The Responsibility of Waste Production: comparison of European Waste Statistics Regulation and Dutch National Waste Registry

Rusne Sileryte^{a,b,c}, Alexander Wandl^a and Arjan
van Timmeren^{a,b}

^aFaculty of Architecture and the Built Environment, Delft University of Technology
Julianalaan 134, Delft, the Netherlands

^bAmsterdam Institute of Advanced Metropolitan Solutions, Gebouw 027W, Kattenburgerstraat
5, Amsterdam, the Netherlands

^cCorresponding author: Rusne Sileryte, r.sileryte@tudelft.nl, orcid=0000-0002-8245-3016

ARTICLE INFO

Keywords:
waste producers
waste statistics regula-
tion
waste statistics
circular economy

ABSTRACT

The announcement of a new Circular Economy Action Plan as part of the European Green Deal policy has created an urgent need for the reliable information on resource flows to monitor and support the transition. An updated Monitoring Framework is set to rely as much as possible on European Statistics, however at this point there are no changes introduced in supranational statistics regulations. This raises a question whether regulations that have been created before the paradigm shift are still able to supply us with statistics necessary to inform policy makers about current successful practices, remaining barriers, positive and negative impacts of the transition and overall progress towards the set goals. This paper focuses on the Waste Statistics Regulation, specifically the relationship between the types of waste and economic activities which are considered to be the waste producers. Dutch National Waste Registry is used as a case study to compare the guidelines on pan-European waste data collection to the actual waste reports. The task of this publication is dual: first, it presents a computational method to link waste producers to their economic activities using a national Trade Registry. Secondly, it explores to which extent the guidelines available in the Waste Statistics Regulation correspond to the operational reality. An extensive discussion of the results provides insights and recommendations for the future guidelines of waste statistics to support circular economy transition.

42 **1. Introduction**

43 Waste generation and its treatment is often the starting point for monitoring
44 the transition towards a circular economy (CE) as it represents the final stage of
45 the undesired linear economy. The ability to prevent waste disposal and generate
46 secondary materials for a long time has signified success in preventing material
47 losses and protecting the environment (Melosi, 2004). However, environmental
48 research in the last two decades has exposed that waste recycling alone is not
49 sufficient to achieve a sustainable economy in the light of increasing global
50 resource scarcity (Geyer et al., 2016) putting waste prevention, material reuse
51 and upcycling higher on the political agendas than ever before (Morseletto,
52 2020).

53 To support the paradigm shift in 2020 the European Commission (EC)
54 has announced a new Circular Economy Action Plan (CEAP) as part of the
55 European Green Deal policy (European Commission, 2020). Streamlining the
56 regulatory framework requires a monitoring framework that is able to inform
57 policy makers about current successful practices, remaining barriers, positive
58 and negative impacts of the transition and overall progress towards the set
59 goals. The new CEAP outlines that an updated Monitoring Framework will
60 rely as much as possible on European Statistics (European Commission, 2020),
61 however at this point does not announce if changes in supranational statistics
62 regulations would be introduced.

ORCID(s): 0000-0002-8245-3016 (R. Sileryte)

63 To date, the EC Regulation No 2150/2002 on Waste Statistics enables
64 Eurostat to collect statistics from member states on 1) waste generation per
65 economic sector and household consumption; 2) waste treatment by waste
66 category and type of treatment and 3) number and capacity of recovery and
67 disposal facilities (per NUTS 2 region) and population served. Each country is
68 free to choose and apply any data collection method as long as it complies to
69 the provided guidelines.

70 A number of reports aimed at evaluating the transition on the EU level em-
71 phasise that significant variations of data quality and the lack of harmonisation
72 in data collection methodologies between the member states hinder effective
73 monitoring and knowledge transfer (Hanssen et al., 2013; Deloitte, 2017; Nuss
74 et al., 2017). At the same time the most common methodology used to provide
75 waste statistics is to scale up data collected from a sample of companies to a
76 whole sector. Waste treatment statistics are often collected directly from the
77 waste treatment facilities and therefore disconnected from the waste producers.

78 This way of reporting statistics works well within the traditional linear
79 economy where waste is a post-factum problem and needs to be dealt with after
80 it has already occurred. However, promoting and supporting such circular econ-
81 omy strategies as waste prevention, design for reuse, prolonged lifespan, etc.
82 involve companies from the full supply chain. Therefore as much information
83 as possible is necessary to understand which interventions can be made to deal
84 with which kinds of waste before it is effectively disposed of.

85 The Netherlands is one of the few member states whose waste data is
86 not based on sample surveys but on consistent waste registration from every

87 company that has a waste permit (Deloitte, 2017). These companies are statu-
88 torily required to register all transported waste, including its producer, waste
89 characteristics, transport methods and final treatment. The caveat of the current
90 system is that companies involved in the waste chain are reported only by their
91 name and address without using unique identifiers able to link the available data
92 with other business registries.

93 The Dutch Chamber of Commerce registry holds information about all
94 companies and their economic activities and could be used to enhance the
95 waste registry with the relevant information. However, there is no key yet
96 that connects both databases. Company names typically have multiple spelling
97 variations, they often have different administrative and operational locations,
98 moreover, multiple companies can be registered at the exact same location.
99 Finally, the waste registry contains spelling and factual errors with regards to
100 companies' names and addresses.

101 Within this scope this paper presents a computational method to link waste
102 producers to their economic activities based on name similarity and geospatial
103 proximity. Secondly, it explores to which extent the guidelines available in the
104 Waste Statistics Regulation are reflected in the available data and if they can
105 be turned into computational rules to improve the quality of linking. Finally,
106 in the discussion consequences for guidelines of waste statistics to support CE
107 transition and legal and operational obstructions affecting policy making and
108 urban planning are explored.

