DTU Library

Statistical analysis of solid waste composition data: Arithmetic mean, standard deviation and correlation coefficients

Edjabou, Maklawe Essonanawe; Martín-Fernández, Josep Antoni; Scheutz, Charlotte; Astrup, Thomas Fruergaard

Published in:
Waste Management

Link to article, DOI:
10.1016/j.wasman.2017.08.036

Publication date:
2017

Document Version
Peer reviewed version

Citation (APA):
Edjabou, M. E., Martín-Fernández, J. A., Scheutz, C., & Astrup, T. F. (2017). Statistical analysis of solid waste composition data: Arithmetic mean, standard deviation and correlation coefficients. *Waste Management*, *69*, 13-23. https://doi.org/10.1016/j.wasman.2017.08.036

Published in Waste Management

Statistical analysis of solid waste composition data: arithmetic mean, standard deviation and correlation coefficients

Maklawe Essonanawe Edjabou[1]*, Josep Antoni Martín-Fernández[2], Charlotte Scheutz[1], Thomas Fruergaard Astrup[1]

1) Department of Environmental Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

2) Dept. Computer Science, Applied Mathematics and Statistics, University of Girona, Campus Montilivi (P4), E-17071 Girona, Spain

*Highlights

Title of paper: Statistical analysis of solid waste composition data: Arithmetic mean, standard deviation and correlation coefficients

The core findings of the paper:

- Data for waste fraction compositions represent closed datasets that require special attention in case of statistical analysis
- Classical statistics are ill-suited to data for waste fraction compositions
- Isometric log-ratio coordinates enable appropriate transformation of waste fraction compositional data prior to statistical analysis.

Abstract

Data for fractional solid waste composition provide relative magnitudes of individual waste fractions, the percentages of which always sum to 100, thereby connecting them intrinsically. Due to this sum constraint, waste composition data represent closed data, and their interpretation and analysis require statistical methods, other than classical statistics that are suitable only for non-constrained data such as absolute values. However, the closed characteristics of waste composition data are often ignored when analysed. The results of this study showed, for example, that unavoidable animal-derived food waste amounted to 2.21±3.12% with a confidence interval of (-4.03; 8.45), which highlights the problem of the biased negative proportions. A Pearson's correlation test, applied to waste fraction generation (kg mass), indicated a positive correlation between avoidable vegetable food waste and plastic packaging. However, correlation tests applied to waste fraction compositions (percentage values) showed a negative association in this regard, thus demonstrating that statistical analyses applied to compositional waste fraction data, without addressing the closed characteristics of these data, have the potential to generate spurious or misleading results. Therefore, ¨compositional data should be transformed adequately prior to any statistical analysis, such as computing mean, standard deviation and correlation coefficients.

Keywords:

Waste composition

Compositional data analysis

Isometric log ratio

Variation array

1. Introduction

Knowledge of the individual material fractions in waste represents the basis of any waste management system planning and development (Christensen, 2011). This information is also crucial for establishing baselines and evaluating the effectivness of environmental policies. Generally, the fractional composition of waste is obtained by conducting waste fraction composition studies and is usually provided as weight percentages of selected materials such as paper, plastic, metal, food waste, etc. (Lagerkvist et al., 2011). Independently of waste characterisation methods, waste fraction composition arithmetic mean and standard deviation are usually provided (European Commission, 2004), thus ignoring the inherent structure of data for waste fraction compositions (Pawlowsky-Glahn et al., 2015). Here, the standard deviation measures the 'spread' of the estimated arithmetic mean (Reimann et al., 2008).

Waste fraction composition data are 'closed' datasets because of the limited sample space (from 0 to 100 i.e. percentages). This is known as the 'constant sum constraint' (Aitchison, 1986), where the percentage of one waste fraction depends on the ratio of the other waste fractions included in the sampled waste stream. Consequently, the percentages of waste fractions are linked to each other intrinsically. Therefore, univariate analysis (composition of waste fractions analysed separately) of waste fraction compositions is inappropriate, because it violates the fundamental assumption of independence of observations (Pawlowsky-Glahn et al., 2015). For example, Hanc et al. (2011) studied the composition of household bio-waste and reported that the yearly percentage of grass amounted to 27.6±30.8% in single-family areas. The mean was 27.6% and its standard deviation 30.8%. The resulting confidence interval (2* standard deviation) of the mean was the interval (-34.0% ; 89.2%), which covers negative percentages, although the values cannot be negative in this case. This problem is described as 'intervals covering negative proportions' (Pawlowsky-Glahn et al., 2015). An increase in the percentage of one waste fraction leads to a decrease in the percentage of another fraction and vice versa, because the sum of the percentage of individual waste fraction is fixed (Reimann et al., 2008).

Data for waste fraction compositions refer to compositional data, which arise in many fields such as geochemistry (mineral composition of rocks), medicine (blood composition) and archaeology (ceramic compositions) (Aitchison, 1994). Here, compositional data carry relative information or a ratio and add up to a constant (1 for proportion, 100 for percentage and 10^4 for ppm (parts per million)) (Aitchison, 1986; Buccianti and Pawlowsky-Glahn, 2011). As further examples, chemical compositionwaste water content, etc. also represent closed datasets (see Aitchison, 1994).

Arithmetic mean and standard deviation are based on the assumption that observations follow normal or symmetrical statistical distribution (Reimann et al., 2008). Numerous – mainly statistical-based – studies show that these estimates are affected considerably when data exhibit small deviations from normal distribution (Reimann et al., 2008; Wilcox, 2012). On the other hand, environmental data including waste fraction composition are often skewed (Reimann et al., 2008), in which case the resulting descriptive statitics may be biased and subsequently lead to wrong conclusions. Nevertheless, most waste characterisation studies report the arithmetic mean and standard deviation of waste fraction compositions, ignoring the natural structure of compositional data (e.g. Hanc et al., 2011; Edjabou et al., 2015; Naveen et al., 2016).

