The responsibility of waste production

Comparison of European waste statistics regulation and Dutch National Waste Registry

Sileryte, Rusne; Wandl, Alexander; van Timmeren, Arjan

DOI
10.1016/j.wasman.2022.07.022

Publication date
2022

Document Version
Final published version

Published in
Waste Management

Citation (APA)
Sileryte, R., Wandl, A., & van Timmeren, A. (2022). The responsibility of waste production: Comparison of European waste statistics regulation and Dutch National Waste Registry. *Waste Management*, *151*, 171-180. https://doi.org/10.1016/j.wasman.2022.07.022

Waste Management 151 (2022) 171–180

Contents lists available at ScienceDirect

Waste Management

journal homepage: www.elsevier.com/locate/wasman

Research Paper

The responsibility of waste production: Comparison of European waste statistics regulation and Dutch National Waste Registry

Rusne Sileryte [a,b,*], Alexander Wandl [a], Arjan van Timmeren [a,b]

[a] Faculty of Architecture and the Built Environment, Delft University of Technology, Julianalaan 134, Delft, the Netherlands
[b] Amsterdam Institute of Advanced Metropolitan Solutions, Gebouw 027W, Kattenburgerstraat 5, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Waste producers
Waste statistics regulation
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 to explore to which extent the guidelines available in the Waste Statistics Regulation correspond to the operational reality. To do so it presents a computational method to link waste producers to their economic activities using a national Trade Registry. An extensive discussion of the results provides insights and recommendations for the future guidelines of waste statistics to support circular economy transition.

1. Introduction

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

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

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

A number of reports aimed at evaluating the transition on the EU level emphasise that significant variations of data quality and the lack of harmonisation in data collection methodologies between the member states hinder effective monitoring and knowledge transfer (Hanssen et al., 2013; Deloitte, 2017; Nuss et al., 2017). At the same time the most common methodology used to provide waste statistics is to scale up data collected from a sample of companies to a whole sector. Waste treatment

* Corresponding author at: Faculty of Architecture and the Built Environment, Delft University of Technology, Julianalaan 134, Delft, the Netherlands.
E-mail address: r.sileryte@tudelft.nl (R. Sileryte).

https://doi.org/10.1016/j.wasman.2022.07.022
Received 19 August 2021; Received in revised form 22 April 2022; Accepted 17 July 2022
Available online 9 August 2022

statistics are often collected directly from the waste treatment facilities and therefore disconnected from the waste producers.

This way of reporting statistics works well within the traditional linear economy where waste is a post-factum problem and needs to be dealt with after it has already occurred. However, promoting and supporting such circular economy strategies as waste prevention, design for reuse, prolonged lifespan, etc. involve companies from the full supply chain. Governments need to create a coherent set of incentives and increase coordination among all relevant stakeholders (OECD, 2021). Therefore as much information as possible is necessary to identify the right stakeholders by understanding which economic sectors they belong to and to understand which interventions can be made to deal with which kinds of waste before it is effectively disposed of. At the moment the expected correspondence between the types of waste and economic sectors that produce them is provided by the "Manual for the Implementation of the Regulation (EC) No 2150/2002 on Waste Statistics". However, it is not known how well these expectations reflect the operational reality of waste production and disposal.

The Netherlands is one of the few member states whose waste data is not based on sample surveys but on consistent waste registration from every company that has a waste permit (Deloitte, 2017). These companies are statutorily required to register all transported waste, including its producer, waste characteristics, transport methods and final treatment. The caveat of the current system is that companies involved in the waste chain are reported only by their name and address without using unique identifiers able to link the available data with other business registries.

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

Within this scope this paper explores to which extent the guidelines available in the Waste Statistics Regulation are reflected in the available data and if they can be turned into computational rules to improve the quality of linking. A computational method is used to link waste producers to their economic activities based on name similarity and geospatial proximity. Finally, a discussion is provided on the consequences that legal and operation discrepancies on the waste producer responsibility might have on supporting CE transition.

