Waste Management 39 (2015) 15–25

Contents lists available at ScienceDirect

Waste Management

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

Waste container weighing data processing to create reliable information of household waste generation

Pirjo Korhonen *, Juha Kaila

Aalto University, School of Engineering, Department of Civil and Environmental Engineering, P.O. Box 12100, FI-00076 Aalto, Finland

ARTICLE INFO

Article history:
Received 10 October 2014
Accepted 16 February 2015
Available online 9 March 2015

Keywords:
Household waste
Data cleaning
Data mining
Dimensioning
Waste management

ABSTRACT

Household mixed waste container weighing data was processed by knowledge discovery and data mining techniques to create reliable information of household waste generation. The final data set included 27,865 weight measurements covering the whole year 2013 and it was selected from a database of Helsinki Region Environmental Services Authority, Finland. The data set contains mixed household waste arising in 6 m^3 containers and it was processed identifying missing values and inconsistently low and high values as errors. The share of missing values and errors in the data set was 0.6%. This provides evidence that the waste weighing data gives reliable information of mixed waste generation at collection point level. Characteristic of mixed household waste arising at the waste collection point level is a wide variation between pickups. The seasonal variation pattern as a result of collective similarities in behaviour of households was clearly detected by smoothed medians of waste weight time series. The evaluation of the collection time series against the defined distribution range of pickup weights on the waste collection point level shows that 65% of the pickups were from collection points with optimally dimensioned container capacity and the collection points with over- and under-dimensioned container capacities were noted in 9.5% and 3.4% of all pickups, respectively. Occasional extra waste in containers occurred in 21.2% of the pickups indicating the irregular behaviour of individual households. The results of this analysis show that processing waste weighing data using knowledge discovery and data mining techniques provides trustworthy information of household waste generation and its variations.

1. Introduction

The waste prevention objectives, set by the European Union (EU), require monitoring municipal solid waste generation in member states (EC, 2008). Monitoring based on waste weighing data provides closer insight into municipal solid waste generation at sources and offers more special and individual information about the quantities and variation of municipal solid waste arising. Reliable data of waste quantities and generation trends are important information e.g. for planning and modelling waste management and to estimate resource management and the workload of waste collection services (Beigl et al., 2008; Rimaityte et al., 2012; Shamshiry et al., 2011; Cherian and Jacob, 2012).

In general, municipal solid waste generation is expressed in quantity of waste generated per capita in different time frames, mostly kg per capita per day or per year, or kg per household per week, and is based on official statistics or sampling data. Additionally, in many studies numerous factors which influence the quantity and composition of solid waste at the household level are also identified (e.g. Beigl et al., 2004; Skumatz, 2008; Dahlén et al., 2009; Denafas et al., 2014). However, the commonly highlighted attributes of waste-related data are uncertainty and unreliability (e.g. Dahlén, 2008; Karadimas and Loumos, 2008; Dahlén and Lagerkvist, 2010; Rada et al., 2013). The lack of reliable and disaggregated waste data is recognised although the data is gathered daily at operational waste management level and modern traceability devices with Global Positioning System (GPS) and General Packet Radio Service (GPRS) technologies allow real-time data collection and transmission (Faccio et al., 2011).

This paper discusses a new aspect to utilise municipal solid waste weighing data. Mixed waste weighing data selected from a database of municipal waste management authority is processed to determine the household mixed waste generation characteristics in residential properties based on the mixed waste quantities related to container capacity. Knowledge discovery in databases and data mining are adopted to create property-based groups, and their waste generation profiles. The main aim of the whole

* Corresponding author. Gsm: +358 505417752.
E-mail addresses: pirjo.korhonen@aalto.fi (P. Korhonen), juha.kaila@kasui.fi (J. Kaila).

http://dx.doi.org/10.1016/j.wasman.2015.02.021
0956-053X/

research is to develop a method to identify trend changes in waste generation from households based on operational data. In the future the method will be employed to evaluate the effect of waste avoidance campaigns and recycling programmes on waste arising both at the property and at different spatial levels.

2. Waste collection data

In waste-related studies the development of municipal solid waste generation models and the evaluation of waste management systems are usually based on municipal solid waste collection data. Parfitt et al. (2001) used operational municipal solid waste collection data to compare regions with their waste management and recycling performances by a hierarchical cluster analysis method. Municipal solid waste generation was defined as kg per household per week. Dahlén et al. (2009) evaluated household waste generation impact factors and collection systems in Swedish municipalities based on waste collection data, and annual statistics of local authorities and waste management companies. Dahlén and Lagerkvist (2010) used official household waste data to evaluate the strengths and weaknesses of weight-based billing in household waste collection systems.

Xu et al. (2013) based their hybrid model on historical municipal solid waste generation time series data (from 2000 to 2009) without demographics and socio-economic factors. They combined grey system theory with seasonal autoregressive integrated moving average (sARIMA) model to forecast seasonal and annual municipal solid waste generation. Navarro-Esbrí et al. (2002) utilised daily and monthly municipal solid waste collection data in their study to predict waste generation by a non-linear dynamics technique producing a result which was comparable to that of the sARIMA methodology. Benitez et al. (2008) collected residential waste samples of households for the waste generation analysis and modelling. In addition to the total weight of waste sampling bags, education level, household size, and income of participating households were included but seasonal variation was not taken into account in the analysis.

Waste collection data is typically discrete data and the variation of individual container weight values is quite large. Dahlén (2008) has pointed out that the precise metadata and the awareness of uncertainty sources of waste collection data increases the quality of input data and provides better research results. The general data problems presented by Dahlén are the result from (a) factors affecting waste generation, (b) technical devices, (c) variation of data registration level, (d) inadequate waste generator-related additional data, and (e) gaps in waste flows which mean that all wastes from households do not end up in the waste flows of municipal waste management. The variation of waste arising from households is also a result of different waste practises in households (Beigl et al., 2008).

Waste management information systems consist of several subsystems with databases such as collection and transfer systems, waste reception systems (e.g. weight bridge systems), and invoicing systems. The collection and transport systems feature logistic systems, such as transport control systems (TCS), which have applications for mobile terminal, software for a map and navigation, data transferring, and office software designed for the driver of a waste collection vehicle (Rada et al., 2010, 2013; Faccio et al., 2011). The data of the waste collection and transfer system are related to waste management services provided for the customers and offer special information to customers. The services are governed by municipal waste management regulations that define for instance separate collections of recyclables and maximum collection frequencies.

3. Material and methods

3.1. Data

3.1.1. Study area

The study area is the operational area of Helsinki Region Environmental Services Authority (HSY) covering five municipalities with an area of 1136 km^2 on the south coast of Finland. At the end of 2013, the population in the study area was in total 1,128,515 inhabitants. The characteristic of the area is mostly urban and sub-urban. Two thirds of the population lives in blocks of flats and only one tenth in single-family houses. From the households in the study area, 350,480 tonnes of municipal solid wastes were collected in the year 2013. Of this 189,488 tonnes (54%) were collected as mixed waste, 38,846 tonnes (11%) as biowaste, 115,526 tonnes (33%) as other recyclables, and 6620 tonnes (2%) as other waste (HSY, 2014a,b).