Despite the importance of arithmetic mean and standard deviation estimates in relation to waste composition, no attempts have been made to address the quality of these estimates.

Correlation coefficients between individual waste fractions are commonly computed in order to investigate relationships between material fractions in mixed waste (e.g. Alter, 1989; Hanc et al., 2011; Naveen et al., 2016), but they are also used to evaluate the quality and the source of elements in chemical compositions of municipal solid waste (e.g. Hanc et al., 2011; Naveen et al., 2016). An illustrative example is the correlation between food waste and packaging materials such as paper, board, plastic and metal. For example, Alter (1989) claimed that an increase in food packaging may decrease food waste occuring in housholds. In contrast, Williams et al. (2012) argued that 20 to 25% of food waste generation is due to packaging. Notwithstanding the relevance of correlation analysis applied to waste fraction compositions, the contradictory results of correlation coefficients (see Alter, 1989 and Williams, 2012) still require explanation.

Overall, computing arithmetic means, standard deviations and correlation coefficients for material fraction compositions may lead to biased results (Aitchison, 1994; Filzmoser and Hron, 2008). Additionally, uncertainty analysis (e.g. Monte Carlo analysis) of these datasets can be a source of concern when the issue of independence between material fraction compositions is either ignored or poorly addressed (Xu and Gertner, 2008).

Several studies have attempted to analyse waste composition data by applying log transformation (Chang and Davila, 2008; Dahlén et al., 2007) or log-logistic transformation (Milke et al., 2008). However, the compositional nature of waste fraction composition remains intrinsic for waste fraction composition data.

The overall aim of this paper is to demonstrate why fractional waste composition data should be transformed appropriately prior to statistical analysis. We compared some commonly encountered classical statistics applied to waste fraction compositions data and the compositional data analysis technique based on log-ratio coordinates, by analysing the fractional compositions of residual household waste in Denmark.

2 Methods and materials

2.1 Study area and waste sampling analysis

We analysed residual household waste collected from 779 single-family areas in Denmark. In these residential areas, paper, board, gardening waste, household hazardous waste, waste electrical and electronic equipment (WEEE) and bulky waste were source-segregated.

The residual household waste was generated over a one-week period, collected directly from households and kept separately for each household. Each waste bin was labelled with the address of the household from where the waste was collected. The waste bins were sealed tightly, to prevent mixing of waste during transportation to the sorting facility. Each household waste bin was weighed and sorted separately, thereby enabling us to obtain data for residual household waste for each house.

Collected residual household waste was sorted manually into the following waste fractions (Table 1): (1) avoidable vegetable food waste (AV), (2) avoidable animal-derived food waste (AA), (3) unavoidable vegetable food waste (UV), (4) unavoidable animal-derived food waste (UA), (5) paper & board (Paper or Pa), (6) plastic packaging (Plastic or Pl), (7) metal packaging (Metal or Me) and (8) other waste fractions (Others or Ot). In the present study, 'paper' refers to paper and board packaging. 'Others' refers to all other waste materials not included in the first seven waste fractions in Table 1. Avoidable food waste constitutes food and drinks that could have been eaten but instead have been disposed of. It consists of avoidable animal-derived (AA) and vegetable (AV) food waste. Unavoidable food waste is food that is not edible under normal conditions (Edjabou et al., 2016) and consists of unavoidable animal-derived (UA) and vegetable (UV) food waste. The detailed sub-fractions included in these waste fractions are presented in Table 1.

In this study, waste fraction composition represents the fractional composition of waste fractions expressed in percentage terms. Waste fraction generation rates are the mass of individual waste fractions in kg per capita per week.

Here (Table 1)

2.2 Overview of statistical analysis: classical statistical analysis

For this study, we computed (1) the arithmetic mean (Mean) of waste fraction compositions, (2) log-transformed (log-Mean), and its back-transformed (*exp(log-Mean)*) shown as Mean-log. We also computed standard deviation (SD), log-transformed (SD-log) and coefficient of variation (CV).

Noticeably, any covariance matrix has in its diagonal the variance ('var') of each variable. The sum of this diagonal, also known as the 'trace' of the matrix, is equal to total variance (Härdle and Simar, 2015) and holds in raw and log transformed of waste fraction composition datasets. Therefore, for each dataset (waste fraction compositions and log transformed), we calculated the total variance and the percentage thereof.

We also investigated the relationship between waste fractions by applying Pearson's correlation analysis to raw and log-transformed data for waste fraction compositions (in percentage) and generation rates (kg waste fraction per capita per week). However, this paper focuses mainly on the waste fraction composition dataset.

2.3 Compositional data analysis: isometric log-ratio approach

We applied statistical analysis to isometric log-ratio (ilr) coordinates, computed based on the sequential binary partition (SBP) (Egozcue et al., 2003). This approach transforms data for waste fraction compositions into an unconstrained, real dataset, thus enabling the use of classical statistics (Filzmoser and Hron, 2008). This, for example, may mean that instead of a dataset with a list of percentages that should always sum up to 100 for each observation, the isometric log-ratio transforms waste fraction composition data into a list of values that are independent and should not sum up to a constant.

Similar to classical log transformation, the isometric log-ratio requires that the data should not contain ‘zero values’. For this study, a waste ‘zero value’ means that a household did not generate any waste during this sampling week. Thus, we assumed that zero values were due to the experimental design, mainly the ‘time limit’ of the sampling campaign. For this reason, zero values were replaced, using ‘imputation based on the log-ratio expectation-maximisation (EM) algorithm’ (lrEM) in the zCompositions package (Palarea-Albaladejo and Martín-Fernández, 2015), which comprises four steps: (1) dataset selection, which can be the waste fraction composition (percentage) or generation rate (kg waste fraction per capita per week). For this study, we used the waste fraction generation rate; nevertheless, the function lrEM is based on compositional data analysis technique and therefore ensures equivalent results regardless of datasets. (2) The descriptive analysis of the zero values was performed using the function zPattern in the zCompositions package. As a result, a graphical representation of the relative frequencies of zero for each waste fraction is provided. (3) Threshold (the detection limit) values should be defined prior to zero replacement. A single value for all waste fractions or varying values can be selected. For this study, a single threshold value was set at 10 g, which is the minimum weight of the weighing scale used for the waste sampling campaign. (4) The new dataset contained non-zero values. In practice, the function lrEM substitutes an observation $\boldsymbol{x}$ with a value of zero by a random observation $\boldsymbol{y}$ in the interval between zero and the threshold value (see Palarea-Albaladejo and Martín-Fernández, 2015, for detailed mathematics underpinning zCompositions).