109 2. Related Work

110 Linking waste production to the responsible economic sectors is a common
111 subject in Material Flow Analysis (MFA) (Brunner and Rechberger, 2016)
112 studies. Those studies aim to quantify material flows and stocks in a system
113 with strictly defined temporal and geographic boundaries. Regional MFAs
114 typically aim to quantify material supply, export, consumption and disposal
115 over a chosen period of time. MFA follows the law of conservation of mass and
116 can be framed as a mass balancing exercise where bottom-up data is combined
117 with top-down highly aggregated numbers (Gao et al., 2020; Nuss et al., 2017).
118 Input-output tables are used to couple financial information with physical waste
119 data and to link waste with economic activity. However, Salemdeeb et al.
120 (2016) discuss that the used method cannot effectively distinguish between
121 direct and indirect waste generation and being a top-down, economy-wide
122 approach aggregating the whole economy into only 21 industrial sectors, it
123 cannot distinguish sufficiently product groups or individual companies.

124 Region-wide granular bottom-up datasets that describe material input-
125 output nodes are rare and no published examples could be found that aim to link
126 waste producers to their economic sectors on a legal entity level. Nevertheless,
127 linking diverse registries of legal entities without a common identifier is a
128 common problem arising in various fields. Identifying records that correspond
129 to the same real-world entity appears under the names of entity resolution,
130 linkage, matching, merge, purge or deduplication (Burdick et al., 2015).

131 A rule-based matching approach using both entity name and address sim-
132 ilarity can be found in such domains as the investigation of health-related

133 behaviours dependent on living environments (Hirsch et al., 2020; Mendez
134 et al., 2014), validating names and addresses of transportation and logistic
135 entities (Guermazi et al., 2020), matching observations across financial datasets
136 (Cohen et al., 2018; Burdick et al., 2015), identifying same entities in patent
137 files (Medvedev and Ulanov, 2011; Magnani and Montesi, 2007). Most of them
138 conclude that domain expert knowledge integration improves or would improve
139 matching results (Pilania and Kumaran, 2019; Cohen et al., 2018; Choi et al.,
140 2017; Antoni et al., 2018; Schild, 2016; Mendez et al., 2014; Magnani and
141 Montesi, 2007).

142 This paper further builds on the existing examples of entity matching using
143 standard computational methods to evaluate name similarity and geospatial
144 proximity between potential matches. Additional rules that could improve the
145 matching results are derived from the "Manual on Waste Statistics" (Eurostat,
146 2013) that guides the data collection process in Member States. The importance
147 is not so much to obtain the highest possible matching score but to understand
148 the reasons behind the unsuccessful matches as they reveal the differences
149 between the official guidelines and the operational reality of data collection
150 and waste disposal.

151 **3. Methods**

152 The computational entity linkage process follows six phases as defined by
153 Köpcke and Rahm (2010): data preprocessing, indexing, pairwise comparisons,
154 classification, manual review, evaluation, and refinement. A random sample
155 of 1000 companies ((8% of the full dataset) evenly distributed throughout the

156 whole geographical study area is used to calibrate the individual parameters
157 of the algorithm. The same sample is used for manual review and validation
158 to evaluate how well each set of matches represents correct links between the
159 entities in two datasets. The first four steps discuss the matching algorithm while
160 the last two steps investigate how it could or could not be improved on the basis
161 of the Waste Statistics Regulation.

162 3.1. Data Sources

163 The first dataset, further referred to as "the LMA dataset", consists of digi-
164 tised waste reports filtered for all waste produced in Amsterdam Metropolitan
165 Areas (AMA) in 2018 according to the registered postcode of a waste producer.
166 The filtered dataset consists of 208,133 reports. The reports are collected with
167 regard to the EU Regulation (EC) No 2150/2002, amended by Regulation (EU)
168 No. 849/2010, which mandates Member States to produce statistics relative to
169 the generation, recovery and disposal of waste. The reports represent a chain
170 of waste management from the original waste producer all the way to the final
171 treatment destination.

172 Waste producers in the LMA dataset often have two related addresses: an
173 administrative address and a waste disposal address. Since the waste disposal
174 address does not necessarily have to be officially associated with the waste
175 disposing party (e.g., in the case of construction companies or other service
176 providers), linking is performed based on the administrative address only.
177 Finally, 8.25% of all waste reports marked as *en route* collection have been
178 excluded from the matching as these represent the same waste stream collected
179 from multiple companies and waste collector instead of the waste disposer

180 registered as a waste producer. In addition to the name and address, entities from
181 the LMA dataset have a list of EWC (European Waste Classification) codes that
182 describe which wastes they have disposed of.

183 If the effective waste disposal address is different from the entity's
184 administrative address and the regulations are followed correctly, using the
185 disposal address for linking the two datasets should point not to the entity
186 responsible for the disposal but to an entity in which premises the waste is gen-
187 erated and could be considered an indirect waste producer. If the administrative
188 address is different from a disposal address, it means that the entity effectively
189 responsible for waste generation has provided a service to the one at whose
190 premises the waste has been generated. It is, however, not obligatory to register
191 the customer who has received the provided service, therefore indirect waste
192 producers are not known and therefore not included in the waste statistics.

193 The second dataset, further referred to as "the KvK dataset", comes from the
194 Dutch Chamber of Commerce register (NL: *Kamer van Koophandel (KvK)*)
195 which is the key register for all businesses and legal entities in the Nether-
196 lands. This is a highly sensitive dataset, therefore only three fields could be
197 used for linking: entity name, address and economic activity code according
198 to the NACE Rev. 2 classification (Nomenclature statistique des ActivitÃ's
199 Ã'conomiques dans la CommunautÃ' EuropÃ'enne).

200 KvK dataset provides all registered addresses of the same legal entity and
201 multiple versions of the their names and their abbreviations. The dataset used
202 for this publication has been limited to the entities registered as active in the

203 AMA in year 2018 and resulted in 358,406 unique combinations of names and
204 addresses.

205 **3.2. Code**

206 The method is implemented in Python 3.7, with the help of the follow-
207 ing scientific software packages: Numpy 1.17, Pandas 1.0, Matplotlib 3.2,
208 Fuzzywuzzy 1.0. All data visualisations are created using Matplotlib Pyplot.
209 A geocoder is created with GeoPy, using a Mapbox service for all data points
210 in this experiment.