2. Related Work

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

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

A rule-based matching approach using both entity name and address similarity can be found in such domains as the investigation of health-related behaviours dependent on living environments (Hirsch et al., 2020; Mendez et al., 2014), validating names and addresses of transportation and logistic entities (Guermazi et al., 2020), matching observations across financial datasets (Cohen et al., 2018; Burdick et al., 2015), identifying same entities in patent files (Medvedev and Ulanov, 2011; Magnani and Montesi, 2007). Most of them conclude that domain expert knowledge integration improves or would improve matching results (Pilania and Kumaran, 2019; Cohen et al., 2018; Choi et al., 2017; Antoni et al., 2018; Schild, 2016; Mendez et al., 2014; Magnani and Montesi, 2007).

This paper further builds on the existing examples of entity matching using standard computational methods to evaluate name similarity and geospatial proximity between potential matches. Therefore, the novel contribution of this work is not in the domain of entity matching but within the discussion regarding the adequacy of the European Waste Statistics Regulation to support the desired transition towards the circular economy. To date, no published study of the waste allocation to the economic sectors according to the Eurostat method could be identified. The lack of such studies is likely influenced by the high sensitivity of the relevant datasets which are typically not available for research purposes. This study is thus the first one to uncover the discrepancies between the legal and operational responsibility for waste production.

3. Methods

To explore to which extent the guidelines available in the Waste Statistics Regulation are reflected in the available data, the companies registered as waste producers are first linked to the trade registry to assign each of them to an economic sector. The computational entity linkage process follows six phases as defined by Köpcke and Rahm (2010): data preprocessing, indexing, pairwise comparisons, classification, manual review, evaluation, and refinement. A random sample of 1000 companies (8% of the full dataset) evenly distributed throughout the whole geographical study area is used to calibrate the individual parameters of the algorithm. The same sample is used for manual review and validation to evaluate how well each set of matches represents correct links between the entities in two datasets.

After the evaluation of the matching algorithm, all matches are assigned to confidence groups according to how likely the matching is to be correct. The group which has been matched with the highest confidence is then used to investigate how the lower confidence matches could or could not be improved on the basis of the Waste Statistics Regulation. Additional rules that could improve the matching results are derived from the "Manual on Waste Statistics" (Eurostat, 2013) that guides the data collection process in Member States. The importance is not so much to obtain the highest possible matching score but to understand the reasons behind the unsuccessful matches as they reveal the differences between the official guidelines and the operational reality of data collection and waste disposal.

3.1. Data Sources

The first dataset, further referred to as "the LMA dataset", consists of digitised waste reports filtered for all waste produced in Amsterdam Metropolitan Areas (AMA) in 2018 according to the registered postcode of a waste producer. The filtered dataset consists of 208,133 reports. The reports are collected with regard to the EU Regulation (EC) No 2150/2002, amended by Regulation (EU) No. 849/2010, which mandates Member States to produce statistics relative to the generation, recovery

and disposal of waste. The reports represent a chain of waste management from the original waste producer all the way to the final treatment destination.

Waste producers in the LMA dataset often have two related addresses: an administrative address and a waste disposal address. Since the waste disposal address does not necessarily have to be officially associated with the waste disposing party (e.g., in the case of construction companies or other service providers), linking is performed based on the administrative address only. Finally, 8.25% of all waste reports marked as *en route* collection have been excluded from the matching as these represent the same waste stream collected from multiple companies and waste collector instead of the waste disposer registered as a waste producer. In addition to the name and address, entities from the LMA dataset have a list of EWC (European Waste Classification) codes that describe which wastes they have disposed of.

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

The second dataset, further referred to as "the KvK dataset", comes from the Dutch Chamber of Commerce register (NL: *Kamer van Koophandel (KvK)*) which is the key register for all businesses and legal entities in the Netherlands. This is a highly sensitive dataset, therefore only three fields could be used for linking: entity name, address and economic activity code according to the NACE Rev. 2 classification (Nomenclature statistique des Activités économiques dans la Communauté Européenne).