Seven coordinates (ilr_1) were computed corresponding to D-1 numbers of partitions. Here, D was eight, namely the number of waste fractions shown in Table 1. The first ilr coordinate was computed by dividing the eight fractions into two groups: food waste and non-food waste. Subsequently, each of the two groups was divided further until each group was represented by one single waste fraction, as indicated in Table 2, where (+1) refers to the group in the numerator, while (-1) is the group appearing in the denominator.

Here (Table 2)

The ilr coordinates were computed based on the formulas shown in Eqs. (1-7). Eq. (1) computed the coordinate (ilr_1) between food waste and non-food waste. Eqs. (2-4) computed the coordinates ilr_2 (vegetable versus animal food waste), ilr_3 (avoidable versus unavoidable vegetable food waste) and ilr_4 (avoidable versus unavoidable animal-derived food waste). Furthermore, the coordinate ilr_5 (paper and metal versus plastic and other) was calculated in Eq. (5), the coordinate ilr_6 between paper and metal was derived in Eq. (6) and the coordinate ilr_7 between plastic and other in Eq. (7).

$$\mathrm{ilr}_1\{\mathrm{AV}, \mathrm{UV}, \mathrm{AA}, \mathrm{UA}\}\mathrm{vs.}\{\mathrm{Pa}, \mathrm{Me}, \mathrm{Pl}, \mathrm{Ot}\} = \sqrt{\frac{4\times4}{4+4}}\,\mathrm{LN}\,\frac{\sqrt[4]{\mathrm{AV}\times\mathrm{UA}\times\mathrm{AA}\times\mathrm{UA}}}{\sqrt[4]{\mathrm{Pa}\times\mathrm{Me}\times\mathrm{Pl}\times\mathrm{Ot}}} \quad (1)$$

$$\mathrm{ilr}_2\{\mathrm{AV}, \mathrm{UV}\}\mathrm{vs.}\{\mathrm{AA}, \mathrm{UA}\} = \sqrt{\frac{2\times2}{2+2}}\,\mathrm{LN}\,\frac{\sqrt[2]{\mathrm{AV}\times\mathrm{UV}}}{\sqrt[2]{\mathrm{AA}\times\mathrm{UA}}} \quad (2)$$

$$\mathrm{ilr}_3\{\mathrm{AV}\}\mathrm{vs.}\{\mathrm{UV}\} = \sqrt{\frac{1\times1}{1+1}}\,\mathrm{LN}\,\frac{\sqrt[1]{\mathrm{AV}}}{\sqrt[1]{\mathrm{UV}}} \quad (3)$$

$$\mathrm{ilr}_4\{\mathrm{AA}\}\mathrm{vs.}\{\mathrm{UA}\} = \sqrt{\frac{1\times1}{1+1}}\,\mathrm{LN}\,\frac{\sqrt[1]{\mathrm{AA}}}{\sqrt[1]{\mathrm{UA}}} \quad (4)$$

$$\mathrm{ilr}_5\{\mathrm{Pa}, \mathrm{Me}\}\mathrm{vs.}\{\mathrm{Pl}, \mathrm{Ot}\} = \sqrt{\frac{2\times2}{2+2}}\,\mathrm{LN}\,\frac{\sqrt[2]{\mathrm{Pa}\times\mathrm{Me}}}{\sqrt[2]{\mathrm{Pl}\times\mathrm{Ot}}} \quad (5)$$

$$\mathrm{ilr}_6\{\mathrm{Pa}\}\mathrm{vs.}\{\mathrm{Me}\} = \sqrt{\frac{1\times1}{1+1}}\,\mathrm{LN}\,\frac{\sqrt[1]{\mathrm{Pa}}}{\sqrt[1]{\mathrm{Me}}} \quad (6)$$

$$ilr_7\{Pl\}vs.\{Ot\} = \sqrt{\frac{1\times1}{1+1}}\,LN\,\frac{\sqrt[1]{Pl}}{\sqrt[1]{Ot}} \quad (7)$$

Here, LN stands for the natural logarithm, and the other abbreviations refer to the waste fractions presented in Table 1. Pa refers to paper and board, Pl to plastic packaging, Me to metal packaging and Ot to other.

The CoDa technique uses the geometric mean of the dataset, which is the 'back-transformed' value of the ilr-arithmetic mean and is calculated as follows:

$$g_m(x) = \left[\prod_{i=1}^{D} x_i\right]^{1/D} = exp\left[\frac{1}{D}\sum_{i=1}^{D} LN(x_i)\right] \quad (8)$$

where $g_m(x)$ is the geometric mean and D is the number of waste fractions (x_i) involved. The natural logarithm is abbreviated as $LN(x_i)$ and its inverse is abbreviated as $exp(x_i)$.

The back transformation of the isometric log-ratio coordinates is calculated simply by reversing the original transformation (Egozcue et al., 2003). The general formula for the back transformation of the isometric log-ratio coordinate (ilr^{-1}) is provided as follows (Felipe et al., 2016):

$$ilr^{-1} = C(\exp(x \cdot \psi)) \quad (9)$$

where ilr^{-1} is the back transformation, $\mathbf{x}$ is the simulated value for the transformation (ilr), ψ is the matrix constructed from the sequential binary partition given in Eqs (1 to 7) and C is the closure operation that provides a closed dataset.