In Finland, according to the Waste Act (646/2011) the municipalities are responsible for offering municipal solid waste collection services to residential properties, public services, and private health and educational services. Mixed waste from every property is collected by on-site or property-close collection systems. There is also separate collection of the waste fractions for material recovery from multi-family properties, and public and private services. The number of fractions to be separated at a source is defined in the municipal waste management regulations according to the number of housing units in a property or the weekly generation of a waste fraction at the service properties. In Helsinki Metropolitan Area residential properties with ten or more housing units are required to organise the separate collection of paper, biowaste, and cardboard, and the properties with 20 or more housing units have to organise the separate collection of paper, biowaste, cardboard, glass and metal in an addition of mixed waste collection. Properties with less than ten households are recommended to take voluntary source-separated waste fractions to drop-off points and compost their biowaste. The authority (HSY) organises the collection of recyclables except for recyclable paper, which is under producer liability and it is forbidden to put recyclable paper into mixed waste containers by regulation (HSY, 2012).

The property owner signs a waste service agreement with HSY for household waste collection and transportation. The contract is saved in a waste management database which includes the identifiable information about the customers and their waste collection points, separate collected waste fractions, number and capacity of bins and containers, type of bins and containers, and emptying schedules. Collection task lists for the drivers of waste collection vehicles are generated from the customer database. The task lists contain all necessary information required to empty the right bins or containers at the property, such as the collection address, and the number of bins or containers at that location. After emptying the driver signs the task and, when a container is weighed, the weight is updated to the task row manually or it is transferred from the scale to the computer in the vehicle by wireless data transfer technology. After the collection route is completed, the emptying data is transferred to the waste management database for billing and other utilisation.

In 2013 the total number of pickups of mixed waste bins and containers was 5,609,466 in the service area of HSY. Table 1 shows

Table 1
Mixed waste pickups from different bin types at the study area. (HSY, 2014b).

Bin/container	Nr of pickups 2013
140–300 L bins and bags	1,352,224
600–660 L bins	4,189,731
Others	67,511
Total	5,609,466

that most of the pickups, over 4 million, concerned traditional mixed waste wheeled bins with 600 L or 660 L capacity (HSY, 2014b). The pickups in the study data are included in the group 'Others' which mainly consist of weighted containers.

3.1.2. Target data set

The original data contained 55,459 weight measurements and it was selected from the waste management database of Helsinki Region Environmental Services Authority. The data contains mixed waste and biowaste amounts in deep collection, front load and ground containers with capacities from 2 to 8 m^3. In the data, 87% of the emptying tasks were from residential properties. Other waste generator sectors were public services (e.g. education, offices, and institutional care), private services (e.g. commercial, industrial, and warehouses) and others (e.g. drop-off points), the shares of which were less than 2% for each. The majority of the waste collection points are at the blocks of flats (82%), and the rest are at row houses (11%), and others 7% (one-dwelling houses, two-dwelling houses, terraced houses and balcony-access blocks).

The data set contains both static and dynamic data of mixed waste collections. Table 2 describes the metadata of the data set variables. The waste collection points are located by *X*,*Y*-coordinates in ETRS-GK25 coordinate system. Additional information is related both to attributes of property as a customer and attributes of collected waste. The timestamp of pickup is the basis to time series analyses for which the collection date was converted to the day-of-year number.

According to the information content of the data set (see Table 2), it is possible to explore waste generation on different levels from a waste container to a region. At the container level the information prior to a pick up is waste type to be collected, container type and number of containers to be emptied, the capacity of container(s), and the location of the container. After emptying, additional information is the weighed amount of collected waste in kilograms, the number of emptied containers, and the emptying timestamp. This real time information is a tool for customer service to control waste collection and inform customers in exceptional situations. The property level data visualises how waste is generated in time, and it can be used to compare with historical data as feedback. Housing type is utilised to compare waste generation with other properties of the same housing type. However, it is not possible to describe waste generation at household level without the identification of households.

3.1.3. Sub-data set for processing

Residential mixed waste collection and weighing data by housing type was selected from the original target data to represent household waste generation in the region. The sub-data set included 43,593 data rows. The general capacity of containers for mixed waste was six cubic metres covering 86% of residential pickups and the general container type was a deep collection container, 95% of all containers (Table 3). The capacity of the container represents the storage space for wastes until collection.

The mixed waste weighing data related to a single 6 m^3 container in each collection point was selected from the sub-data set for further processing. This selection was made to unify the data and to enable to compare the waste arising time series in analysis. The other reason for the selection was that the weight values in the original data set are summarised on the waste collection point level and thus the container-based values are not available for those points with more than one container. The final data set includes 27,865 pickups of 579 deep collection containers and 13 front load containers and they cover 64% of residential mixed waste pickups of the target data. The locations of the containers selected for the analysis are shown in Fig. 1.

Table 2
Metadata of waste weighing data set.

Variable	Data type	Explanation	Used in analysis
Contract area	Static	Municipal waste management authority has divided Helsinki Metropolitan Area in smaller areas for organising MSW collection with private waste collection and transfer companies	
Agreement number	Static	The identification number of the customer in MWMA customer database	
Pickup address	Static	Street address of a waste collection point (it is not necessarily equal with official street address of customer)	√
Waste collection point number	Static	Property might have several waste collection points according to number of buildings	√
Waste fraction	Static	The type of waste fraction to be collected (in this dataset mixed waste or bio waste)	√
Type of container	Static	A deep collection container, a front load container, or ground container	√
Capacity of container	Static	2, 3, 4, 6, and 8 cubic metres (m^3)	√
Number of containers	Static	Number of containers at the waste collection point for the waste fraction	√
Number of emptied containers	Dynamic	Number of containers emptied in one pick up event	√
Total capacity of emptied containers	Dynamic	Number of emptied containers multiplied by the capacity of containers (calculated in database)	√
Total amount of waste	Dynamic	Total amount of collected waste at a waste collection point (summarised in database)	√
X	Static	*X*-coordinate of waste collection point (ETRS-GK25 coordinate system, longitude)	√
Y	Static	*Y*-coordinate of waste collection point (ETRS-GK25 coordinate system, latitude)	√
LoadID	Dynamic	Identification of load	
Unloading place	Dynamic	Waste treatment facility where waste is transported	
Housing type	Static	Housing type code according to building types based on the building classification (Statistics Finland, 1994)	√
Timestamp	Dynamic	Date and time when assignment is signed	√

Table 3
Number and types of different containers for household mixed waste in the sub-data set.

Container type	Capacity				
	2 m^3	3 m^3	4 m^3	6 m^3	8 m^3
Deep collection container	13		110	1099	2
Front load container	1		13	15	34
Ground container		3			
Total	14	3	123	1114	36

3.2. Methods

3.2.1. Knowledge discovery process

Knowledge discovery process and data mining are processes for digging and identifying previously unknown, potentially useful, advantageous relationships and patterns from large data sets. The discovery process is described as steps from database to the

