste facilities (e.g., landfills, offsite sorting facilities [OSFs] or public fill
175 banks) was implemented in 2006. According to this scheme, the main contractor is charged
176 HK\$200 for every ton of non-inert waste it dumps in landfills; HK\$175 per ton for mixed inert
177 and non-inert waste accepted by OSFs; and HK\$71 per ton for inert waste accepted by public
178 fills (raised from HK\$125, HK\$100, and HK\$27 respectively in April 2017). As a policy
179 system, the CWDCS together with its enforcement measures has been praised for its efficiency
180 in construction waste minimization (Lu et al., 2015).

181
182 At the same time, the CWDCS has incentivized illegal dumping. Illegal disposal of one load
183 of construction waste immediately saves contractors between HK\$405 to HK\$3,750 in tipping
184 fees, depending on the volume and type of waste. This does not include savings in transport
185 costs (normally HK\$800-1,500 per trip) and waiting time at government facilities. In response
186 to a Legislative Council (LegCo) query, the Environment, Transport and Works Bureau (2006)
187 reported that 508 complaints of construction waste illegal dumping were received between 20
188 January 2006 (the CWDCS implementation date) and 31 May 2006, a significant rise from the
189 101 received in the same period in 2005. After that, fly-tipping reports have continuously
190 become epidemic. Hong Kong's Audit Commission (2016) recently found that public reports
191 of illegally dumped construction materials increased a phenomenal 328% in 2015, rising from
192 1,517 to 6,499. In that year, 6,300 tons of illegally dumped construction materials were cleared
193 by government departments. Without quick abatement, such waste can cause severe
194 environmental damage. For example, environmentalists have warned that wetland fauna and
195 mangroves are particularly vulnerable to illegal dumping (Lau, 2016).

196
197 The structure of the big data is illustrated in Fig. 1, which comprises:

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

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

A table of 9,338,423 records in CWDCS (2011-2017)