Total variance ($totvar(\mathbf{x})$) is introduced to provide a global measure of spread (Pawlowsky et al., 2008) and measures the variation between individual waste fraction compositions included in the dataset. Total variance is computed as:

$$totvar(\boldsymbol{x}) = \frac{1}{D}\sum_{i=1}^{D-1} \sum_{j=i+1}^{D} var\left(LN\frac{x_i}{x_j}\right) \quad (10)$$

The relationship between pairs of waste fractions is analysed by means of a variation array, calculated as:

$$A = \begin{bmatrix} 0 & v_{12} & \cdots & v_{1D} \\ e_{21} & 0 & \cdots & v_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ e_{D1} & e_{D2} & \cdots & 0 \end{bmatrix} \quad (11)$$

where,

$$e_{ij} = E\left(\ln\frac{x_i}{x_j}\right) \quad (12) \qquad \text{and} \qquad v_{ij} = var\left(\ln\frac{x_i}{x_j}\right) \quad (13)$$

The variation array (Aitchison, 1986) was introduced to provide a solution to the problem of computing correlation coefficients for compositional data. We computed the variation array using both waste fraction compositions and generation rates.

2.4 Software for data analysis

First, the data were explored and zero values imputed using the R package 'zCompositions' (Palarea-Albaladejo and Martín-Fernández, 2015). The ilr coordinates and their back transformation, as well as variation array, were computed with CoDaPack (Thió-Henestrosa and Comas-Cufi, 2011). Thereafter, the most commonly used methods employed for describing and analysing waste data, such as mean, standard deviation, coefficients of variation and correlation tests (European Commission, 2004), were carried out in R (R Core Team, 2017). Among other packages implemented in R, the 'StatDA' (Filzmoser, 2015) software package was used for plotting.

3 Results

3.1 Exploration of data for waste fraction compositions

Figure 1 displays the graphical output of the zero values analysis. The columns show the analysis of zero values by waste fraction. The data in Figure 1 can be grouped into two parts. The first is a rectangle, containing squared boxes coloured in dark grey, where waste fractions have zero values, and light grey for non-zero values. The number of squared boxes per column is the total combinations of zero values for each household involved as a function of waste fraction. The second is bar plots on the top (in dark grey), which show the percentage frequency of zero values by waste fraction, whereas bar plots on the right (in light grey) present the percentage frequency of non-zero values for all possible combinations of household and waste fractions. For example (see bar plots on the top in dark grey), the percentage frequency of zero was 5.35% for avoidable vegetable food waste (see first column), and 2.94% for unavoidable food waste (see second column). Regarding bar plots on the right-hand side of the rectangle (in light grey), 64.45% of observations (households) have non-zero values for all waste fractions (first line), and 8.31% are non-zero values, except for the avoidable animal derived-food waste fraction.

Here (Figure 1)

Subsequently, the zero value detected was replaced prior to computing the log-ratio coordinates and undertaking normal log transformation. For example, the minimum values for the four food waste fractions (zero values) were replaced by 5.7 g for avoidable vegetable food waste, 5.8 g for unavoidable vegetable food waste, 2.8 g for avoidable animal-derived food waste and 1.6 g for unavoidable animal-derived food waste. Note that here the replaced values are between zero and 10 g. A comparison of the datasets before and after zero replacement showed quite a similar distribution, demonstrating that the distribution of the dataset is preserved despite containing many zero values (SM Figure 1, SM Tables 2 and 3).

Figure 1 also presents a detailed overview of household waste fraction generation patterns; for example, only 1.3% and 0.3% of the households did not generate plastic packaging or paper, respectively. Noticeably, for vegetable food waste, only 5.2% and 2.9% of the households (see Figure 1, vertical bars) did not generate AV and UV, respectively. On the other hand, the percentage of households that did not generate animal-derived food waste was 15.2% for AA and 14.6% for AU (see Figure 1, vertical bars). These data indicate that vegetable food waste occurred more often than animal-derived food in Danish houses.

3.2 Mean and standard deviation of waste fraction compositions

The distribution of the waste fraction compositions for all households is shown in Figure 2. Asymmetry is evident in the boxplot of each waste fraction, because the distance from the median (horizontal bar in the rectangular box) to the fifth percentiles (bottom horizontal bar (Figures 2 and 4) or vertical bar on the left (Figure 3)) is smaller than the distance between the median to the 95th percentiles (upper horizontal bar (Figures 2 and 4) or vertical bar on the right (Figure 3)), as shown in Figure 2. Thus, the data for each waste fraction were positively skewed and also contained potential outliers, which are defined as unusually large or small values in a sample of observation (Wilcox, 2012). Here, outliers are shown in Figure 3 as circles above the upper horizontal bar, and these outliers lead to bias in the arithmetic mean and inflate the standard error. Thus, robust statistical techniques have been developed to deal effectively with this problem, though these methods are not included in this study.

A detailed analysis of vegetable food waste (AV and UV) is provided in Figure 3 as an example. Figures 3a and 3b illustrate a combined histogram and boxplot of waste fraction composition and log transformation for avoidable vegetable food waste, while Figures 3c and 3d represent unavoidable vegetable food waste in the same regard. These figures reveal asymmetric distribution despite log transformation. Conversely, the ilr coordinates are distributed symmetrically (see Figure 4).

Here (Figure 2)

Here (Table 3)

Here (Figure 3)

The arithmetic means (Mean) based on waste fraction compositions sum up to 100, whereas the arithmetic means based on log-transformed (Log-mean) data sum up to 14. As a result, the means of the log-transformed data are difficult to interpret and apply because of the change in scale (USEPA, 2006). This problem could be solved by Mean-log', which is obtained by 'back transforming' the log-transformed mean (Mean-log=*exp(Log-Mean-log)*). The arithmetic mean, log-mean and mean-log were computed from an asymmetric dataset, which led to biased parameter estimation and incorrect results (Reimann et al., 2008; Wilcox, 2012).