KvK dataset provides all registered addresses of the same legal entity and multiple versions of the their names and their abbreviations. The dataset used for this publication has been limited to the entities registered as active in the AMA in year 2018 and resulted in 358,406 unique combinations of names and addresses.

3.2. Code

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

3.3. Data Preprocessing

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

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

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

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

3.4. Indexing

The goal of indexing is to reduce the quadratic complexity by effective pair candidate generation. Trying to compute the name similarity and geographic proximity between each of the LMA and each of the KvK entities would result in more than 4,5 billion pairwise comparisons. Besides an extensive computational time, such an effort would not add significant quality to the result. Increased pool of matching possibilities tends to result in less confident matches and more frequent linking due to accidental similarity. Therefore, to reduce the matching pool, the potential matches are evaluated only if they are within a certain radius from the LMA entity location.

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

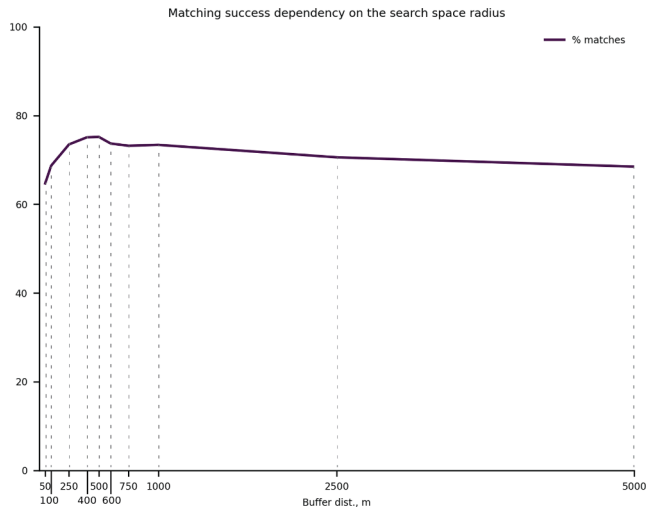

Fig. 1. Algorithm matching success ratio dependency on the search radius distance.

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

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

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

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

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

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

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

3.7. Manual Review

To validate how well each set of matches represents correct links between the entities in two datasets, a manual inspection is carried out on the sample of 1000 LMA entities. Linkage quality is indicated by manually assigning one of the 5 tags to each match as can be seen in Table 2. Inspection is performed based on the similarity between an actor name and the linked company name and the correspondence between the name and the assigned economic activity. No additional search using other data sources is performed. The manual inspection serves not only the evaluation of algorithm accuracy, it also provides insights behind the unsuccessful matches.

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:

1. **Failed geolocation**. If address geolocation in one of the two datasets results in a point that is not within 500 m of the actual address and the entity name is not identical in both datasets, the entities will not get matched. However, upon inspection of the two datasets, it appears that only 1.3% of points that represent the same postcode lie more than 500 m apart from each other.
2. **Heavily misspelled or an alternative name.** Besides the cases of heavily misspelled entity names, sometimes an alternative name or an old company name is used that is not similar to the one registered at the Chamber of Commerce. E.g., *Hotel Campenile* can be found under the name of *Hotel Gaasperpark B.V.* or *Milieustraat Almere* vs. *Recyclingperron Almere Poort*. Assuming, that the entity address is still correct and got 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 5 m 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.