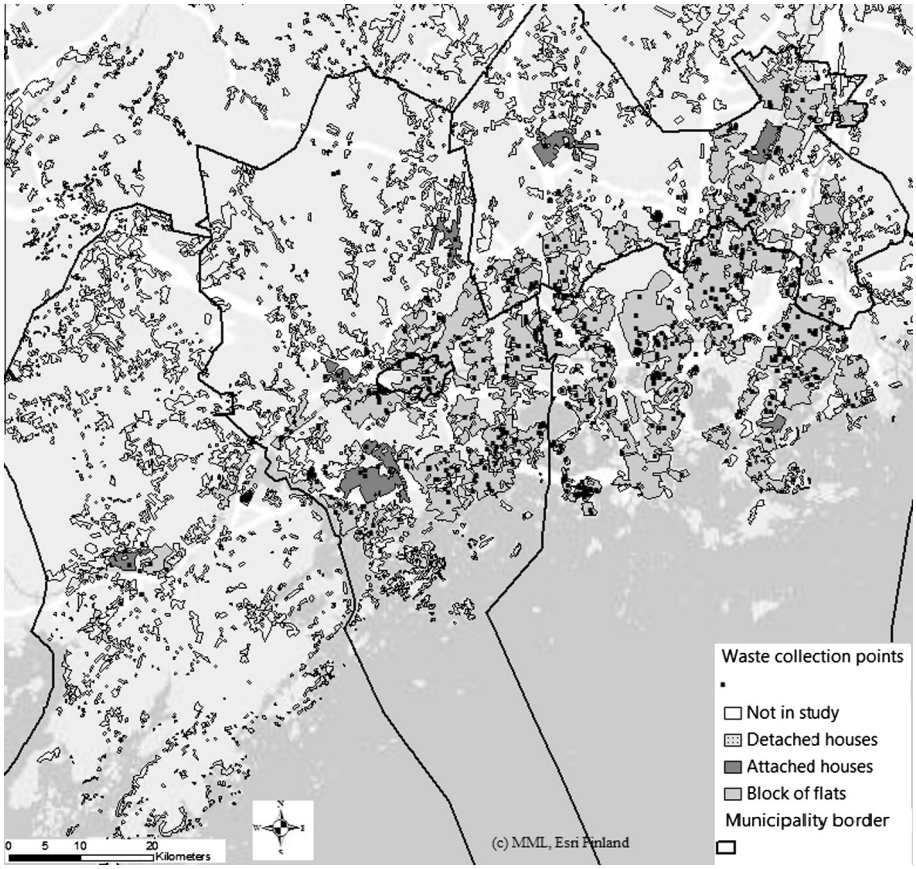

Fig. 1. Map of study area with the location of containers in the study.

discovered knowledge of a domain. The first step is to select a target data set from a database on which knowledge discovery is performed. Secondly, the target data set is cleaned and pre-processed for further processing. In the cleaning process incorrect and incomplete values in data are detected and corrected or removed. In the third step, data mining methods are adopted to discover different patterns and relationships in the data set. In this phase, data exploration tools, such as summaries and visualisation, are utilised to understand the data. Fourth, the results of data mining are converted to knowledge by evaluation and interpretation. Finally, the useful knowledge is published via user applications (Fayyad et al., 1996). In this paper the steps from one to tree are exploited in processing and evaluating the waste weighing data. Data processing was conducted using the program packages MATLAB® R2013b and Microsoft Excel 2010.

3.2.2. Data cleaning

Data cleaning includes strategies for handling missing values and errors, like inconsistency values and outliers, aiming at reliable data and avoiding false conclusions. The missing values are generally caused by technical problems in weighing a container, or the driver is unable to sign the assignment due to a computer failure. In latter cases, the emptying task is updated into the database manually in office afterwards. The missing weight values are replaced in original data with the number of containers or the total capacity of a container, mostly by value 1. Not only the number of missing values was studied but also how they are distributed in the study area: are there certain containers of which waste amounts are commonly missing or are missing values randomly distributed. Missing values indicate the completeness of studied data and in further its accuracy (Gorla et al., 2010).

The inconsistency values of weights are out-of-scale values which differ significantly from previous or next emptying weights of the container. Thus, a method of distinguishing inconsistency values as human errors from essential outliers or noise as a part of normal variation is necessary otherwise some of relevant data is lost (Skutan and Brunner, 2012). According to several studies, the amounts of collected waste vary highly at the property level and per capita (Nuortio et al., 2006; Dahlén, 2008; Dahlén and Lagerkvist, 2010). The variation is the consequence of seasonal variation, temporal changes in a collection schedule and different waste practises in households (Beigl et al., 2008), or differences in measuring the amount of waste (Dahlén and Lagerkvist, 2010). Basu and Mechesheimer (2005) grouped outliers in two groups: additive outliers and innovative outliers. An additive outlier is a result of human error or technical breakdown of the system while an innovative outlier is caused by a change in the system. The detected abnormal low or high values in waste related data are typically additive outliers as a result of measurement errors (Beigl et al., 2004; Fellner et al., 2007; Clavreul et al., 2012) or an inadequate sample preparation (Skutan and Brunner, 2012). Innovative outliers are abnormal values which might be an outcome of changes in the system or some external factors.

The identification method of outliers and inconsistently values is a critical issue when modelling waste generation using waste collection data. The main question is how to identify outliers as abnormal values and not as errors. In waste related studies, the outlier detection method has been used to identify values which deviate more than 30% from the median or the trend line (Lebensorger and Beigl, 2011) while in data cleaning an outlier has been defined as a value which is more than 2 standard deviations from the mean (Hellerstein, 2008). However, in this study the

method of identifying the inconsistently high values in waste weighing data is based on technical limits of (deep collection) containers recommended by manufacturers.

A deep collection container consists of well body, lifting bag and lid. Two technical limit values are available for the lifting bag: safe working load and maximum lifting capacity. The maximum lifting capacity is the amount of waste which a lifting bag can carry without breaking. The safe working load describes the load value which is safe for the whole lifting operation and it is based on the maximum load recommended by the manufacturer of the container. The maximum lifting capacity of the lifting bag in this study is 2200 kg (J. Salli, personal communication, May 13, 2014) while the safe working load for the lifting bag is 1250 kg. These limit values are used in the data cleaning process.

3.2.3. Data mining

In this study, descriptive data mining techniques, like descriptive statistics and distribution, were used to understand how the selected data set represents the mixed waste arising from households living in different housing types in the region. Analysing how the measured weights are in compliance with the dimensioning values for the containers based on the variation of mixed household waste arising in time indicates the reliability of the weighing data. The curves of seasonal variation were created with MATLAB by fitting the time series data with *smooth* function with *lowess* (locally weighted scatter plot smooth) method and span value 0.1, which uses 10% of all data points in each sub-data set. Regression weights for each data point within the span were computed and then a weighted linear least-squares regression using a first degree polynomial was calculated. The normal mixed waste variation range was defined by two different smoothed curves:

- range *a*: smoothed median ± 30% of median value,
- range *b*: smoothed median ± two times smoothed standard deviation.

The median was selected as the base for variation because it is more resistant to outliers and noise than mean in discrete data sets (Painter et al., 2011).

The distribution of the pickup weights by housing type was analysed with kernel distribution which has the advantage that it produces a smooth, continuous probability curve while histogram places the values in discrete bins. The kernel distribution is generally used when any assumptions about the distribution of data are avoided.

4. Results and discussion

4.1. Missing values

The final data set included 27,865 pickups in 2013 and the weight value was missing in 97 pickups, which is 0.35% of the total. Six properties covered 53% of these missing values which might indicate some difficulties in weighing the containers or other problems at these properties. Similar results were achieved when investigating the missing values with the load identifier (loadID); there were only from one to three missing weight values per load. However, in one day one route had eight missing values of which seven were the first pickups. This might have been caused by some technical problems with the crane scale or the computer in the collection vehicle. The missing values were detected in four contract areas of nine and 55% of missing values were in one contract area. The results show that most of the measurement problems occur only at a few properties. The missing values were removed from the dataset before further processing.