On the contrary, the 'Mean-ilr' (mean based on isometric log-ratio coordinates) (see Table 3) was computed from symmetrical data, thus suggesting that the log-ratio coordinates enable a data analyst to obtain symmetric distribution of data, as shown in Figure 4. Importantly, while log-ratio transformation enables one to remove the constant sum constraint, the ‘Mean-ilr’ for waste fractions sums up to 100. Similar to classical statistics, robust methods have been developed for the statistical analysis of compositional data (Templ et al., 2011), though these methods are not included in this study.

Here (Figure 4)

The standard deviation, total variance and percentage of variance estimates were calculated and are shown in Table 4. The results indicate that the standard deviation values for the raw waste fraction composition are very high compared to their corresponding arithmetic mean (Mean in Table 3). In particular, the standard deviation of animal-derived food waste (AA and AV) and metal packaging are higher or equal to the corresponding arithmetic mean, thereby generating very high variation value coefficients (e.g. 155% for metal packaging, 141% for unavoidable animal-derived food waste, 99% for avoidable animal-derived food waste). The resulting confidence intervals (Mean ± 2* SD) were (-6.78; 20.74) and

(-4.03; 8.45) for AA and AV, respectively, including negative percentages. These results highlight some of the pitfalls of computing standard deviations for waste fraction compositions. Moreover, the estimated percentages of variances for waste fractions varied when the raw dataset for waste fraction compositions (% Var) was log-transformed (% Var-log). The highest variance percentages were found for the fractions other (%Var= 31.43%) and avoidable animal-derived food waste (%Var-log=33.24%) in raw and log-transformed datasets, respectively. On the other hand, the lowest variance percentages were found for unavoidable animal-derived food waste (%Var=1.47%) and other (%Var-log=2.74%) in the raw and log-transformed datasets, correspondingly. These incoherent results indicate that while log transformation could indeed help to achieve normality, the calculated variance becomes impossible to compare after transformation, as demonstrated by Filzmoser et al. (2009).

Overall, it is questionable whether standard deviation values are informative in the case of most sets of waste composition data, due to the dual issues of non-normality and the constant-sum constraint. First, the standard deviation ignores the compositional nature of waste fraction composition data (composition of waste fractions should add up to 100). Second, most coefficients of variation (CV %) provided in Table 4 are extremely high, thus restricting their application in environmental modelling (Ciroth et al., 2013). As a solution, total variance (see Eq. 9) that measures overall data homogeneity (or variation) can be calculated (Pawlowsky et al., 2008). Here, total variance expresses variation in the dataset for each waste fraction. Thus, the contribution of each waste fraction to total variation is provided in percentage terms (clr-Var %), as shown in Table 4.

Here (Table 4)

Based on the compositional data analysis technique, total variance (totvar) from Eq. (9) amounted to 9.25, as shown in Table 4. The waste fraction contributing to the highest variation in the dataset was avoidable animal-derived food waste (24.73%), followed by unavoidable animal-derived food waste (18.84%) and metal packaging (14.81%), suggesting that the generation of these fractions by Danish households varies substantially.

On the other hand, paper (5.27%) and plastic packaging (5.53%) made a small contribution to total variance. A possible interpretation for this finding could be that metal packaging materials are source-sorted by a wider variety of households than paper and plastic packaging, and therefore they do not vary much in the fraction that ends up in residual household waste bins. However, a characterisation of total household waste including source-segregated waste (e.g. paper, metal, plastic) could provide a better interpretation of these results, thereby demonstrating that total variance enables the analyst to compare systematically variations among waste fraction compositions, which is difficult for classical standard deviation and coefficient of variation estimates.

3.3 Relationship between waste fractions: Pearson's correlation test

Table 5 presents the pairwise correlation coefficients between waste fractions, computed using datasets of (1) percentage composition (Percentage %) and (2) generation rates (kg/capita/week). A negative correlation coefficient between waste fractions means an inverse relationship, whereas a positive correlation coefficient means these fractions vary in the same direction (when the value of one waste fraction increases, the value of the other fraction increases too, and vice versa). Moreover, while a correlation coefficient of value ±0.5 shows a strong relationship between the two waste fractions, a value of 1 means a perfect correlation. A correlation coefficient is statistically significant when the p-value is less than 0.5.

Here (Table 5)

Based on the waste fraction generation rates, we found a positive and significant correlation coefficient between ‘Other’ and the seven remaining waste fractions, as shown in Table 5. In contrast, we found negative and significant correlation coefficients between these fractions when the Pearson’s correlation test was applied to waste fraction compositions (Percentage %).

Figure 5 illustrates the results of the correlation test applied to waste fraction composition data. Figure 5 shows that the Pearson’s correlation test applied to the waste fraction generation dataset provided a positive correlation coefficient between avoidable food waste (UA, UV, AA and AV) and plastic packaging. These results are consistent with those of Williams et al. (2012), suggesting that a reduction in plastic packaging materials may lead to a reduction in avoidable vegetable food waste. In contrast, the results of the Pearson’s correlation applied to the waste fraction compositions dataset showed a negative correlation between the same waste fractions, except for UA. These results are in good agreement with those obtained by Alter (1989), and similar results were obtained when the Pearson’s correlation test was applied to log-transformed data. Note here that the signs and the values of the correlation coefficients depend on the datasets, even though a Pearson’s correlation test was applied to log-transformed data (SM Table 1). These results pose an interpretation dilemma. First, a reduction in plastic packaging may contribute to food waste reduction, due to the positive correlation between these waste fractions, although, on the other hand, an increase in the use of plastic packaging may contribute to a reduction in household food waste because of the negative correlation coefficient. Moreover, while these correlation coefficients were statistically significant, their estimates were somewhat different (see Figure 4 and Table 5).

Here (Figure 5)

3.4 Variation array with CoDa

The variation array was computed using Eq. (10) and is shown in Table 6. Note that the same variation array was obtained when using either the waste fractions generation rates (kg/capita/week) or waste fraction compositions (percentage %), and therefore the relationship between waste fractions is interpreted independently of waste datasets.