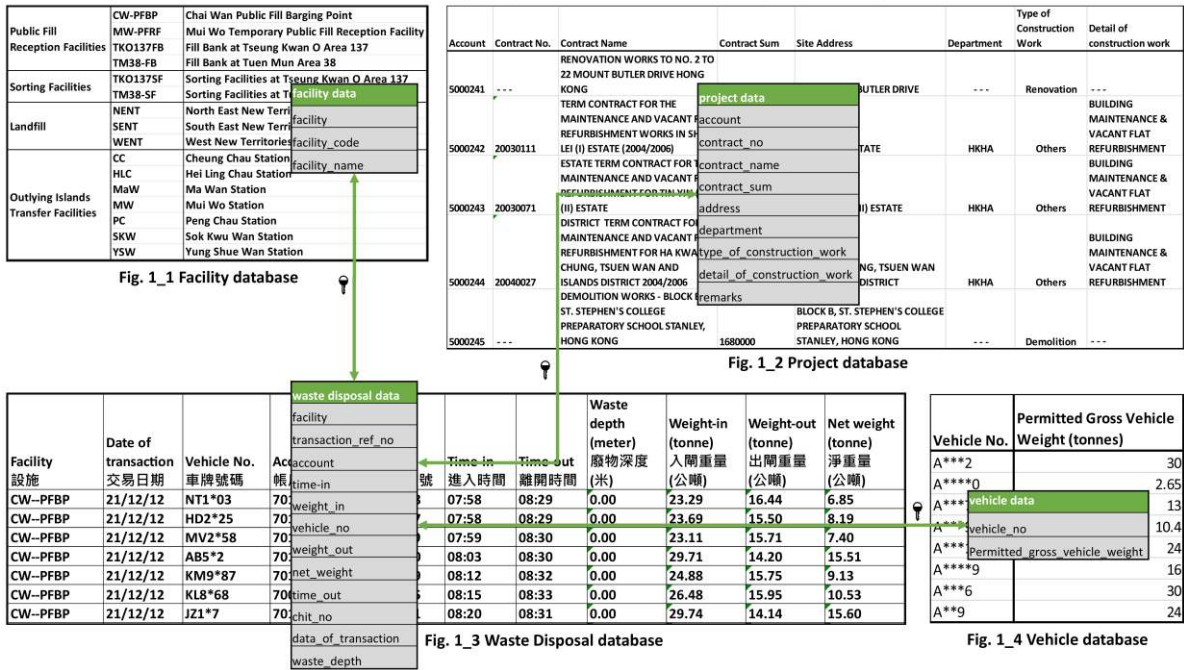


Fig. 1 The big data structure and example records

214
215
216

217 4. Methods

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

225

226 4.1 Developing a set of red-flag indicators

227 To develop red-flag indicators, illegal dumping behavior is characterized by adopting a mixed
228 method approach. Since 2013, the research team has conducted a series of research projects
229 with construction clients (both public and private), main contractors, government departments
230 (e.g., the EPD and Construction Industry Council), LegCo members, waste haulers, unions,

231 environmentalists, and other informants to try to understand the motivations for illegal
 232 dumping and offenders' behavior.

233

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

246

247 Table 1. List of indicators for predicting illegal dumping activities

ID	Name	Unit	Source and calculation
<i>I</i> ₁	Time spent in a facility	Minute	Difference between 'departure time' and 'entering time'
<i>I</i> ₂	Dumping weight	Ton	Difference between 'departure weight' and 'entering weight'
<i>I</i> ₃	Rest/absent days between two working periods	Day	The number of absent days from last dumping record
<i>I</i> ₄	The number of clients served per day	1	The counts of project accounts/clients associated with the same hauler per day
<i>I</i> ₅	Loading ratio	%	Dumping weight/maximum capacity
<i>I</i> ₆	Dumping depth	m	Excessive depth of waste defined in the "waste disposal" database
<i>I</i> ₇	Dumping weight by facility type	Ton	Dumping weight according to type of facility (e.g., landfill, OSF, public fill)
<i>I</i> ₈	Percentage of dumping weight by facility type	%	Percentage of dumping weight according to type of facility
<i>I</i> ₉	Dumping count by facility type	1	Dumping transaction counts by different types of facility per day

248

249 **4.2 Developing an analytical model**

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

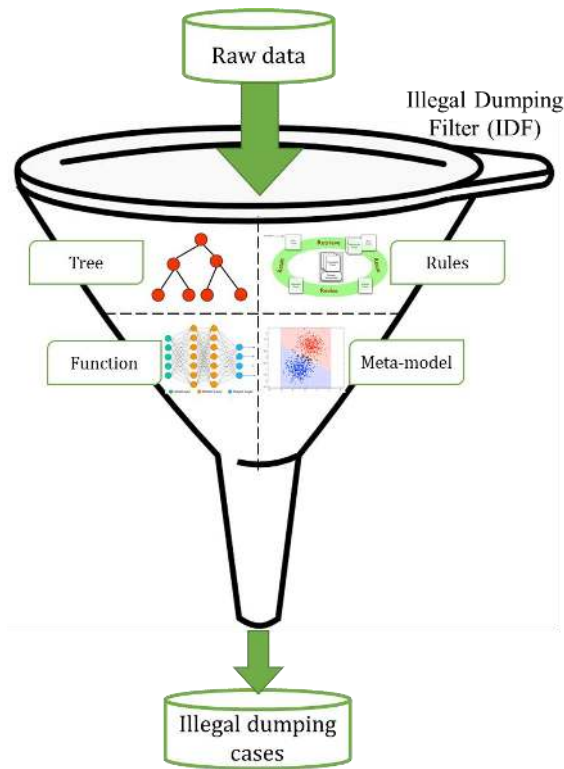


Fig. 2 An illustration of the Illegal Dumping Filter (IDF) in this study

264
265
266

267 **4.3 Training, evaluating, and calibrating the big data analytical model**

268 The third step is to train, evaluate, and calibrate the model before it can be applied to the big
269 data set to identify illegal dumping cases. The sample, mainly comprising cases of illegal
270 dumping convictions, will be used as the experimental/target group while a comparable sample
271 will be used as the control group. It is critical that effectiveness of models is gaugeable. In
272 math language, the effectiveness of the models can be gauged by precision rate, recall rate,
273 and F₁-measure, which is the weighted average of Precision and Recall and is considered more
274 accurate than if they are used individually. The research team needs to adjust the variable
275 settings in the given software platform until a satisfactory result is reached.