211 **3.3. Data Preprocessing**

212 The data preprocessing stage assures that data from all sources have the
213 same format. Filtering, cleaning and harmonisation steps are necessary to
214 identify suspicious entries, correct the obvious errors, and filter out entries that
215 cannot be fixed. The same data preprocessing is applied to both LMA and KvK
216 datasets.

217 Filtering controls if all fields of the provided addresses have a valid format,
218 e.g., street and city names are supposed to be composed of at least 3 alphabetic
219 characters and postcodes must follow a Dutch postcode pattern of 4 numerals
220 and two Latin letters.

221 Cleaning and harmonisation deals with the problem of spelling variations
222 that include partial or full abbreviations, different word order, hyphenation,
223 spacing, etc. Since LMA dataset is not based on any official registry, the same
224 entity often has its name spelled differently if a report has been submitted by
225 a different person. Spelling mistakes are also common. Subsidiary companies

226 often have slight variations between their names that indicate different services
227 and activities.

228 Geolocation (or geocoding) is the conversion of addresses into unique points
229 with geographic coordinates. This step is necessary to compute the geographic
230 proximity between the LMA entities and their potential equivalents in the
231 KvK dataset. Geocoding is prone to errors that happen if an address is not
232 complete, misspelled, corresponds to multiple points or it is simply not included
233 in the service database. To validate geocoding results and rectify the errors
234 the Dutch postcode districts (NL: *Postcodegebied*) are used. Postcode districts
235 are polygons that include all addresses within the same first four digits of a
236 postcode. If a point falls within its own postcode polygon, then the location
237 is considered valid. Otherwise, the geolocation is considered invalid and a
238 postcode polygon centroid is assigned instead of the geolocated point. This
239 rectification ensures that in case the geolocation has failed due to an incorrectly
240 spelled address, an entity is located in the proximity of its counterpart in the
241 other dataset and can still be matched based on the name similarity.

242 **3.4. Indexing**

243 The goal of indexing is to reduce the quadratic complexity by effective pair
244 candidate generation. Trying to compute the name similarity and geographic
245 proximity between each of the LMA and each of the KvK entities would
246 result in more than 4,5 billion pairwise comparisons. Besides an extensive
247 computational time, such an effort would not add significant quality to the
248 result. Increased pool of matching possibilities tends to result in less confident
249 matches and more frequent linking due to accidental similarity. Therefore, to

250 reduce the matching pool, the potential matches are evaluated only if they are
251 within a certain radius from the LMA entity location.

Figure 1: Algorithm matching success ratio dependency on the search radius distance.

252 A series of empirical tests using the data sample have been performed to
253 choose an optimum search radius. Figure 1 shows that the ratio of successful
254 matches peaks at 500m and steadily decreases with the further increase of
255 the radius. This phenomenon is caused by the further explained probabilistic
256 linkage method due to which a higher number of probable matches reduces
257 the overall matching confidence, throwing a larger number of matches to be
258 discarded as not confident enough. It must be noted that the ratio of successful
259 matches does not indicate the ratio of correct matches. However, by manually
260 comparing the differences between matches at 500m buffer radius and 5000m
261 buffer radius, it could be noticed that both correct and incorrect matches get

$$d_{Levenshtein(x,y)}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} lev_{x,y}(i-1,j) + 1 \\ lev_{x,y}(i,j-1) + 1 \\ lev_{x,y}(i-1,j-1) + 1_{x_i \neq y_j} \end{cases} & \text{otherwise} \end{cases} \quad (1)$$

where $1_{x_i \neq y_j}$ indicator refers to 0 when $x_i = y_j$ and refers to 1 otherwise. It is compared between the first i characters of x and the first j characters of y

discarded due to reduced confidence. Moreover, a 500m radius provides a good balance between urban and rural areas where the distances between different entities tend to range from a few meters to a few hundred meters.

3.5. Pairwise comparison

The similarity level of record pairs within a search space is determined according to two criteria: name similarity and geospatial proximity. The two criteria are not combined into a single indicator but used to complement each other while deciding the confidence of a potential match.

Name similarity is computed using the Levenshtein Distance (Levenshtein, 1966). It is one of the oldest metrics that indicates how two sequences of words resemble each other. Levenshtein distance is described as the minimum number of edits (insertions, deletions, or substitutions) required to mutate one string into the other. It is evaluated using a dynamic programming algorithm. The bigger the Levenshtein distance between two strings, the more distinct those strings are. Equation 1 calculates the Levenshtein distance between two strings x and y .

Levenshtein Distance is used to calculate the Levenshtein Similarity Ratio. Using the ratio allows normalising the distance against the length of the string,

$$r_{Levenshtein(x,y)}(i,j) = (|x| + |y|) - d_{Levenshtein(x,y)}(i,j) / |x| + |y| \quad (2)$$

where $|x|$ and $|y|$ are the lengths of sequence x and sequence y respectively

so that the number does not fluctuate given inputs with different sizes. The ratio can be computed using Equation 2.

Levenshtein Distance has been chosen against other name similarity metrics due to its ability to compare strings of different length and indicate if one string is contained by the other (especially relevant in cases where one registry includes only the trademark and the other one specifies it in more detail, e.g. *Boskalis* vs. *Boskalis Amsterdam*). It is also able to return a high similarity value in case of spelling mistakes and typos, and distinguish between anagrams.

Geographic proximity is calculated as a Euclidean distance between two points expressed in a local coordinate system based on metric units. It serves two purposes:

1. When the name similarity indicator cannot effectively distinguish between multiple probable matches, geographically closer match is considered more probable to be the correct one;
2. In those cases where name similarity is not sufficient to match with any of the potential counterparts in the other dataset, geospatial proximity allows assigning economic activity based on the economic activities present in its immediate surroundings.