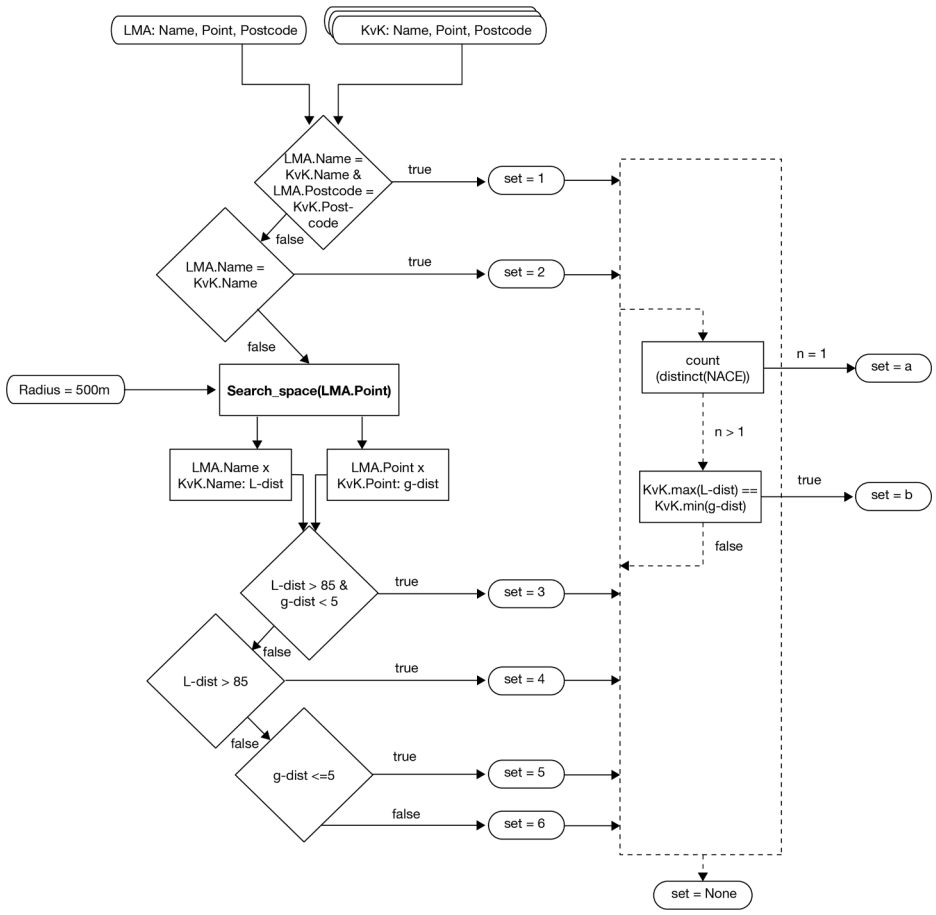

Fig. 2. A series of tests applied as a waterfall approach on each LMA entity and its potential counterparts in the KvK dataset. The algorithm results into 6 sets of matches with two subsets each, where each subsequent subset has lower matching confidence.

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 for the name similarity and 5 m for the geospatial proximity.	
Set 4. Similar names	Only name similarity above 85 is considered.	
Set 5. Similar locations	Only geospatial proximity below 5 m 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.

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

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 first two subsets of the matching algorithm. On the other hand, in case the match has failed due

to a heavily misspelled or an alternative name but the address is correct, the economic activity could be assigned by taking into account entities in the immediate context as in set 6. However, in many cases there are multiple activities happening at the same place.

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

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

> "The principal activity of a statistical unit (e.g., an enterprise) is defined as the one that contributes most to its value added. <… > Therefore, in order to assign the generated waste to the correct NACE activity, the unit to be considered should be the unit that actually generates the value added and that also causes the waste rather than the unit of the customer. For instance, waste arising from the construction of a building should be assigned to the activities of the construction company itself (NACE F) rather than to the activity of the future building owner (e.g. services). As already mentioned, the waste should be attributed to the sector which generates it and hands it over to the waste management sector or takes it directly to a dump or treatment site.".

"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" provides guidance on classification of waste according to EWC-Stat categories (Eurostat, 2010). The document provides all waste categories, their definitions and the NACE rev. 2 code which refers to the most probable economic activities to produce the described waste. The correspondence table that has been used for the experiments has been taken from a conversion table using 20 WStatR (Waste Statistics Regulation) items. The document provides: 1) correspondence between WStatR items and respectively, EWC codes that contain those items, and 2) correspondence between WStatR items and respectively NACE codes that may produce waste containing them. This means that in some cases also other economic activities could be the source of the respective waste.