276

277 **5. Data analyses and findings**

278 The IDF was first trained on a sampled data with a binary value, i.e., True and False, for the
279 target label ‘committed illegal dumping or not’. A target group included six trucks engaged in

280 illegal dumping based on local news and video clips recorded by environmental activists. The
 281 control group was six non-offending trucks of a similar model and loading capacity. The two
 282 groups accounted for 36,678 dumping records between January 2011 and December 2017;
 283 very big data that might help identify hidden illegal dumping patterns or anomalies. The data
 284 of the two groups was selected into an independent table in MySQL (Version 5.7). Table 2
 285 shows an excerpt of the training sample of yearly statistics of waste dumping behaviors based
 286 on Table 1. The first column in Table 2 indicates the target group ('True') or the control group
 287 ('False'). For the indicators $I_1, I_2, I_3, I_4, I_5,$ and $I_6,$ seven statistics, i.e., the minimum, 5%
 288 percentile, average, maximum, 95% percentile, sum, and standard deviation, were calculated
 289 using MySQL functions, e.g., $avg()$ and $max()$, for each indicator. For the indicators $I_7, I_8,$ and
 290 $I_9,$ four-yearly statistics by facility types, i.e., the transaction counts for land fill, public fill,
 291 sorting, and islands, were computed. The final training sample of the IDF, as shown in Table
 292 2, was a 'monster' data table consisting of 55 columns and 57 rows, with personal or privacy
 293 data anonymized in comma separated vector (CSV) format.

294
 295 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	

296
 297
 298

296
 297 The next step was to identify the behavioral drivers of illegal dumping by trying linear models.
 298 A straightforward and easy-to-understand metric of the driving factors is Pearson's linear

299 correlation coefficient. The correlations between the 54 indicators and the label in Table 2
 300 were first tested, using IBM SPSS (version 24.0). Table 3 lists the eleven indicators showing
 301 statistical significance at the level 0.01 (two-tailed). The eleven indicators are statistics of three
 302 types of indicators, i.e., the duration in facility (I_1), the number of daily clients served (I_4), and
 303 waste depth (I_6). In other words, the three indicators were more related with the drivers of
 304 illegal dumping. Three statistics of I_4 , i.e., I_4^{avg} , $I_4^{95\%}$, and I_4^σ had moderate negative
 305 correlations, while all the rest had weak negative correlations. To sum up, a truck with illegal
 306 dumping behaviors usually had fewer daily clients, less time spent at the government facilities,
 307 and less waste depth in the government's waste records.