4.2. Inconsistent values and outliers

Fig. 2 shows five weighing values which exceed the maximum lifting capacity of the lifting bag (2200 kg) and therefore they are impossible. These values might have been caused by human errors when weight values were fed into the database. Comparing the high values with the prior weights of the same container shows that the order of magnitude difference was most probably caused by an additional zero which multiplies the weight, e.g. 2900 kg while two prior weights were 260 kg and 210 kg.

The inconsistently high values were removed from the data set because the number of these values was very low and the impact on the results of further analysis was considered insignificant. The weights between the maximum lifting capacity and the safe working limit in Fig. 2 are possible weights in exceptional cases or they might also be errors. They were considered as outliers and were not removed from the data set.

Weight values which were less than 30 kg were regarded as inconsistently low values. This limit value is 10% of the average pickup weight (300 kg) of mixed waste in a 6 m^3 deep collection container in Helsinki region. The search resulted in 68 pickups which were investigated one by one. The investigation showed that most of the inconsistently low values might have been feeding errors supposing that a zero at the end of the weight value is missing. Values of 10 kg or less were detected in cases where the value was fed within a few minutes after previous weighing at the same location. These low values were assumed to be errors in entering the weights manually into the computer or there might have been extra waste outside the container. Some low weights were identified due to confusion in waste collection because the container was emptied on a previous day. In these cases, the waste amounts were correct according to short accumulation time (one day). The weights less than 30 kg were removed from data set prior to further analysis.

4.3. Noise in waste weighing data

Noise in waste collection data is an essential feature in household waste generation as a result of external factors as well as both individual and collective behaviour of households. Seasonal variations, which are a typical result of collective behaviour of households, have been demonstrated in waste related studies (e.g. Navarro-Esbrí et al., 2002; Dahlén and Lagerkvist 2008; Salhofer et al., 2008; Gómez et al., 2009; Xu et al., 2013; Denafas et al., 2014). The noise points are outside the 'normal' waste generation range. To identify noise points the normal variation of mixed waste generation in the data set has to be defined. The weight values in the cleaned data set were box plotted to visualise the variation of daily mixed waste quantities in emptied containers. Fig. 3 shows that the median of daily quantities varies from 200 kg to 400 kg with a few exceptions. Waste collection is generally performed on weekdays from Monday to Friday, but there are also exceptional pickups on Saturdays and even Sundays due to midweek holidays, such as Easter and Midsummer.

The data (Fig. 3) is consistent with the dimensioning principles of Helsinki Metropolitan Area waste management authorities. For bin and container dimensioning purposes it is assumed that each person generates in average 1.9 m^3 of mixed household waste annually equalling 170 kg per person per year (HSY, 2014a). To take into account the seasonal variations and some reserve capacity it is assumed that the average filling level of containers is 50–70% when emptied. According to these dimensioning principles, the median weight of waste in a 6 m^3 container should be 260–357 kg when emptied.

In Fig. 4 the weight data is compared with outlier detection methods presented by Lebensorger and Beigl (2011) (range a), and by Hellerstein (2008) (range b).

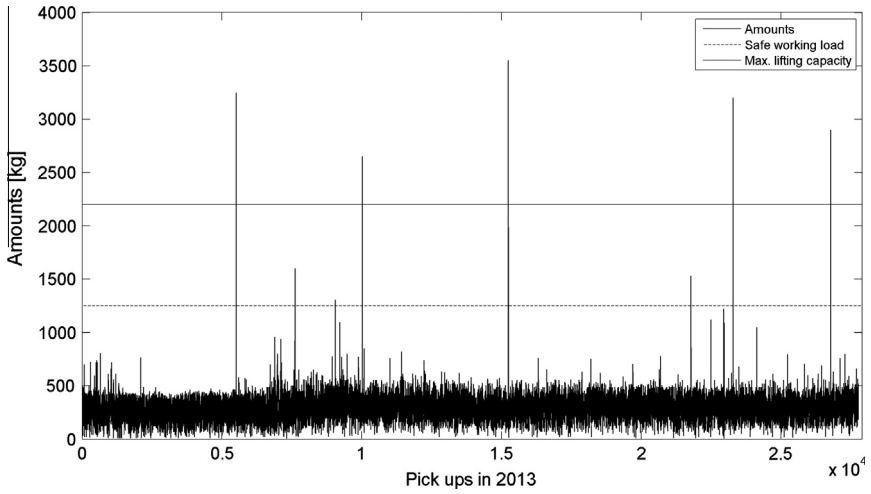

Fig. 2. Weighing data and safe working load (lower) and maximum lifting capacity (upper) of a 6 m^3 container.

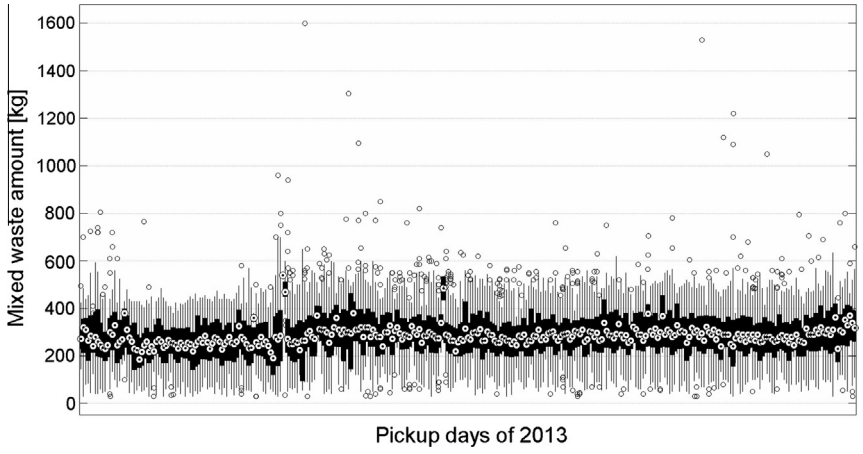

Fig. 3. Daily variation of mixed waste quantities in data set (missing values, inconsistently high and low values are removed).

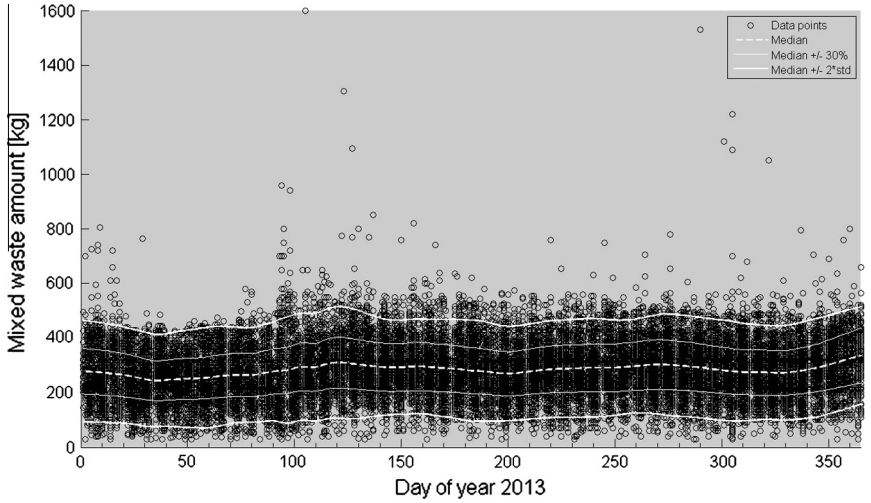

Fig. 4. Two ranges for definition of normal household mixed waste arising variation.