3.6. Classification

A probabilistic entity linkage method is developed using a waterfall approach for generating matched subsets of data, where the subsets are defined

301 by gradually looser match identification criteria (Cohen et al., 2018). A series
302 of tests are applied in a specific order to evaluate if a potential link satisfies
303 the criteria. If an entity passes the test, it is removed from the pool and does
304 not need to go through the following tests. While each successive set of criteria
305 produces a larger number of potential links, the overall confidence level of those
306 links is lower. Before moving to the next, more loose criterion, the algorithm
307 removes those entities that have already satisfied the previous criteria. This
308 process continues until all entities are linked or until further loosening of the
309 criteria results in a linkage of unacceptable quality as illustrated in Figure 2.

310 There are five main tests (plus the remaining category) and two nested tests
311 that split the matches into twelve subsets with decreasing confidence. If neither
312 of the two nested tests are passed, then the match is considered insufficiently
313 confident for the set in question and is passed to the next test. During a manual
314 inspection of the potential matches for the 1000 sample data points, it has
315 been noticed, that the first accidental matches within the search space start
316 occurring below the name similarity ratio of 85 and geospatial distance of 5m
317 . An overview of the subsets can be seen in Table 1. If an entity does not pass
318 any of the tests, it is considered unmatched and gets assigned to an "Unknown
319 Economic Activity".

320 **3.7. Manual Review**

321 To validate how well each set of matches represents correct links between
322 the entities in two datasets, a manual inspection is carried out on the sample of
323 1000 LMA entities. Linkage quality is indicated by manually assigning one of
324 the 5 tags to each match as can be seen in Table 2. Inspection is performed based

The Responsibility of Waste Production

Figure 2: A series of tests applied as a waterfall approach on each LMA entity and its potential counterparts in the KvK dataset. The algorithm results into 6 sets of matches with two subsets each, where each subsequent subset has lower matching confidence

325 on the similarity between an actor name and the linked company name and
 326 the correspondence between the name and the assigned economic activity. No
 327 additional search using other data sources is performed. The manual inspection
 328 serves not only the evaluation of the algorithm accuracy, it also provides
 329 insights behind the unsuccessful matches.

Table 1

An overview of the confidence subsets and their criteria.

Set	Description
Set 1. Same name and postcode	All LMA entities are compared to all KvK entities on the basis of the exact same name and postcode (exact string matching). This test is applied before reducing the search space.
Set 2. Same name	Applies in those cases when an entity name is not misspelled but the address does not match any of the officially registered ones. This test is also applied before reducing the search space.
Set 3. Similar names and locations	A search space is created for each of the LMA entities to reduce the computational runtime. Then the name similarity and geospatial proximity indicators are computed and the threshold is set to 85 fro the name similarity and 5m for the geospatial proximity.
Set 4. Similar names	Only name similarity above 85 is considered.
Set 5. Similar locations	Only geospatial proximity below 5m is considered.
Set 6. Context-based probability	The remaining matches are checked for the two nested tests as described below.
Subset	a. Unanimous NACE code b. Most similar names and locations
	If an LMA entity matches multiple KvK entities according to the test criteria, however, all of them are registered under the same NACE code.
	If the most similar name belongs to the geographically closest KvK entity.

3.8. Evaluation

The algorithm finds a correct match in at least 68% of all cases (or at least 84% of all matched cases). The remaining 32% fail to find their counterparts in the KvK dataset due to various reasons. Upon the manual investigation of selected failed linkages from all subsets and interviews with the LMA data providers, three main reasons of failure could be distinguished:

Table 2

Results and criteria of the manual inspection performed on the sample 1000 entities

Tag	Description	Example	% of the sample
2	Probably correct NACE code	<i>Eetcafe 't Weesperplein</i> and <i>Cafe diner 't Weesperplein</i>	60%
1	Likely correct NACE code	<i>Optisport Almere B.V.</i> and <i>Sportstudio Buiten</i>	3%
0	Impossible to say if it is correct or incorrect	<i>VvE Tuin van Houten</i> and <i>Bold Innovations B.V.</i>	9%
-1	Likely incorrect NACE code	<i>Titania Asset Advise B.V.</i> and <i>Frank a Do</i>	4%
-2	Probably incorrect NACE code	<i>VISCON GLAS</i> and <i>Ferid's Grill V.O.F.</i>	5%
na	Unmatched		19%

336 **1. Failed geolocation.** If address geolocation in one of the two datasets
337 results in a point that is not within 500m of the actual address and the
338 entity name is not identical in both datasets, the entities will not get
339 matched. However, upon inspection of the two datasets, it appears that
340 only 1.3% of points that represent the same postcode lie more than 500m
341 apart from each other.

342 **2. Heavily misspelled or an alternative name.** Besides the cases of heavily
343 misspelled entity names, sometimes an alternative name or an old com-
344 pany name is used that is not similar to the one registered at the Chamber
345 of Commerce. E.g., *Hotel Campenile* can be found under the name
346 of *Hotel Gaasperpark B.V.* or *Milieustraat Almere* vs. *Recyclingperron*
347 *Almere Poort*. Assuming, that the entity address is still correct and got
348 correctly geolocated, this error should not account for more than 20% of

the unmatched cases. This estimation is based on the number of matches that cannot be validated neither as correct nor as incorrect within the 5a and 5b subsets where matched entities are within a 5m radius.

3. Inconsistent address registration. Upon manual investigation of the unmatched entities, it occurs that often LMA dataset refers to an address which in fact is not the address which is registered at the KvK dataset. These are often operational instead of administrative addresses, which means that a great amount of confusion exists regarding which address is legally considered the company's administrative address. This error should account for the remaining 79% of the unmatched or incorrectly matched cases.

There are no observable geographical patterns between the matched, wrongly matched, and unmatched entities as all groups appear to be equally distributed throughout the whole study area.