An experiment has been performed to estimate if the correspondence table between NACE and EWC codes derived from the supplement could be used to prefilter unlikely economic activities from the KvK dataset. From the sample of 1000 entities that got linked to economic activities using the KvK dataset, 681 have been manually confirmed as correct. The waste content of these 681 actors and their linked economic activities have been tested for their presence in the NACE-EWC correspondence table. Since some actors dispose of more than one kind of waste, in total, there are 1186 unique EWC-NACE combinations to 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 the NACE-EWC codes as described by the guidelines would not improve and rather inhibit the current matching algorithm.

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

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			

4. Results

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

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

Entities that have been matched with high confidence have been tested for their correspondence to the Waste Statistics Regulation as explained in subSection 3.9 Refinement. On the NACE section level, the high confidence matches have resulted in 1920 unique combinations of NACE sections and 6-digit EWC codes. Out of them 46,3% of the combinations do not appear in the guidelines. This means that over half of

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 postcode	**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	

the entities matched with high confidence have disposed of and reported waste that is not considered typical to their primary economic activity. The non-typical combinations account for 42,8% of all reported waste mass from those entities. The full overview can be seen in Fig. 3.

Fig. 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 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 non-compliant NACE-EWC combinations in EWC sections 17 and 18 is consistent with the overall higher number of actors that dispose of this type of waste as can be seen in Fig. 4. However, actors that remain unmatched 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.

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

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

It must be acknowledged that if the same algorithm was applied to the same problem in a different country, its performance might be drastically different. Differences might be caused by the distinct registration terms in both trade and waste registries as each country decides on those terms individually. Lack of harmonisation between the registries causes every country to adapt a different strategy to allocate its waste to the economic sectors this way hindering cross-comparison.

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

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

The second risk of a statistical blind spot that got exposed during the experiments is related to the question which entity needs to be effectively considered as responsible for waste production. While the Waste Statistics regulation clearly states that it must be the ”unit that actually generates the value added and that also causes the waste rather than the unit of the customer”, the waste registry shows that this rule is often disregarded and the registered waste producer is that 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.

Furthermore, 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 economic 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

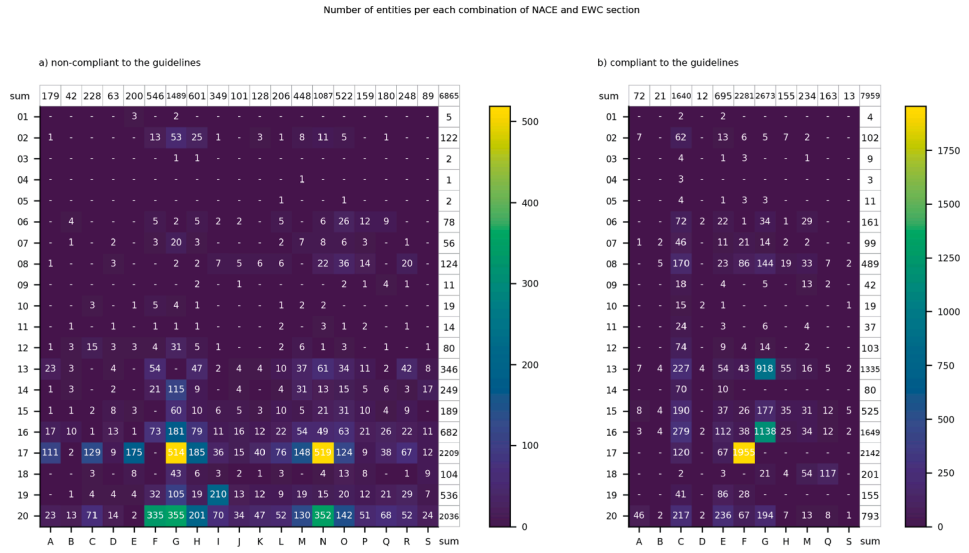

Fig. 3. Number of entities per each combination of NACE and EWC sections that have reported waste under EWC code which is a) not considered typical to their NACE section; b) is considered typical to their NACE section.