Fig. 4 shows that range *a* is quite narrow for overall daily pickup weights whereas range *b* covers most of pickups. Range *a* define 38% of all data points as outliers and range *b* covers most of the data points leaving only 5% of all data as outliers. By definition, outliers should represent values, which are outside the normal variation. When the measured value is above the normal variation it indicates that extra waste has been put to the container. Based on waste quality studies in Helsinki region, the average share of extra waste is 6 w-% (HSY, 2013) leading to the conclusion that range *b* is too wide when the effects of seasonal variation and differences between individual collection points are taken into account. Thus variation range *a* is better from the waste collection point of view when the same households dispose of their wastes into the same container as a rule.

4.4. Effect of housing type

The data set was grouped by housing type which is found to have an effect on waste arising in earlier studies (Emery et al., 2003; Martin et al., 2006; Dahlén, 2008; Timlett and Williams, 2011). First, the main statistics were calculated for different housing types from the final data set (Table 4). The median of mixed waste pickup weights in the data set is the lowest from households living in two-family dwellings. However, it is important to notice that the data set represent household mixed waste generation mostly from new residential properties because deep collection containers have become more popular in the region. Especially, the centralised mixed waste collection containers are installed in new residential areas of one-family and two-family dwellings where households from several properties share these container(s) at a single waste collection point instead of each family having its own small bin. Generally, at these points also containers for recyclable paper (e.g. newsprint), cardboard, and biowaste are located.

The housing types were grouped into three groups according to the similarities of housing types: *the detached houses* group includes one- and two-family houses, *the attached houses* group includes row houses and terraced houses, and *the block of flats* group includes balcony-access blocks and blocks of flats (Statistics Finland, 1994). The graph in Fig. 5 shows that the mixed waste arising from households living in different housing types is almost normally distributed. The households living in the detached houses generate mixed waste in wider range than others. The right tail of the curve is in all groups quite the same, but left tail is wider in the detached houses group. This distortion to the left indicating positive skewness has been found to be common in solid waste generation data (Tchobanoglous et al., 1993). Actually, the distribution of mixed waste arising from households in the attached houses group is closest to normal distribution in the studied data set. Although waste weight values are positive numbers the probability density curve is a sum of individual density curves created for each data point which results in negative possible values.

4.5. Seasonal variation

Seasonal variation is observed as high and low peaks in the curves of median and mean for year 2013 pickup weights (Fig. 6). After Christmas and New Year holidays in the beginning of year the waste generation in households decreases and is on its' lowest in the beginning of February when the smoothed median value is 244 kg. The mixed waste generation is above the overall median value (280 kg) from April to the end of June covering days from 97 to 178. In this period there are four midweek holidays (Easter – four days, the First of May – one day, Ascension Day – one day, and Midsummer – two days) and also the end of the school year. The first high peak is around the First of May with smoothed median value 310 kg and two minor peaks are after Easter (294 kg) and in the beginning of June (296 kg). These variations are connected to collective similarities in the consumption behaviour of households, but also spring with warm weather effects on activities in some households which causes increased waste generation. The lowest generation (269 kg) is between days 178 and 217 on the main summer holiday period which is from Midsummer to the end of July (6 weeks) in Finland. Mixed waste amounts are rising during the first half of August, when the school year starts, reaching the top (303 kg) in the end of September. The last peak is in December during Christmas holidays (3 days) reaching the highest smoothed median of 335 kg on the day 365 when wastes generated during holidays are collected.

4.6. Variation at collection point level

The mixed waste weighing data time series for each collection point were compared with normal mixed waste variation range taking seasonal variation into account. For each waste collection point the range limits were set as range *a*: the upper limit value was set as 30% above and the lower limit value as 30% under the smoothed median value of the pickup day. Then the waste collection points were categorised into groups analysing the pickup weight time series according to the limits. The groups are:

- Optimal container capacity: waste collection points where pickup weight time series were within the normal variation as a rule.
- Over-dimensioned container capacity: waste collection points where pickup weight time series were below the normal variation as a rule, but followed the general pattern of seasonal variation presented in Fig. 6.
- Under-dimensioned container capacity: waste collection points where pickup weight time series were above the normal variation as a rule, but followed the general pattern of seasonal variation presented in Fig. 6.
- Household mixed waste with occasional extra waste: waste collection points where pickup weight time series were within the normal variation as a rule with occasional weights over the range *a* upper limit value for the pickup day.

Fig. 7a shows the time series of a collection point where the container capacity dimensioning is optimal. All weights are within the normal variation range (dash lines) and the distribution curve follows seasonal variation (*see* Fig. 6). The optimal dimensioning ensures that there is always space in a container for a waste bag of household and over filling is avoided. Over-dimensioned

Table 4
Descriptive statistics of mixed waste arising from different housing types in 6 m^3 containers.

Housing type	Code	Pickups	Min (kg)	Max (kg)	Mean (kg)	Standard deviation (kg)	Median (kg)
One-dwelling houses	11	433	30	525	247	112	260
Two-dwelling houses	12	777	30	560	239	110	235
Row houses	21	3741	40	1090	283	95	280
Terraced houses	22	261	50	770	284	90	265
Balcony-access blocks	32	673	45	750	280	100	285
Block of flats	39	21810	30	1600	286	100	280

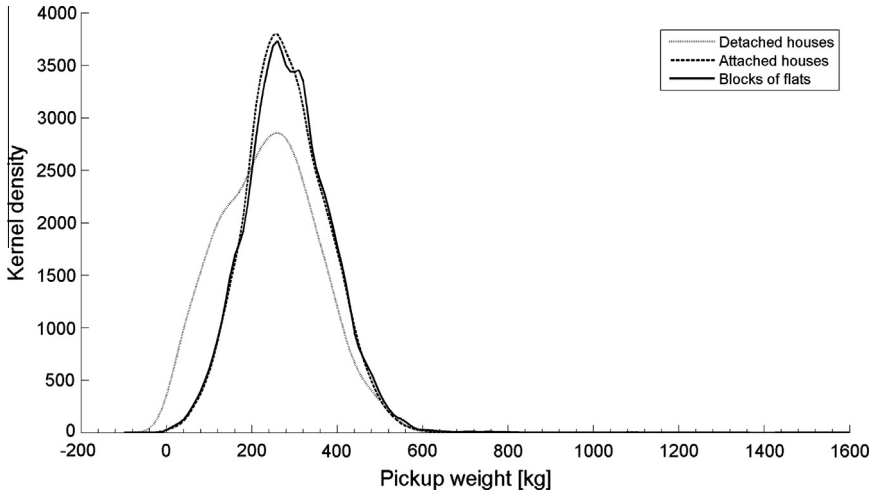

Fig. 5. The kernel distribution of mixed waste amounts in deep collection container with six cubic metres capacity in three housing type groups in year 2013. (Missing values and inconsistently high values are removed.)