3.9. Refinement

Given that the majority of failed matches are caused by the inconsistencies of address registrations in different databases, a part of the remaining unmatched entities could still be linked to a correct economic activity by performing a name similarity check in a significantly wider geographical radius. The challenge with this approach lies in the high probability of accidental matches. If an entity's registration address cannot be trusted and used for matching, the name becomes the only information field that can be used. However, it is obvious that the name must have spelling differences with its counterpart in the KvK dataset, otherwise it would have fallen in one of the

373 first two subsets of the matching algorithm. On the other hand, in case the match
374 has failed due to a heavily misspelled or an alternative name but the address is
375 correct, the economic activity could be assigned by taking into account entities
376 in the immediate context as in set 6. However, in many cases there are multiple
377 activities happening at the same place.

378 Along with the entity's name and location, it is known which type of waste
379 has been disposed of. It can be expected that certain types of waste can be
380 produced only by certain types of economic activities. And certain economic
381 activities (e.g., IT or financial services) are not supposed to produce any
382 other than office waste. Therefore, the further described experiment explores
383 if this field could be used to limit the search space and improve the matching
384 confidence.

385 According to the EC Regulation No 2150/2002 on Waste Statistics:

386 "The principal activity of a statistical unit (e.g., an enterprise) is
387 defined as the one that contributes most to its value added. <...>
388 Therefore, in order to assign the generated waste to the correct
389 NACE activity, the unit to be considered should be the unit that
390 actually generates the value added and that also causes the waste
391 rather than the unit of the customer. For instance, waste arising
392 from the construction of a building should be assigned to the
393 activities of the construction company itself (NACE F) rather than
394 to the activity of the future building owner (e.g. services). As
395 already mentioned, the waste should be attributed to the sector

396 which generates it and hands it over to the waste management sector
397 or takes it directly to a dump or treatment site."

398 "Guidance on classification of waste according to EWC-Stat categories:
399 Supplement to the Manual for the Implementation of the Regulation (EC) No
400 2150/2002 on Waste Statistics" provides guidance on classification of waste
401 according to EWC-Stat categories (Eurostat, 2010). The document provides
402 all waste categories, their definitions and the NACE rev. 2 code which refers
403 to the most probable economic activities to produce the described waste. The
404 correspondence table that has been used for the experiments has been taken
405 from a a conversion table using 20 WStatR (Waste Statistics Regulation)
406 items. The document provides: 1) correspondence between WStatR items and
407 respectively, EWC codes that contain those items, and 2) correspondence
408 between WStatR items and respectively NACE codes that may produce waste
409 containing them. This means that in some cases also other economic activities
410 could be the source of the respective waste.

411 An experiment has been performed to estimate if the correspondence table
412 between NACE and EWC codes derived from the supplement could be used to
413 prefilter unlikely economic activities from the KvK dataset. From the sample
414 of 1000 entities that got linked to economic activities using the KvK dataset,
415 681 have been manually confirmed as correct. The waste content of these 681
416 actors and their linked economic activities have been tested for their presence in
417 the NACE-EWC correspondence table. Since some actors dispose of more than
418 one kind of waste, in total, there are 1186 unique EWC-NACE combinations to

Table 3

Comparison of NACE-EWC combinations obtained from the manually validated part of LMA data and "Guidance on classification of waste according to EWC-Stat categories: Supplement to the Manual for the Implementation of the Regulation (EC) No 2150/2002 on Waste Statistics"

	Level Section	2-digit NACE	4-digit NACE
Total unique combinations in the manually validated part of LMA data	653	900	1186
% combinations not mentioned in the guidelines	43.49%	60.15%	77.10%
Number of entities whose...			
all EWC codes are not mentioned	246	321	445
at least one EWC code is mentioned	99	98	67
all EWC codes are mentioned	269	195	102
...in the respective NACE section of the guidelines			

be compared with the guidance document. The test results can be seen in Table 3.

Comparison has been performed on three different levels:

1. NACE sections consisting of headings identified by an alphabetical code,
2. NACE divisions consisting of headings identified by a two-digit numerical code,
3. NACE classes consisting of headings identified by a four-digit numerical code.

As it can be seen from Table 3, the guidelines cover only a quarter of NACE-EWC combinations at the most detailed level that are available in the manually validated part of the LMA data. Even at the section level, 2246 out of 681 tested entities do not belong to the NACE codes that are mentioned as possible sources of disposed waste. These results suggest that using the correspondence between

432 the NACE-EWC codes as described by the guidelines would not improve and
433 rather inhibit the current matching algorithm.

434 While this experiment does not lead to an improved linking between the
435 two datasets it does reveal discrepancies between the waste registration data
436 and the official guidelines, therefore the same experiment is further repeated
437 and analysed on high confidence matches within the full dataset.

438 **4. Results**

439 There have been 12,655 entities with a valid name and address identified as
440 primary waste producers within AMA in 2018 according to the LMA dataset.
441 These entities have been linked to the legal entities in the KvK dataset using
442 the above described algorithm.

443 Match distribution within the confidence subsets is very similar to that of
444 the random sample as can be seen in Table 4. A total of 5403 actors (42.7%)
445 have been matched with high confidence, 4630 actors have been matched with
446 low confidence (36.58%) and 2622 actors (20.72%) remain unmatched.

447 Entities that have been matched with high confidence have been tested
448 for their correspondence to the Waste Statistics Regulation as explained in
449 subsection 3.9 Refinement. On the NACE section level, the high confidence
450 matches have resulted in 1920 unique combinations of NACE sections and 6-
451 digit EWC codes. Out of them 46,3% of the combinations do not appear in
452 the guidelines. This means that over half of the entities matched with high
453 confidence have disposed of and reported waste that is not considered typical
454 to their primary economic activity. The full overview can be seen in Figure 3.

Table 4

Entity assignment to different confidence subsets according to the probability that the link to a NACE code is correct. Comparison between full and sample datasets.