308

309 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

310

311

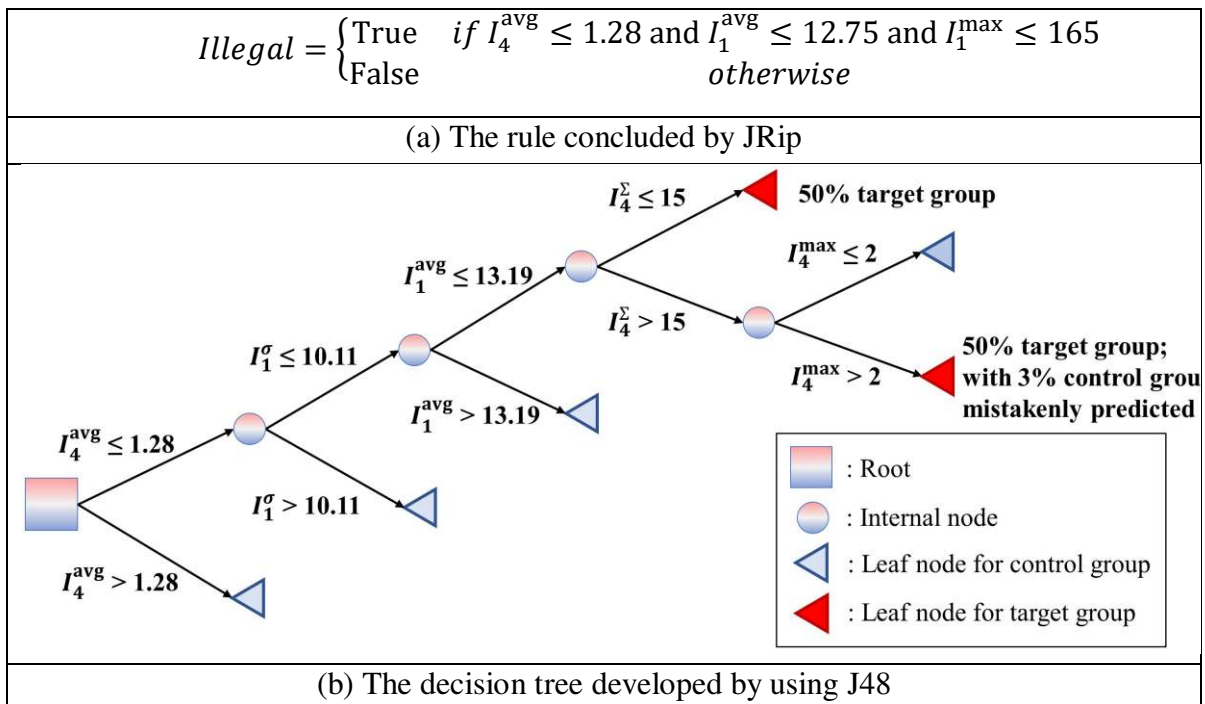
312 The training data was further processed using Weka (version 3.9), which is an open source
 313 data mining software program (Frank et al., 2009). Data mining methods can discover
 314 nonlinear models of correlations, which approximates the illegal dumping behaviors better
 315 than the linear correlations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