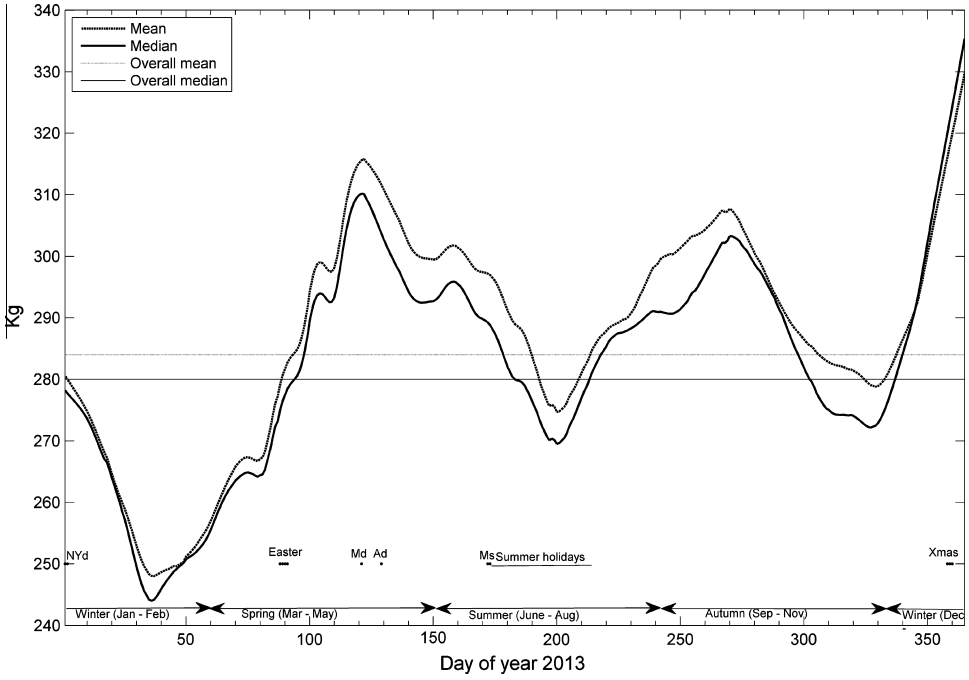

Fig. 6. Seasonal variation of mixed household waste arising in 2013 presented by smoothed median and mean of mixed waste quantities in 6 m^3 containers.

container capacity (Fig. 7b) in the above grouping means that the amount of waste in the container is less than expected when the dimensioning principles used in Helsinki Metropolitan Area (HSY, 2014a) are taken into account. As stated earlier, the median weight should be 260–357 kg for 6 m^3 containers. Reasons for over-dimensioning can be for instance: the number of people served by the container is smaller than expected, the households generate less waste than average, or the collection frequency is not optimal and the container is emptied too often. Same type of reasoning apply also to under-dimensioning (Fig. 7c), but this time the number of people served by the container is bigger than expected, the households generate more waste than average, or the pickup interval is too big. Waste in a full six cubic metres container of ordinary mixed household waste weighs about 540 kg, and the quantity of waste in an under-dimensioned container exceeds the full container weight limit several times during a year causing over filling.

Fig. 7d shows an example of the collection weight time series at the collection point where the container capacity is optimal but there are occasionally extra wastes. Commonly, extra waste is generated when people move in and out, and when they renovate their homes. So it is possible that extra waste can be found also in containers with weights between normal variation limits.

The total quantity of extra waste exceeding the upper limit of normal variation was evaluated by studying each waste collection point separately comparing weights against their moving medians +30%. First, the overall median (280 kg) was set to 1 and then the coefficients for each collection day were calculated from the seasonal variation curve. Secondly, the normal variation upper limit values for each collection point were calculated multiplying the overall median of the point with both the collection day coefficients and constant 1.3. This procedure is based on the assumption of collective behaviour among households represented by seasonal

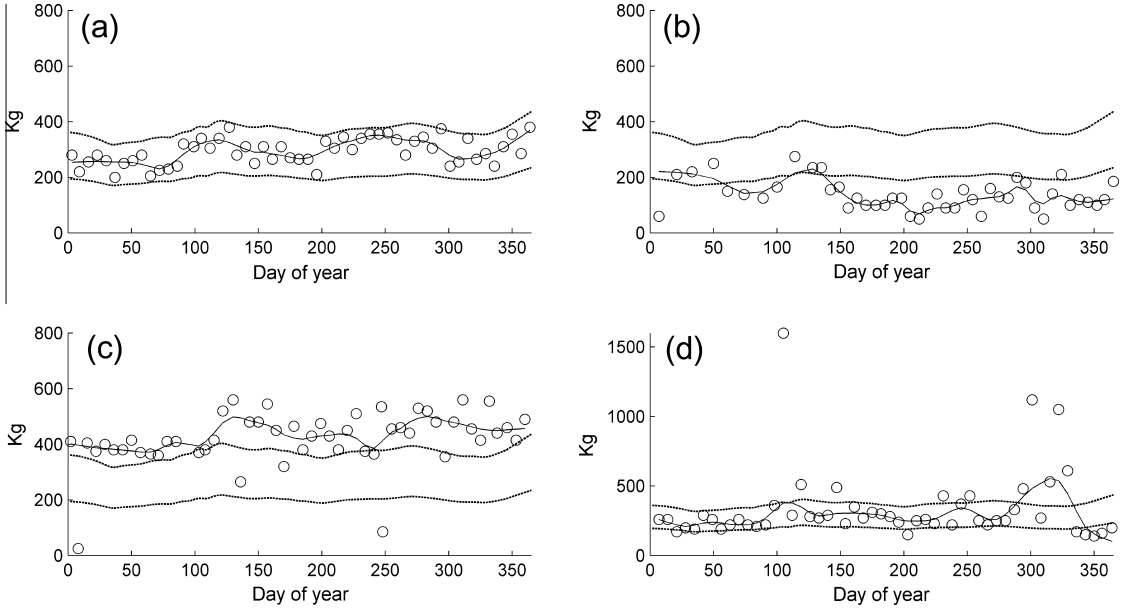

Fig. 7. Examples of time series in different dimensioning groups. (○ measured weight, — moving median for the collection point in question, ---- normal variation limits).

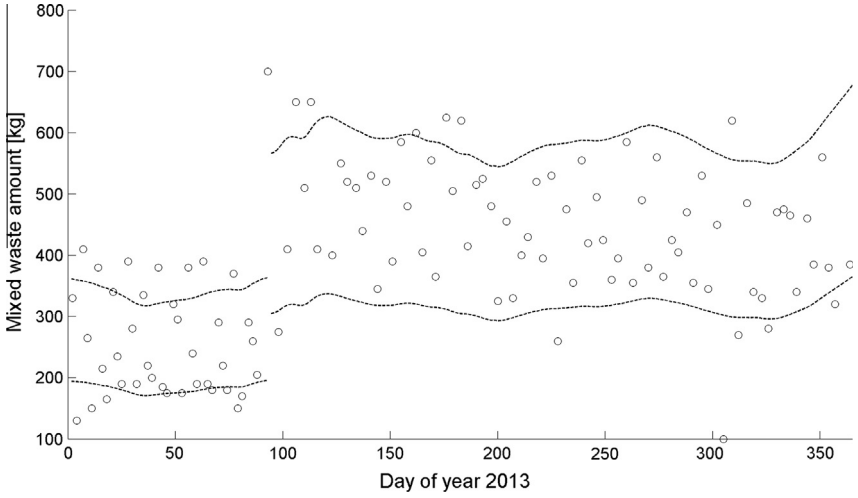

Fig. 8. A level shift in mixed waste time series caused by a change in waste collection system. (○ measured weight, — moving median for the collection point, ---- normal variation limits).

variation. The amount of extra waste was calculated as the difference between pick up weight and the upper limit of collection day at each collection point. With this method, the share of extra waste in containers was 2.4 w-% of collected mixed waste. As stated earlier, the total amount of extra waste in mixed household waste is about 6 w-% (HSY, 2013). Because extra waste is found also in containers with weight values within normal variation, it can be concluded that the method presented above gives plausible information of the weight values above normal variation.