Subset		Confidence	Actors in sample dataset	%	Actors in full dataset	%
Same name and post-code	1a	Highest	332	33.2	3967	31.35
Same name	2a	Highest	67	6.7	869	6.87
	2b	Highest	9	0,9	72	0.57
Similar name and location	3a	High	35	3.5	354	2.8
	3b	High	0	0	3	0.02
Similar name	4a	High	5	0,5	132	1.04
	4b	High	0	0	6	0.05
Similar location	5a	Low	231	23.1	2756	21.78
	5b	Low	112	11.2	1654	13.07
Most similar name is the closest	6a	Lowest	2	0.2	24	0.19
	6b	Lowest	15	1.5	196	1.55
Unmatched			192	19.2	2622	20.72
Total			1000		12655	

Figure 3 reveals that the most common guideline non-compliances occur in the EWC chapter 17: Construction and Demolition Wastes and chapter 20: Municipal Wastes. Regarding the NACE sections, most non compliant combinations occur within section G: Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles, section N: Administrative and Support Service Activities and section O: Public Administration and Defence, Compulsory Social Security. These insights suggest that the guideline which asks to allocate waste generation to the company that contributes most to the economic value at the time of waste generation is the one that is most often violated. Following this guideline, sections N and O, which mostly contain

The Responsibility of Waste Production

Figure 3: Number of entities per each combination of NACE and EWC sections that have reported waste under EWC code which is a) not considered typical to their NACE section; b) is considered typical to their NACE section

administrative services should not be generating any other than office waste of insignificant quantities. Meanwhile, construction and demolition waste should be generated by the extractive industries (A, B and C), waste management activities (E) or construction and demolition services (F) only.

The high number of noncompliant NACE-EWC combinations in EWC sections 17 and 20 is consistent with the overall higher number of actors that dispose of this type of waste as can be seen in Figure 4. However, actors that remain unmatched and or are matched with low confidence are proportionally slightly more common among those disposing of construction and demolition, and municipal waste than other types of waste. Otherwise, the proportional distribution between the different confidence groups and the unmatched actors stays very similar between all EWC sections.

The Responsibility of Waste Production

Figure 4: Number of entities per each EWC section (2-digit code) and their matching subsets

5. Discussion

The presented algorithm is able to reliably determine the primary economic activity of less than half of the registered waste producers in the Amsterdam Metropolitan Area. The suboptimal performance of the algorithm can be attributed to the quality of the used datasets. First of all, the waste registration dataset does not use a common identifier system to recognise the same legal entities or locations among the waste producers which permits multiple name and address spelling variations. Secondly, the trade registry does not contain all

485 operational addresses and alternative company names that are used in the waste
486 registry. Therefore computational methods to match the two datasets have only
487 limited capabilities to mitigate the poor data quality.

488 The most straightforward recommendation to improve the data quality is
489 for the waste registry to request companies filing waste reports to provide
490 their unique identifier used in the trade registry (in case of the Netherlands -
491 KvK number). Using the unique identifiers waste producers could be on-the-fly
492 connected to their economic activities provided in the trade registry. However,
493 this approach still leaves a few potential caveats. First, the approach would not
494 help processing the historical data that has existed before the implementation of
495 the on-the-fly entity matching. Second, the question remains if the self-assigned
496 economic activities provided by the companies at the time of registering their
497 business are the ones that are effectively responsible for producing certain kinds
498 of waste.

499 The described algorithm and experiments have demonstrated that an entity
500 matching algorithm is limited by the lack of corresponding entities between
501 the two datasets. However, comparisons between matched and unmatched
502 entities in terms of reported waste types, their geographical distribution, and
503 sample versus full dataset did not expose any differences that would suggest
504 the unmatched actors would show different statistical patterns. Therefore, it is
505 safe to assume that given that the algorithm performs well on approximately
506 half of the dataset, the successfully matched half could be used as a substantial
507 sample to scale the statistics to the full waste quantity.

508 However, if the goal of collecting detailed waste statistics on a EU level is
509 to improve policies to ensure that the transition to a circular economy can be
510 accelerated, the lack of reliable highly granular current or historical information
511 may have negative consequences. On the one hand, it may hinder the visibility
512 of emerging small-scale good practices which need to be further fostered to
513 ensure their adoption in the wider economy. On the other hand, the less detailed
514 information is available, the harder it is to notice the effects of changing demand
515 and production processes of one economic sector upon waste generation in
516 another (Salemdeeb et al., 2016). For example, it is expected that the shift from a
517 product-based to a service-based economy will increase resource productivity
518 and reduce waste production. To monitor whether this is effectively the case
519 it would be necessary to know not only that certain wastes are generated by
520 service providers instead of manufacturers but which types of services they
521 are and which economic activities they are expected to substitute. Yet the
522 experiments have revealed that the more actors are considered, the more unique
523 combinations of NACE and EWC codes can be found. Therefore, using a
524 representative sample would only help monitoring major shifts after they have
525 already occurred and not their state of emergence.

526 The second risk of a statistical blind spot that got exposed during the exper-
527 iments is related to the question which entity needs to be effectively considered
528 as a responsible waste producer. While the Waste Statistics regulation clearly
529 states that it must be the "unit that actually generates the value added and that
530 also causes the waste rather than the unit of the customer", the waste registry
531 shows that this rule is often disregarded and the registered waste producer is that

The Responsibility of Waste Production

entity which eventually pays the waste management costs. E.g. Construction and demolition waste is often reported by the companies whose core business is not related to construction, food-related waste is reported by the companies who provide catering for their own employees only and various wastes are reported by the enterprise subsidiaries whose main activities are providing financial administration. Instead of pointing out multiple violations of the regulation, these insights rather question the regulation itself.

The requirement to attribute waste to the unit which generates the most financial value added places the burden of the waste production on those companies whose business model is directly related to the amount of waste produced, meaning that more produced waste should directly correlate with increased revenues. Using restrictive policies to stimulate those companies to reduce their waste production might cause undesired backlash. To enable the strategy where companies are encouraged to change their business models in a way that used resources are not discarded but kept in the economy, the consumers of their products or services need to be stimulated to choose for a more sustainable alternative. In that case the information about the customer is as important as the information about the provider.