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

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

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

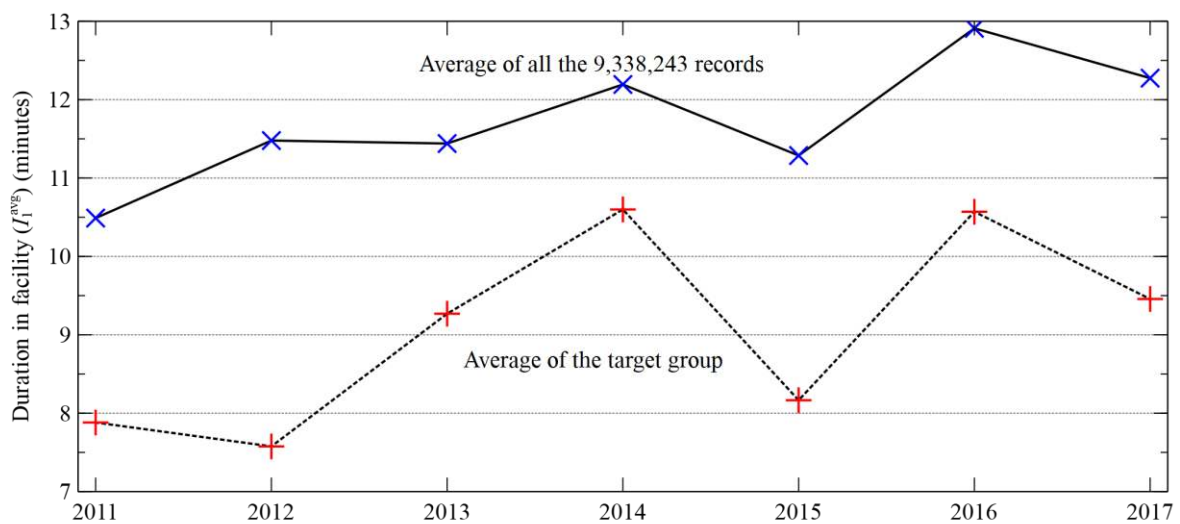
336



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

338

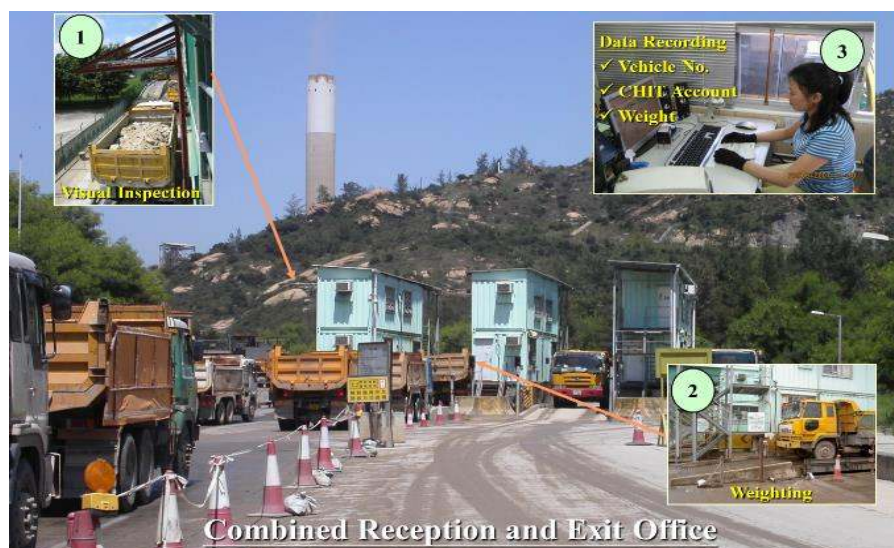
339 There were similar behavior analytical results in the linear Pearson’s correlations model
 340 analyses and in the nonlinear models (i.e., the rules and the decision tree). Firstly, illegal
 341 dumping records had few regular clients, e.g., $I_4^{\text{avg}} \leq 1.28$, in all the results. This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



355
 356 Fig. 4 Slow increments of the average time spent in waste facilities
 357

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

364



365

366

367

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

368 The IDF can also classify suspected illegal dumping records by applying the concluded
369 reasoning models, e.g., the rule and the decision tree in Fig. 3. First, the models for IDF were
370 selected using 10-fold cross-validation experiments, which are well-established for model
371 selection (Fushiki, 2011). Over 30 classification methods of four types were tested, including:
372 (1) tree, (2) rule, (3) function, and (4) meta-model. Table 4 lists the best method selected for
373 each type and the performance metrics including precision, recall, and F_1 -measure. The best
374 method for tree models was J48 with 0.843 precision, 0.842 recall, and 0.842 F_1 -measure; JRip
375 was the selected method for rule models, yet with a slight lower-level performance. Both
376 decision trees and rules can be interpreted by humans, as shown in Fig. 3. The selected method

377 for the function model was Radial Basis Function (RBF) classifier (Frank, 2014), which
 378 returned a high-level performance of 0.862 precision, 0.860 recall, and 0.860 F₁-measure.
 379 Random Committee (Lira et al., 2007), a meta-model method that employs random trees as a
 380 low-level method for evolutionary tuning, returned the same performance as J48. The results
 381 of the latter two methods were not interpretable directly. Visibility of the classification models,
 382 as shown in Fig. 3, is important for domain experts to understand and verify the IDF model.
 383 Therefore, J48 can be used as the method for training the IDF model and classifying all the
 384 yearly truck records beside the target and control groups.

385

386 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

387

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

396

397 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							Suspicion score (%)
	2011	2012	2013	2014	2015	2016	2017	

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

398 N.A. indicates no available data

399

400 **6. Discussion**

401 ***6.1 The trilogy of big data analytics for illegal dumping identification***

402 Too often, the media play up big data's power to tackle crime, corruption, and fraud, adding
403 little to knowledge on how to actually apply big data to solve these contemporary urban issues.

404 Based on previous studies, this paper formalises the methodology of using big data analytics
405 for urban crime identification as a 'trilogy' of 'Identifying indicators/monitors of anomalies',
406 'Developing a big data analytical model', and 'Model training, calibration, and evaluation'.