The variation array is divided into two triangles. The upper triangle shows ratios or proportionalities between waste fractions as pairwise log-ratio variances (variance $\ln(X_i/X_j)$ (see Eq. (12)). The lower triangle presents the pairwise log-ratio means (Mean $\ln(X_j/X_i)$ (see Eq. (13)). Here, the numerator is denoted by columns (X_i), and denominator (X_j) is illustrated by rows. Moreover, the sign (+ or −) of the log-ratio mean values indicates the direction of the ratio between the relevant fractions.

Here (Table 6)

Log-ratio variance values highlighted in grey (the value is close to zero) indicate an almost constant ratio, whereas log-ratio variance values in bold and highlighted in grey (usually value is closed to zero) can be assumed to be zero, suggesting an absolutely constant ratio (Pawlowsky-Glahn et al., 2015). On the other hand, log-ratio variance values that are very much higher than zero are highlighted in red, and these indicate no relationship between the two relevant fractions, because their ratios vary significantly.

For example, the mean log-ratio between plastic packaging and paper and board was negative *{(mean (log(Plastic/Paper))= -1.4)}* (here, *Plastic* is X_j from a row in Table 6 and *Paper* is X_i from a column in Table 6), indicating that the households placed more mass of plastic packaging than paper and board waste into their residual waste bins. Furthermore, the variance in their log-ratio is small (0.77), suggesting a strong relationship between these fractions. This relationship has a negative ratio, which can be calculated as follows:

plastic/paper=exp(-1.4)=0.25

This result suggests that the ratio between discarded (1) plastic and (2) paper and board in residual household waste is constant and estimated at 0.25. This information could be used for future developments in waste generation, i.e. to identify the effects of new regulations and policies addressing packaging materials.

The results shown in Table 6 indicate that the mean log-ratio between avoidable animal-derived food waste and unavoidable vegetable food waste was negative (-1.35). However, the variance in their log-ratio was high (4.21), thereby suggesting that the compositions of these fractions are not proportional. In this case, the ratio between these fractions is not constant.

Overall, the compositions of these pairs of waste fractions are highly dependent: (1) unavoidable vegetable food waste (UV) and paper (Paper), (2) paper (Paper) and plastic packaging (Plastic) and (3) plastic packaging (Plastic) and other waste fractions (Other). However, no relationship between avoidable food waste fractions (AV and AA) and material packaging (paper, plastic and metal) was identified. For example, from the results in Table 7, it is apparent that the ratio between avoidable animal-derived food waste and packaging materials (plastic, paper and metal) is highly variable (very high log-ratio variance painted in red). Similarly, the ratio between avoidable vegetable food waste and packaging materials (plastic, paper and metal) is not constant. These values indicate no constant ratios between these fractions, signifying that there is no relationship between these fractions based on the analysis of residual waste taken from the 779 households.

4. Discussion

From the data in Table 3, arithmetic means of waste fractions composition were influenced by the fact that the assumption of normal distribution was violated (see Figure 4). These results are consistent with previously published studies, which concluded that the arithmetic mean is an inappropriate means of estimating central values of compositional data (Filzmoser et al., 2009; Pawlowsky-Glahn et al., 2015; van den Boogaart et al., 2013). Consequently, any evaluation (e.g. prevention, reduction or recycling of waste) or modelling (e.g. life cycle assessment) based on the arithmetic mean of waste fraction composition may lead to potentially wrong conclusions, because they are based on erroneous estimates. While the log transformation of waste composition may help solve the problem of normality, its value is limited because it relies on a univariate method, which ignores that the compositions of waste fractions account for the limited data, i.e. from 0 to 100.

The results from the variation array were not in agreement with those from the Pearson’s correlation tests applied to both raw and log-transformed data. The correlation test applied to waste fraction generation rates provided positive correlation coefficients. On the other hand, negative correlation coefficients were obtained when the correlation analysis was applied to the composition of waste fractions in percentage terms. The positive correlation coefficients were due to the size of the mass effect of waste fractions (kg/capita/week), explaining why most waste fractions are positively and significantly correlated with each other. The size effect of mass was removed by calculating the correlation coefficient based on the percentage composition of waste fractions. This then generated negative correlation coefficients because of the constant sum constraint (Aitchison, 1986; Pearson, 1897). As a solution, the relationship between food waste fractions and material packaging can be evaluated by the variation array through a compositional data analysis technique. Log-ratio coordinates remove the constant sum constraint and enable the determination of the relationship between waste fractions, independently of the unit. Another advantage of the variation array is that the pairwise ratio between waste fractions could be back-transformed to a desired unit and adequately interpreted while preserving the structure of the original data (Filzmoser and Hron, 2008; Pawlowsky-Glahn et al., 2015). The advantage in this approach is that the variation array of both waste datasets (percentage composition and mass per waste fraction per household) generates the same results because of the log-ratio transformation.

Computing the arithmetic mean (mean-ilr), total variance and variance array based on CoDa technique is a not straightforward undertaking. However, numerous tools that do not require advanced programming skills are freely available (Templ et al., 2011; Thió-Henestrosa and Comas-Cufi, 2011; van den Boogaart, 2008). Therefore, we urge practitioners and researchers within solid waste management to address adequately the constant sum constraint problem when analysing solid waste composition data (Filzmoser et al., 2009).

5. Conclusions

This study is a first attempt to address the problem associated with the statistical analysis of waste fraction composition data. Based on a systematic comparison of the arithmetic mean and standard deviation applied to waste fraction composition data, it was demonstrated that these statistical parameters may generate erroneous and misleading results when applied to fractional percentages (i.e. percentage of paper, board, food waste, etc.). Moreover, correlation coefficients based on raw or general transformation of data depend strongly on the type of waste dataset. As a solution, isometric log-ratio coordinates approximate the symmetrical distribution of data and remove the total constant sum constraint, which restricts the application of classical statistics to waste fraction composition. As a result, statistical analysis applied to log-ratio coordinates generates consistent results independently of the selected data type. The arithmetic means of waste fractions, based on the isometric log-ratio, summed up to 100. The variation array provides a ratio between waste fractions and offers consistent conclusions regardless of the data type.