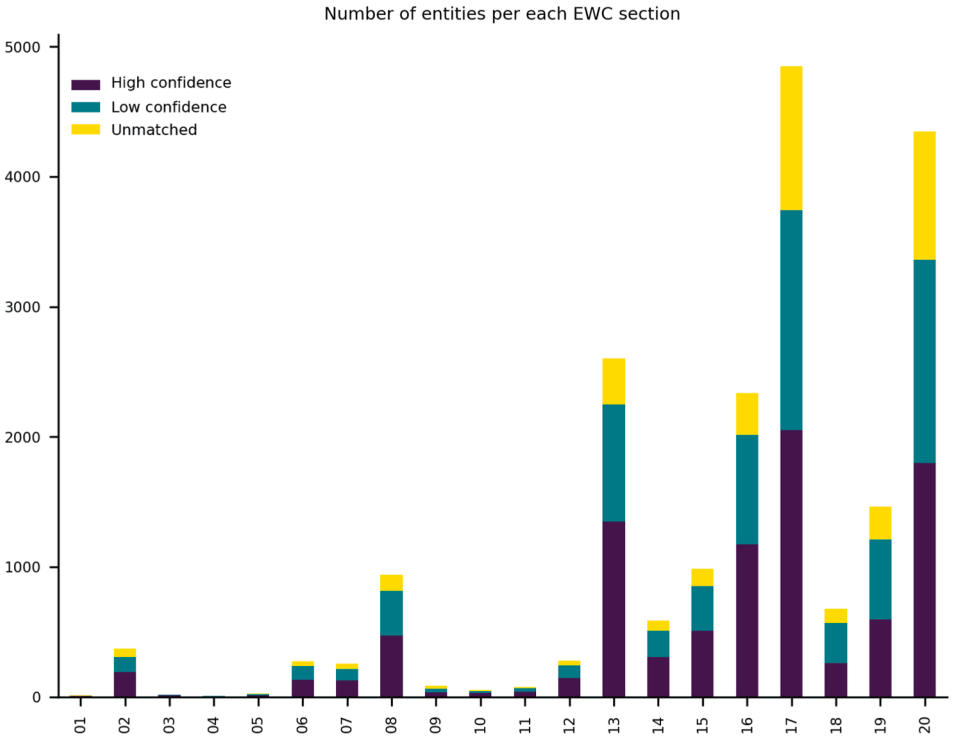

Fig. 4. Number of entities per each EWC Section (2-digit code) and their matching subsets.

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

The manual validation of a random sample of 1000 waste producers

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

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

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

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

Finally, Waste Statistics Regulation suggests using the national Trade registry as a reference for the economic activities of the waste producers. However, this experiment has demonstrated that a Trade Registry in the Netherlands does not sufficiently correspond to the operational reality in terms of company data. Moreover, the primary economic activities assigned at the Trade registry might not be the ones that actually cause waste generation. Future research should include other Member States and their waste data collection methods to ensure that the Waste Statistics Regulation is able to support the required variety of geographical contexts and compile supra-national datasets necessary to support a circular economy transition.

7. Data and Code Availability

Datasets used for this publication have been obtained under two Horizon2020 projects: REPAiR[1] (Resource Management in Peri-Urban Areas) and CINDERELA[2] (New Circular Economy Business Model for More Sustainable Urban Construction). The data that support the findings of this study are available from the Waste Registry Division of the Dutch Ministry of Infrastructure and Public Works (NL: *Landelijk Meldpunt Afvalstoffen (LMA)* in the Netherlands. Restrictions apply to the availability of these data, which were used under license for this study. Data are not available from the authors and can only be accessed directly from the Ministry. All code developed for this experiment is available open source as part of the GitHub repository here: https://github.com/rusne/lma-data-pipeline/tree/master/nace-ewc.

[1] http://h2020repair.eu/

[2] https://www.cinderela.eu/

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The author would like to thank the Waste Registry Division of the Dutch Ministry of Infrastructure and Public Works for sharing the data for the research purposes,especially Tjerk ter Veen for explaining all data collection subtleties. Furthermore thanks to the geoFluxus software development team who have invested their time and resources into this research, too.

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

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 Electron. J. 1–8. https://doi.org/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., Mislang, 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.