4.7. Other sources of noise in data

The used method of evaluating container dimensioning enables also to reveal the impacts on pickup weights when a change in the waste collection system has been implemented. Fig. 8 illustrates the effect on time series when the collection frequency has been changed. From the beginning of the year until day 102, the container capacity is optimal with several pick up weights outside the normal variation limits. After the collection frequency was decreased from three times a week (2–2–3 day interval) to twice a week (3–4 day interval) the pickup weights are higher because of longer accumulation time between pickups. According to Basu and Mechesheimer (2005) the values after day 102 would have been identified as innovative outliers causing a level shift, but according to this analysis they are quite normal values when the system-change causing the under-dimensioning is taken into account using the normal variation limits of the waste collection point.

Some waste collection points could not be categorised to any of the above groups. These points were defined as 'grey data'. Commonly they were points from which only some weight values existed and thus complete time series were not available. In one case, the weighing data was connected to the one-family dwelling housing type while the container is located near a public sports ground and the waste is collected twice a week indicating that also other wastes than ordinary household waste were put into the container. Thus, because of uncertainty of the origin of the waste, this waste collection point was included in the 'grey data' category.

Table 5
The summary of pickups in the studied waste weighing data set.

Household mixed waste	Shares of pickups (%)
Normal generation	
Optimal container capacity	65.0
Over-dimensioned	9.5
Under-dimensioned	3.4
Normal generation with occasional extra waste	21.2
Normal generation total	99.0
'Grey data'	0.4
Missing values and errors (inconsistently high and low values)	0.6
Total	100.0

5. Summary

After all waste collection points in the data set were categorised in groups, the sum of pickups in each group were calculated for evaluation. The summary of the evaluation is presented in Table 5 where the shares of missing values, errors, 'grey data', and pickups in different dimensioning groups are calculated. The share of pickups from waste collection points with optimal container capacity and normal generation with occasional extra waste show that the container dimensioning principles used by the regional authority HSY are relevant for household mixed waste generation. However, nearly 12% of the pickups are from points where waste services are over-dimensioned or under-dimensioned and thus service optimisation should be considered for improved service quality and cost efficiency.

Table 5 provides evidence that 99% of the waste weighing data covers the normal variation and occasional extra waste of household mixed waste generation and thus it is reliable and certain. The original weighing data is gathered from waste containers, which enable weighing during the pickup event which might limit the representativeness of data set. However, the waste collection points in the data set are located randomly over the region representing different residential areas.

6. Conclusions

In this study, the waste weighing data represents how wastes accumulate from households into waste containers at waste collection points over time. Actually, the data reveals more about household practices with their wastes rather than waste generation in individual households. The generated waste from different households ends up in a waste container at different rates. This causes, as for, variation in composition and amounts of mixed waste which leads to the variation in emptying weights. The assumption is that the households dispose of their wastes with regularity into the container(s) at the same waste collection point close to their homes. Furthermore, earlier research in the waste field has studied numerous factors which are expected to correlate with waste collection results, but according to Beigl et al. (2008) it is impossible to identify the impacts which have caused the varying collection quantities. This is true considering the irregular noise in weighing data. Collective similarities in behaviour, however, can be detected as has been shown in the seasonal variation analysis.

The waste weighing data used in this study is continuous and contains detailed information of mixed waste arising from households living in different housing types. The studied data set is gathered at the waste collection point level and it covers a whole year including relatively few missing weight values and errors. Characteristic to mixed waste arising at the waste collection point level is a wide variation between pickups which appears as outliers and noise in waste arising time series. This study provides evidence that the variation in waste weighing data is mostly a result of households' waste behaviour rather than the unreliability of data as suggested in some previous studies (Dahlén et al., 2009). The result shows also that almost all of the variation in waste weighing data can be explained by analysing the data with knowledge discovery and data mining techniques.

The low share of errors and missing values indicates good completeness of the waste weighing data. The data was found to be accurate and trustworthy, and thus provide reliable information for the planning of waste management services to improve household waste practices and monitoring the effects of these improvements on waste arising. The data cleaning method used in this study is utilised in future to create more analytical knowledge about household waste arising in urban areas.

Acknowledgements

The authors would like to thank Helsinki Region Environmental Services Authority for the provision of waste weighing data and the European Regional Development Fund (ERDF) in Southern Finland for the financial support.

References

Basu, S., Mechesheimer, M., 2005. Automatic outlier detection for time series: an application to sensor data. Knowl. Inform. Syst. 11 (2), 137–154. http://dx.doi.org/10.1007/s10115-006-0026-6.

Beigl, P., Wassermann, G., Schneider, F., Salhofer, S., 2004. Forecasting municipal solid waste generation in major European cities. In: Pahl-Wostl, C., Schmidt, S., Jakeman, T. (Eds.), iEMSs 2004 International Congress: "Complexity and Integrated Resources Management". Osnabrueck, Germany. <http://www.iemss.org/iemss2004/pdf/regional/beigfore.pdf> (retrieved July, 2014).

Beigl, P., Lebersorger, S., Salhofer, S., 2008. Modelling municipal solid waste generation: a review. Waste Manage. 28, 200–214. http://dx.doi.org/10.1016/j.wasman.2006.12.011.

Benitez, S.O., Lozano-Olvera, G., Morelos, R.A., de Vega, C.A., 2008. Mathematical modeling to predict residential solid waste generation. Waste Manage. 28, S7–S13. http://dx.doi.org/10.1016/j.wasman.2008.03.020.

Cherian, J., Jacob, J., 2012. Management models of municipal solid waste: a review focusing on socio economic factors. Int. J. Econ. Finan. 4 (10), 131–139. http://dx.doi.org/10.5539/ijef.v4n10p131.

Clavreul, J., Guyonnet, D., Christensen, T.H., 2012. Quantifying uncertainty in LCA-modelling of waste management systems. Waste Manage. 32, 2482–2495. http://dx.doi.org/10.1016/j.wasman.2012.07.008.

Dahlén, L., 2008. Household Waste Collection. Factors and Variations. Doctoral Thesis, Department of Civil, Mining and Environmental Engineering Division of Waste Science and Technology Luleå University of Technology, Luleå, Sweden.

Dahlén, L., Lagerkvist, A., 2008. Methods for household waste composition studies. Waste Manage. 28, 1100–1112. http://dx.doi.org/10.1016/j.wasman.2007.08.014.

Dahlén, L., Lagerkvist, A., 2010. Pay as you throw. Strengths and weaknesses of weight-based billing in household waste collection systems in Sweden. Waste Manage. 30, 23–31. http://dx.doi.org/10.1016/j.wasman.2009.09.022.

Dahlén, L., Åberg, H., Lagerkvist, A., Berg, P., 2009. Inconsistent pathways of household waste. Waste Manage. 29, 1798–1806. http://dx.doi.org/10.1016/j.wasman.2008.12.004.

Denafas, G., Ruzgas, T., Martuzevičius, D., Shmarin, S., Hoffmann, M., Mykhaylenko, V., Ogorodnik, S., Romanov, M., Neguliaeva, E., Chusov, A., Turkadze, T., Bochoidze, I., Ludwig, C., 2014. Seasonal variation of municipal solid waste generation and composition in four East European cities. Resour. Conserv. Recy. 89, 22–30. http://dx.doi.org/10.1016/j.resconrec.2014.06.001.