Acknowledgments

The authors acknowledge the Danish Strategic Research Council for financing this study via the IRMAR (Integrated Resource Management & Recovery) Project (No. 11-116775). We wish to thank Compositional Data Analysis and Related methods (CoDa- RETOS) for their support through the project MINISTERIO ECONOMÍA Y COMPETIVIDAD (Ref: MTM2015-65016-C2-1-R (MINECO/FEDER, UE). The Danish Environmental Protection Agency (EPA) and Claus Petersen from ECONET AS are also acknowledged for their valuable support and contributions.

References

Aitchison, J., 1994. A Concise Guide to Compositional Data

Analysis. Lect. Notes-Monograph Ser. 24, 73–81.

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK).Reprinted (2003) with additional material by The Blackburn Press, Caldwell, NJ.

Alter, H., 1989. The Origins of Municipal Solid Waste: the Relations Between Residues From Packaging Materials and Food. Waste Management 7, 103–114. doi:10.1177/0734242X8900700115.

Buccianti, A., Pawlowsky-Glahn, V., 2011. Compositional Data Analysis, Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd, Chichester, UK. doi:10.1002/9781119976462.

Chang, N.-B., Davila, E., 2008. Municipal solid waste characterizations and management strategies for the Lower Rio Grande Valley, Texas. Waste Management 28, 776–94. doi:10.1016/j.wasman.2007.04.002.

Christensen, T.H., 2011. Introduction to Waste Management. in Solid Waste Technology and Management 3–16, in: Christensen, T.H. (Ed.), Solid Waste Technology & Management, Volume 1 & 2. Chichester, UK. doi:10.1002/9780470666883.ch1.

Ciroth, A., Muller, S., Weidema, B., Lesage, P., 2013. Empirically based uncertainty factors for the pedigree matrix in ecoinvent. International Journalof Life Cycle Assessment 1–11. doi:10.1007/s11367-013-0670-5.

Dahlén, L., Vukicevic, S., Meijer, J.-E., Lagerkvist, A., 2007. Comparison of different collection systems for sorted household waste in Sweden. Waste Management 27, 1298–305. doi:10.1016/j.wasman.2006.06.016.

Edjabou, M.E., Boldrin, A., Scheutz, C., Astrup, T.F., 2015. Source segregation of food waste in office areas: Factors affecting waste generation rates and quality. Waste Management 46, 94–102. doi:10.1016/j.wasman.2015.07.013.

Edjabou, M.E., Petersen, C., Scheutz, C., Astrup, T.F., 2016. Food waste generation and composition in Danish households. Waste Management 52, 256–268. doi:10.1007/s13398-014-0173-7.2.

Egozcue, J.J., Pawlowsky Glahn, V., Mateu-Figueras, G., Barceló Vidal, C., 2003. Isometric logratio for compositional data analysis. Mathematical Geosciences 35, 279–300. doi:10.1023/A:1023818214614.

European Commission, 2004. Methodology for the Analysis of Solid Waste (SWA-tool). User Version 43, 1–57.

Filzmoser, P., Hron, K., 2008. Correlation Analysis for Compositional Data. Mathematical Geosciences. 41, 905–919. doi:10.1007/s11004-008-9196-y.

Filzmoser, P., Hron, K., Reimann, C., 2009. Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. Science of the Total Environment407, 6100–6108. doi:10.1016/j.scitotenv.2009.08.008.

Peter Filzmoser (2015). StatDA: Statistical Analysis for Environmental Data. R package version 1.6.9. https://CRAN.R-project.org/package=StatDA.

Hanc, A., Novak, P., Dvorak, M., Habart, J., Svehla, P., 2011. Composition and parameters of household bio-waste in four seasons. Waste Management 31, 1450–60. doi:10.1016/j.wasman.2011.02.016.

Härdle, W., Simar, L., 2015. Applied Multivariate Statistical Analysis Applied Multivariate Statistical Analysis, Technometrics. doi:10.1198/tech.2005.s319

Lagerkvist, A., Ecke, H., Christensen, T.H., 2011. Waste Generation and Characterization Waste Characterization : Approaches and Methods, in: Christensen, T.H. (Ed.), Solid Waste Technology & Management, Volume 1 & 2. Chichester, UK. doi:10.1002/9780470666883.ch5.

Milke, M.W., Wong, V., McBean, E. a., 2008. Describing variability of MSW composition data with the log-logistic distribution. Waste ManagementRes. 26, 355–361. doi:10.1177/0734242X08089464.

Naveen, B.P., Mahapatra, D.M., Sitharam, T.G., Sivapullaiah, P.V., Ramachandra, T.V., 2016. Physico-chemical and biological characterization of urban municipal landfill leachate. Environ. Pollut. doi:10.1016/j.envpol.2016.09.002.

Palarea-Albaladejo, J., Martín-Fernández, J.A., 2015. ZCompositions - R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems. 143, 85–96. doi:10.1016/j.chemolab.2015.02.019.

Pawlowsky-Glahn, V., Egozcue, J., Tolosana-Delgado, R., 2015. Modeling and Analysis of Compositional Data. John Wiley & Sons.

Pawlowsky, G. V, Egozcue, J., Delgado, R.T., 2008. Lecture notes on compositional data analysis 1–100.

Pearson, K., 1897. Mathematical Contributions to the Theory of Evolution. - On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs. Proc. R. Soc. London 60, 489–498. doi:10.1098/rspl.1896.0076.

R Core Team, 2017. R: A Language and Environment for Statistical Computing.

Reimann, C., Filzmoser, P., Garrett, R.G., Dutter, R., 2008. Statistical Data Analysis Explained: Applied Environmental Statistics With R, Statistical Data Analysis Explained: Applied Environmental Statistics With R. John Wiley & Sons, Ltd, Chichester, UK. doi:10.1002/9780470987605.