Another important consideration is waste ownership. In the economy where waste is considered a burden, the company that causes its production tends to take the responsibility. However, in a circular economy where redundant materials are considered an asset instead of waste and therefore may have financial value, it is more likely that the ownership will stay with the company that has paid for the materials before they have become redundant. Moreover,

sectors with long supply chains tend to generate more indirect waste which gets distributed over a number of different supporting economic activities.

Finally, it must be noted that no official correspondence table exists that relates NACE and EWC codes, meaning that there is no guidance on which types of waste should be expected from which type of companies. Having WStatR items as a significantly less detailed intermediary layer between the two detailed classification systems causes unlikely combinations to be valid (e.g. Glass waste connects EWC code "15 01 07 glass packaging" with NACE code "4310 Demolition and site preparation"). At the same time, it excludes a large number of possible combinations for the sake of clarity. When data across the EU is collected using sample surveys and combinations of registries, a high-quality correspondence between NACE and EWC codes would not only help to control the quality of the statistics but also improve data consistency and consequently knowledge transfer regarding the circular economy transition.

6. Conclusions and Recommendations

A Dutch national waste registry dataset for the year of 2018 limited to the Amsterdam Metropolitan Area has been used to explore which economic activities effectively produce which types of waste and if these activities can be held responsible for waste production. The Dutch trade registry has been used as a reference dataset which connects company names and addresses to their primary economic activities according to the NACE codes. The conducted experiments have demonstrated that using geospatial proximity in combination with name similarity is able to speed up the linking process in comparison to

579 only using textual information on the entity's name and address. In addition,
580 considering geospatial proximity next to the traditional approach of name
581 similarity for legal entity matching is able to limit the search space for each
582 individual entity and therefore solve multiple data quality issues.

583 The manual validation of a random sample of 1000 waste producers has
584 shown that the algorithm is able to correctly assign the primary economic
585 activity to approximately 43% of all waste producers. Additional 37% of
586 the actors are assigned an economic activity with low confidence, while the
587 remaining 20% cannot be assigned at all. The reasons behind the suboptimal
588 assignment success rate stem mostly from the quality of underlying datasets
589 and their limitations. Waste registry dataset contains multiple entities with
590 misspelled names and addresses, and addresses that do not point to the actual
591 addresses of registration in the KvK trade registry. At the same time, the KvK
592 dataset provides outdated or incomplete data points where not all relevant
593 addresses are present.

594 The attempt to refine the algorithm on the basis of the type of waste an
595 entity disposes of has proven unsuccessful. Comparing the available guidance
596 and conversion tables to the NACE-EWC combinations obtained from the high
597 confidence linking subsets has revealed that even at the least detailed level
598 of economic activity classification, roughly half of all actors do not comply
599 with the regulation guidelines. The lack of compliance can be explained by
600 the unrealistic expectations of the guidelines set by a lack of a high-quality
601 correspondence table and a non-operational definition of the waste producer.

602 No statistical differences could be observed between the matched and un-
603 matched parts of the entities, therefore the waste production statistics obtained
604 from the matched part could be scaled to the full dataset to show which
605 economic sectors have produced which quantities and types of waste. However,
606 this method is not able to provide a more detailed representation on an economic
607 activity instead of sector level.

608 Based on the described experiment, the following recommendations can be
609 made regarding the Waste Statistics Regulation and further related research.
610 First, a guidance document that provides high-level-of-detail correspondence
611 between EWC and NACE codes would provide a control mechanism for the
612 consistency of the reported statistics, especially given that every member state
613 applies different data collection and reporting methods. The definition of a
614 waste producer should be chosen in light of which statistics are necessary to
615 support the transition towards a circular economy and not which entity needs
616 to be charged for the waste management costs as typical to the linear economy.
617 Additionally, waste statistics could collect not only data related to the waste
618 disposing entity but also to the economic activities that have preceded the
619 disposal.

620 Finally, Waste Statistics Regulation suggests using the national Trade reg-
621 istry as a reference for the economic activities of the waste producers. However,
622 this experiment has demonstrated that a Trade Registry in the Netherlands does
623 not sufficiently correspond to the operational reality in terms of company data.
624 Moreover, the primary economic activities assigned at the Trade registry might
625 not be the ones that actually cause waste generation. Future research should

626 include other Member States and their waste data collection methods to ensure
627 that the Waste Statistics Regulation is able to support the required variety of
628 geographical contexts and compile supra-national datasets necessary to support
629 a circular economy transition.

630 **Data and Code Availability**

631 Datasets used for this publication have been obtained under two Hori-
632 zon2020 projects: REPAiR¹ (Resource Management in Peri-Urban Areas) and
633 CINDERELA² (New Circular Economy Business Model for More Sustainable
634 Urban Construction). The data that support the findings of this study are avail-
635 able from the Waste Registry Division of the Dutch Ministry of Infrastructure
636 and Public Works (NL: *Landelijk Meldpunt Afvalstoffen (LMA)* in the Nether-
637 lands. Restrictions apply to the availability of these data, which were used under
638 license for this study. Data are not available from the authors and can only be
639 accessed directly from the Ministry. All code developed for this experiment is
640 available open source as part of the GitHub repository here: [https://github.](https://github.com/rusne/lma-data-pipeline/tree/master/nace-ewc)
641 [com/rusne/lma-data-pipeline/tree/master/nace-ewc](https://github.com/rusne/lma-data-pipeline/tree/master/nace-ewc)

642 **Acknowledgements**

643 The author would like to thank the the Waste Registry Division of the Dutch
644 Ministry of Infrastructure and Public Works for sharing the data for the research
645 purposes, especially Tjerk ter Veen for explaining all data collection subtleties.

¹<http://h2020repair.eu/>

²<https://www.cinderela.eu/>

Furthermore thanks to the geoFluxus software development team who have invested their time and resources into this research as well.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 688920 and No 776751.

CRediT authorship contribution statement

Rusne Sileryte: All study elements. **Alexander Wandl:** Study supervision. **Arjan van Timmeren:** Study supervision.