Eurostat, 2010. Guidance on classification of waste according to EWC-Stat categories.

Eurostat, 2013. Manual on waste statistics: A handbook for data collection on waste generation and treatment. URL: http://bookshop.europa.eu.

Gao, C., Gao, C., Song, K., Fang, K., 2020. Pathways towards regional circular economy evaluated using material flow analysis and system dynamics. Resour. Conserv. Recycl. 154, 104527. https://doi.org/10.1016/j.resconrec.2019.104527.

Geyer, R., Kuczenski, B., Zink, T., Henderson, A., 2016. Common Misconceptions about Recycling. J. Ind. Ecol. 20, 1010–1017. https://doi.org/10.1111/jiec.12355.

Guermazi, Y., Sellami, S., Boucelma, O., 2020. Address Validation in Transportation and Logistics: A Machine Learning Based Entity Matching Approach. volume 1323. Springer International Publishing. URL: https://doi.org/10.1007/978-3-030-65965-3_21, doi:10.1007/978-3-030-65965-3_21.

Hanssen, O.J., Stenmarck, A., Dekhtyar, P., O'Connor, C., Ostergren, K., 2013. Review of EUROSTATs reporting method and statistics. Fredrikstad, Norway: FUSIONS project.

Hirsch, J.A., Moore, K.A., Cahill, J., Quinn, J., Zhao, Y., Bayer, F.J., Rundle, A., Lovasi, G. S., 2020. Business Data Categorization and Refinement for Application in Longitudinal Neighborhood Health Research: a Methodology. Journal of Urban Health. https://doi.org/10.1007/s11524-020-00482-2.

Köpcke, H., Rahm, E., 2010. Frameworks for entity matching: A comparison. Data and Knowledge Engineering 69, 197–210. https://doi.org/10.1016/j.datak.2009.10.003.

Levenshtein, V.I., 1966. Binary Codes Capable of Correcting Deletions. Insertions and Reversals. Soviet Physics Doklady 10, 707.

Magnani, M., Montesi, D., 2007. A study on company name matching for database integration. Technical Report Technical Report UBLCS-07-15. The University of Bologna Department of Computer Science Research Technical Reports. Bologna, Italy.

Medvedev, T., Ulanov, A., 2011. Company names matching in the large patents dataset. Technical Report. HP Laboratories.

Melosi, M.V., 2004. Garbage in the Cities: Refuse reform and the Environment. University of Pittsburgh Pre.

Mendez, D.D., Duell, J., Reiser, S., Martin, D., Gradeck, R., Fabio, A., 2014. A methodology for combining multiple commercial data sources to improve measurement of the food and alcohol environment: Applications of geographical information systems. Geospatial Health 9, 71–96. https://doi.org/10.4081/gh.2014.7.

Morseletto, P., 2020. Targets for a circular economy. Resour. Conserv. Recycl. 153, 104553. https://doi.org/10.1016/j.resconrec.2019.104553.

Nuss, P., Blengini, G.A., Haas, W., Mayer, A., Nita, V., Pennington, D., 2017. Development of a Sankey Diagram of material flows in the EU economy based on Eurostat data: Monitoring of non-energy & non- food material flows in the EU-28 for the EC Raw Materials Information System (RMIS). URL: doi: 10.2760/642511, doi: 10.2760/642511.

OECD, 2021. Towards a more resource-e?cient and circular economy. OECD Publishing, Paris, 1 – 53.

Pilania, A., Kumaran, G.M.M., 2019. Comparative study of name matching algorithms, in: Proceedings of the 2019 6th International Conference on Computing for Sustainable Global Development, INDIACom 2019, pp. 1174–1178.

Salemdeeb, R., Al-Tabbaa, A., Reynolds, C., 2016. The UK waste input-output table: Linking waste generation to the UK economy. Waste Manage. Res. 34, 1089–1094. https://doi.org/10.1177/0734242X16658545.

Schild, C.J., 2016. Linking 'Orbis' Company Data with Establishment Data from the German Federal Employment Agency. SSRN Electron. J. https://doi.org/10.2139/ssrn.3549262.