407 This paper enriches the trilogy through a vivid case study.

408

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

416

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

429

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

438

439 ***6.2 Prospects and challenges of big data analytics for identifying illegal dumping***

440 The predictions, as shown in Table 5, can only be used for filtering possible offenders. Similar
441 to big data analytics in other urban crime identification cases (e.g., corruption in public

442 procurement, or credit card fraud), they cannot be used for prosecution. Direct evidence must
443 be obtained from other means. That does not mean the post-mortem analyses using big data
444 are useless. Rather, they can be used as important information for follow-up interventions to
445 combat illegal dumping, such as opening more gates at waste disposal facilities. Government
446 departments have debated using GPS to track all waste hauling trucks but such a measure
447 would be prohibitively expensive. However, the measure could be piloted in highly suspected
448 vehicles as a means of deterrence.

449

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

456

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

466

467 Informed consent, premised on the liberal tenets of individual autonomy, freedom of choice
468 and rationality, is a cornerstone of personal data regulation and ethics (Cheung, 2016).
469 However, researchers cannot possibly obtain consent from every waste hauler passively
470 leaving data as a part of their operations. Traditional de-identification approaches (e.g.,
471 anonymization, pseudonymization, encryption, or data sharding) to protect privacy and
472 confidentiality and allow analysis to proceed are problematic in big data, as even anonymized
473 data can be re-identified and attributed to specific individuals (Ohm, 2009). De-identification
474 is not always helpful as companies can be re-identified from records in other databases.
475 Researchers thus need to start thinking more clearly about accountability of big data analytics,
476 identifying methods, predictions and inferences that can be considered ethical and those that
477 are not.

478

479 **7. Conclusions**

480 Illegal dumping of construction waste has long plagued cities around the world, and its
481 surreptitious nature has presented a major challenge to the identification of suspected cases.
482 Utilizing more than nine million waste disposal records over the past eight years in Hong Kong
483 and a decision tree as the major analytical tool, this research identified 546 waste hauling
484 trucks suspected of involvement in illegal dumping. Through big data analytics, previously
485 unknown characteristics of illegal dumpers were identified: for example, they are freelance,
486 and less patient in queuing at government waste disposal facilities. These characteristics exist
487 alongside known motivations such as saving time and cost, or simply convenience. Although
488 the analytical results cannot be used as evidence to prosecute suspected offenders, they offer
489 important decision-support information for follow-up interventions to combat illegal dumping.
490

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

501

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

506

507 **Declarations of interest: none**

508

509 **References**

510 Agrawal, A. (2006). Engaging the inventor: Exploring licensing strategies for university
511 inventions and the role of latent knowledge. *Strategic Management Journal*, 27(1), 63-79.
512 Audit Commission (2016). Management of abandoned construction and demolition materials.
513 Chapter 4 of the Director of Audit's Report No. 67, Audit Commission.

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

517 Chen, M., Mao, S., and Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*,
518 19(2), 171-209.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

572 Lu, W., Chen, X., Peng, Y., and Shen, L.Y. (2015). Benchmarking construction waste
573 management performance using big data. *Resources, Conservation and Recycling*. 105(A),
574 49-58.

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

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

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

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

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

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

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

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

594 Padhy, R. P. (2013). Big data processing with Hadoop-MapReduce in cloud systems.
595 *International Journal of Cloud Computing and Services Science*, 2(1), 16.

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

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

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

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

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

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

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

613 Scherer, R. (1995). On the prowl with the sanitation police. *Christian Science Monitor*, 87, 10-
614 12.

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

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

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

625 Transparency International (2017). Brazil: Open data just made investigating corruption easier.
626 Available at:
627 [https://www.transparency.org/news/feature/brazil_open_data_just_made_investigating_](https://www.transparency.org/news/feature/brazil_open_data_just_made_investigating_corruption_easier)
628 [corruption_easier](https://www.transparency.org/news/feature/brazil_open_data_just_made_investigating_corruption_easier).

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

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

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

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

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

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

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

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

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

652
653
654

655 **Appendix A. The training data in this paper**

656
657 (see Supplementary Interactive Plot Data Appendix A.csv)

658

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

660

661 (see Supplementary Interactive Plot Data Appendix B.csv)