References

- Antoni, M., Koller, K., Laible, M.C., Zimmermann, F., 2018. Orbis-ADIAB: From record linkage key to research dataset: Combining commercial company data with administrative employer-employee data. Technical Report. Research Data Centre (FDZ) of the German Federal Employment Agency (BA) as the Institute for Employment Research. URL: https://econpapers.repec.org/RePEc:iab:iabfme:201804_en, doi:10.5164/IAB.FDZM.1804.en.v1.
- Brunner, P.H., Rechberger, H., 2016. Practical handbook of material flow analysis. volume 1. CRC press.
- Burdick, D., Hernandez, M., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L.C., Stanoi, I., Vaithyanathan, S., Das, S.R., 2015. Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study. SSRN Electronic Journal , 1–8doi:10.2139/ssrn.2666384.
- Choi, S.C.T., Lin, Y., Mulrow, E., 2017. Comparison of public-domain software and services for probabilistic record linkage and address standardization. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 10344 LNAI, 51–66. doi:10.1007/978-3-319-69775-8_{_}3.
- Cohen, G., Friedrichs, M., Gupta, K., Hayes, W., Lee, S.J., Marsh, B., Mislant, N., Shaton, M.O., Sicilian, M., 2018. The U.S. Syndicated Loan Market: Matching Data. SSRN Electronic Journal doi:10.2139/ssrn.3297815.
- Deloitte, 2017. Resource Efficient Use of Mixed Wastes - Fact sheet Denmark - Construction and Demolition Waste Management in Denmark. September.
- European Commission, 2020. A new Circular Economy Action Plan For a cleaner and more competitive Europe. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1583933814386&uri=COM:2020:98:FIN>.

The Responsibility of Waste Production

- 675 Eurostat, 2010. Guidance on classification of waste according to EWC-Stat categories.
- 676 Eurostat, 2013. Manual on waste statistics: A handbook for data collection on waste generation and treatment. URL:
677 <http://bookshop.europa.eu>.
- 678 Gao, C., Gao, C., Song, K., Fang, K., 2020. Pathways towards regional circular economy evaluated using material
679 flow analysis and system dynamics. Resources, Conservation and Recycling 154, 104527. URL: <https://doi.org/10.1016/j.resconrec.2019.104527>, doi:10.1016/j.resconrec.2019.104527.
- 680
- 681 Geyer, R., Kuczenski, B., Zink, T., Henderson, A., 2016. Common Misconceptions about Recycling. Journal of
682 Industrial Ecology 20, 1010–1017. doi:10.1111/jiec.12355.
- 683 Guermazi, Y., Sellami, S., Boucelma, O., 2020. Address Validation in Transportation and Logistics: A Machine
684 Learning Based Entity Matching Approach. volume 1323. Springer International Publishing. URL: http://dx.doi.org/10.1007/978-3-030-65965-3_21, doi:10.1007/978-3-030-65965-3_{_}21.
- 685
- 686 Hanssen, O.J., Stenmarck, A., Dekhtyar, P., OâĂŽConnor, C., Ostergren, K., 2013. Review of EUROSTATs reporting
687 method and statistics. Fredrikstad, Norway: FUSIONS project .
- 688 Hirsch, J.A., Moore, K.A., Cahill, J., Quinn, J., Zhao, Y., Bayer, F.J., Rundle, A., Lovasi, G.S., 2020. Business Data
689 Categorization and Refinement for Application in Longitudinal Neighborhood Health Research: a Methodology.
690 Journal of Urban Health doi:10.1007/s11524-020-00482-2.
- 691 Köpcke, H., Rahm, E., 2010. Frameworks for entity matching: A comparison. Data and Knowledge Engineering 69,
692 197–210. URL: <http://dx.doi.org/10.1016/j.datak.2009.10.003>, doi:10.1016/j.datak.2009.10.
693 003.
- 694 Levenshtein, V.I., 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics
695 Doklady 10, 707.
- 696 Magnani, M., Montesi, D., 2007. A study on company name matching for database integration. Technical Report
697 Technical Report UBLCS-07-15. The University of Bologna Department of Computer Science Research Technical
698 Reports. Bologna, Italy.
- 699 Medvedev, T., Ulanov, A., 2011. Company names matching in the large patents dataset. Technical Report. HP
700 Laboratories.
- 701 Melosi, M.V., 2004. Garbage in the Cities: Refuse reform and the Environment. University of Pittsburgh Pre.
- 702 Mendez, D.D., Duell, J., Reiser, S., Martin, D., Gradeck, R., Fabio, A., 2014. A methodology for combining
703 multiple commercial data sources to improve measurement of the food and alcohol environment: Applications
704 of geographical information systems. Geospatial Health 9, 71–96. doi:10.4081/gh.2014.7.
- 705 Morseletto, P., 2020. Targets for a circular economy. Resources, Conservation and Recycling 153, 104553. URL:
706 <https://doi.org/10.1016/j.resconrec.2019.104553>, doi:10.1016/j.resconrec.2019.104553.
- 707 Nuss, P., Blengini, G.A., Haas, W., Mayer, A., Nita, V., Pennington, D., 2017. Development of a Sankey Diagram of
708 material flows in the EU economy based on Eurostat data: Monitoring of non-energy & non- food material flows
709 in the EU-28 for the EC Raw Materials Information System (RMIS). URL: <https://dx.doi.org/10.2760/>

The Responsibility of Waste Production

- 710 642511, doi:10.2760/642511.
- 711 Pilania, A., Kumaran, G.M.M., 2019. Comparative study of name matching algorithms, in: Proceedings of the 2019 6th
712 International Conference on Computing for Sustainable Global Development, INDIACom 2019, pp. 1174–1178.
- 713 Saleemdeen, R., Al-Tabbaa, A., Reynolds, C., 2016. The UK waste input-output table: Linking waste generation to the
714 UK economy. *Waste Management and Research* 34, 1089–1094. doi:10.1177/0734242X16658545.
- 715 Schild, C.J., 2016. Linking 'Orbis' Company Data with Establishment Data from the German Federal Employment
716 Agency. *SSRN Electronic Journal* doi:10.2139/ssrn.3549262.