EC – European Commission, 2008. Directive 2008/98/EC of the European Parliament and of the Council of 19 November 2008 on Waste and Repealing Certain Directives (Waste Framework Directive). Official Journal 22/11/2008, L 312/3.

Emery, A., Griffiths, A., Williams, K., 2003. An in depth study of the effects of socio-economic conditions on household waste recycling practices. Waste Manage. Res. 21, 180–190.

Faccio, M., Persona, A., Zanin, G., 2011. Waste collection multi objective model with real time traceability data. Waste Manage. 31, 2391–2405. http://dx.doi.org/10.1016/j.wasman.2011.07.005.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery in databases. AI Mag. 17 (3), 27–34, <http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf> (retrieved September, 2013).

Fellner, J., Cencic, O., Rechberger, H., 2007. A new method to determine the ratio of electricity production from fossil and biogenic sources in waste-to-energy

plants. Environ. Sci. Technol. 41, 2579–2586. http://dx.doi.org/10.1021/es0617587.
Gómez, G., Meneses, M., Ballinas, L., Castells, F., 2009. Seasonal characterization of municipal solid waste (MSW) in the city of Chihuahua, Mexico. Waste Manage. 29, 2018–2024. http://dx.doi.org/10.1016/j.wasman.2009.02.006.
Gorla, N., Somers, T.M., Wong, B., 2010. Organizational impact of system quality, information quality, and service quality. J. Strateg. Inform. Syst. 19 (3), 207–228. http://dx.doi.org/10.1016/j.jsis.2010.05.001.
Hellerstein, J.M., 2008. Quantitative Data Cleaning for Large Databases. United Nations Economic Commission for Europe (UNECE). <http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf> (retrieved March, 2013).
HSY Helsinki Region Environmental Services Authority, 2012. Waste Management Regulations. <http://www.hsy.fi/jatehuolto/Documents/Palvelut/Kiinteiston_jatehuolto/Jatehuoltomaaraykset 2012.pdf> (retrieved May, 2012).
HSY Helsinki Region Environmental Services Authority, 2013. Quality and Quantity of Household Mixed Solid Waste in the Helsinki Metropolitan Area 2012. <http://www.hsy.fi/tietoahsy/Documents/Julkaisut/2_2013_pks_kotitalouksien_sekajatteen_maaja_ja_laatu_lr.pdf> (retrieved March, 2013).
HSY Helsinki Region Environmental Services Authority, 2014a. Waste Flows of Helsinki Metropolitan Area. <http://www.pksjatevirrat.fi/?mo=stats&y=2012&rm=5&view=54> (accessed 17.09.14).
HSY Helsinki Region Environmental Services Authority, 2014b. Waste Statistics 2013. <http://www.hsy.fi/jatehuolto/Documents/Ymparisto/Laatu/HSY_jatehuollon_vuositilasto_2013.pdf> (accessed 16.04.14).
646/2011 Jätelaki. Waste Act. Suomen säädöskokoelma. <http://www.finlex.fi/fi/laki/alkup/2011/20110646>.
Karadimas, N.V., Loumos, V.G., 2008. GIS-based modelling for the estimation of municipal solid waste generation and collection. Waste Manage. Res. 26, 337–346. http://dx.doi.org/10.1177/0734242X07081484.
Lebensorger, S., Beigl, P., 2011. Municipal solid waste generation in municipalities: quantifying impacts of household structure, commercial waste and domestic fuel. Waste Manage. 31, 1907–1915. http://dx.doi.org/10.1016/j.wasman.2011.05.016.
Martin, M., Williams, I.D., Clark, M., 2006. Social, cultural and structural influences on household waste recycling: a case study. Resour. Conserv. Recy. 48 (4), 357–395. http://dx.doi.org/10.1016/j.resconrec.2005.09.005.
Navarro-Esbrí, J., Diamadopoulos, E., Ginestar, D., 2002. Time series analysis and forecasting techniques for municipal solid waste management. Resour. Conserv. Recy. 35, 201–214, PII: S0921-3449(02)00002-2.
Nuortio, T., Kytöjoki, J., Niska, N., Bräysy, O., 2006. Improved route planning and scheduling of waste collection and transport. Expert Syst. Appl. 30, 223–232. http://dx.doi.org/10.1016/j.eswa.2005.07.009.
Painter, R., Watson, V., Kheder, A., 2011. Robust statistical analysis for MSW characterization studies. J. Civ. Environ. Eng. 1, 102. http://dx.doi.org/10.4172/2165-784X.1000102.
Parfitt, J.P., Lovett, A.A., Sünnenberg, G., 2001. A classification of local authority waste collection and recycling strategies in England and Wales. Resour. Conserv. Recy. 32, 239–257, PII: S0921-3449(01)00064-7.
Rada, E.C., Grigortu, M., Ragazzi, M., Fedrizzi, P., 2010. Web oriented technologies and equipments for MSW collection. In: Proceedings of the International Conference on Risk Management, Assessment and Mitigation, pp. 150–153.
Rada, E.C., Ragazzi, M., Fedrizzi, P., 2013. WEB-GIS oriented system viability for municipal solid waste selective collection optimization in developed and transient economies. Waste Manage. 33, 785–792. http://dx.doi.org/10.1016/j.wasman.2013.01.002.
Rimaityte, I., Ruzgas, T., Denafas, G., Račys, V., Martuzevicius, D., 2012. Application and evaluation of forecasting methods for municipal solid waste generation in an Eastern-European city. Waste Manage. Res. 30, 89–98. http://dx.doi.org/10.1177/0734242X10396754.
Salhofer, S., Obersteiner, G., Schneider, F., Lebersorger, S., 2008. Potentials for the prevention of municipal solid waste. Waste Manage. 28, 245–259. http://dx.doi.org/10.1016/j.wasman.2007.02.026.
Shamshiry, E., Nadi, B., Bin Mokhtar, M., Komoo, I., Hashim, H.S., Ahya, N.Y., 2011. Forecasting generation waste using artificial neural networks. In: Proceedings of the 2011 International Conference on Artificial Intelligence, vol. 2, pp. 770–777.
Skumatz, L.A., 2008. Pay as you throw in the US: Implementation, impacts, and experience. Waste Manage. 28, 2778–2785. http://dx.doi.org/10.1016/j.wasman.2008.03.033.
Skutan, S., Brunner, P.H., 2012. Metals in RDF and other high calorific value fractions from mechanical treatment of MSW: analysis and sampling errors. Waste Manage. Res. 30 (7), 645–655. http://dx.doi.org/10.1177/0734242X12442740.
Statistics Finland, 1994. Classification of Buildings. <http://www.stat.fi/meta/luokitukset/rakennus/001-1994/koko_luokitus_en.html>.
Tchobanoglous, G., Theisen, H., Vigil, S.A., 1993. Integrated Solid Waste management. Engineering Principles and Management Issues. McGraw-Hill International Editions, Civil Engineering Series, Singapore, ISBN 0-07-112865-4.
Timlett, R., Williams, I.D., 2011. The ISB model (infrastructure, service, behaviour): a tool for waste practitioners. Waste Manage. 31, 1381–1392. http://dx.doi.org/10.1016/j.wasman.2010.12.010.
Xu, L., Gao, P., Cui, S., Liu, C., 2013. A hybrid procedure for MSW generation forecasting at multiple time scales in Xiamen City, China. Waste Manage. 33, 1324–1331. http://dx.doi.org/10.1016/j.wasman.2013.02.012.