Templ, M., Hron, K., Filzmoser, P., 2011. robCompositions: an R-package for robust statistical analysis of compositional data, in: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis. Theory and Applications. John Wiley & Sons, Chichester (UK), pp. 341–355.

Thió-Henestrosa, S., Comas-Cufi, M., 2011. CoDaPack v2 USER ' s GUIDE.

USEPA, 2006. Data Quality Assessment: Statistical Methods for Practitioners, U.S. Envir. ed. U.S. Environmental Protection Agency, Washigton, D.C.

van den Boogaart, K.G., 2008. "compositions": A unified R package to analyze compositional data. Comput. Geosci. 34, 320–338.

van den Boogaart, K.G., Tolosana-Delgado, R., Boogaart, K.G. Van Den, Tolosana-Delgado, R., 2013. Analyzing Compositional Data with R. Springer Berlin Heidelberg, Berlin, Heidelberg. doi:10.1007/978-3-642-36809-7.

Wilcox, R., 2012. Introduction to Robust Estimation and Hypothesis Testing, Introduction to Robust Estimation and Hypothesis Testing. Elsevier. doi:10.1016/B978-0-12-386983-8.00008-1.

Williams, H., Wikström, F., Otterbring, T., Löfgren, M., Gustafsson, A., 2012. Reasons for household food waste with special attention to packaging. Journal of Cleaner Production. 24, 141–148. doi:10.1016/j.jclepro.2011.11.044.

Xu, C., Gertner, G.Z., 2008. Uncertainty and sensitivity analysis for models with correlated parameters. Reliability Engineering and System Safety 93, 1563–1573.

doi:10.1016/j.ress.2007.06.003.

Table 1: List of residual waste fractions and components included

Waste fractions	Components
Avoidable vegetable food waste (AV[1])	Cooked food (e.g. rice, pasta, potatoes, etc.) Fresh fruit, fresh carrots and potatoes, bread, cereals
Avoidable animal-derived food waste (AA[1])	Cooked eggs, rest of food containing meat, fish, etc. Canned meat and fish,
Unavoidable vegetable food waste (UV[1])	Residues from fruits, vegetables, coffee grounds Eggs not cooked, dairy products, not cooked meat and fish, etc.
Unavoidable animal-derived food waste (UA[1])	Leftovers containing meat, fish, skins and bones, etc. Cheese rinds, eggs shells, other non-edible mixed animal and vegetable products
Paper and board (Paper:Pa[1])	Advertisements , Books & booklets, Magazines & Journals, Newspapers Office paper, Phonebooks, Miscellaneous paper, Corrugated boxes Beverage cartons, Folding boxes, Miscellaneous board
Plastic packaging (Plastic:Pl[1])	Packaging plastics, such as PET/PETE, HDPE, PVC/V , LDPE/LLDPE, PP, PS, others, etc
Metal packaging (Metal;Me[1])	Metal packaging containers (ferrous and non-ferrous) Composites
Others (Ot[1])	Gardening waste, glass packaging, other/special glass, Table and kitchen ware glass, Non-packaging metals Non-packaging plastic, plastic film Miscellaneous combustible waste, inert (other non-combustible), special waste

[1] *Refers to abbreviation of waste fractions in equations and figures and other tables in the present paper*

Table 2: Signs code of the sequential binary partition applied to the residual household waste fractions: Balance code, (+1) means that the fraction is assigned to the first group (numerator), (-1) to the second group, and 0 the fraction is not included in the partition in this balance

Coordinates	**Residual household waste fractions**							
	AV[a]	UV[b]	AA[c]	UA[d]	Paper[e]	Metal[f]	Plastic[g]	Other[h]
ilr_1	+1	+1	+1	+1	-1	-1	-1	-1
Ilr_2	+1	+1	-1	-1	0	0	0	0
Ilr_3	+1	-1	0	0	0	0	0	0
Ilr_4	0	0	+1	-1	0	0	0	0
Ilr_5	0	0	0	0	+1	+1	-1	-1
Ilr_6	0	0	0	0	+1	-1	0	0
Ilr_7	0	0	0	0	0	0	+1	-1

[a]*Avoidable vegetable food waste*
[b]*Unavoidable vegetable food waste*
[c]*Avoidable animal-derived food waste*
[d]*Unavoidable animal-derived food waste*
[e]*Paper and board;* [f]*Metal packagin.;* [g]*Plastic packaging;* [h]*grouped waste fraction (see Table 1 for waste fractions)*

.

Table 3: Comparison of arithmetic means computed based on raw data (Mean), log transformed data (Log-Mean), back-transformed data (Mean-log) and back-transformed isometric log-ratio mean (Mean-ilr)

Waste fractions	Classical statistics				CoDa-technique
	Mean[a]	Log-mean[b]	Mean-log[c]	Median	Mean-ilr[d]
Avoidable vegetable food waste	15.57	2.32	10.14	13.84	13.3
Unavoidable vegetable food waste	17.03	2.47	11.87	15.22	15.5
Avoidable animal-derived food waste	6.98	1.13	3.09	5.11	4.0
Unavoidable animal-derived food waste	2.21	-0.06	0.94	1.08	1.2
Paper and board	20.79	2.91	18.28	18.52	23.9
Metal packaging	2.12	0.09	1.09	1.44	1.4
Plastic packaging	5.51	1.50	4.49	4.76	5.9
Other	29.80	3.28	26.59	26.30	34.8
Total	100.00	13.63	76.49	86.27	100.0
Wet waste kg per household per week	10.41		8.80	9.52	
Wet waste kg per person per week	4.00		3.45	3.42	

[a]*Arithmetic mean from raw data,*
[b]*Arithmetic mean for log-transformed data;*
[c]*Arithmetic mean based on back-transformation of the log-transformed data;*
[d]*Arithmetic mean for ilr coordinates, which is back-transformed*

Table 4 Comparison of standard deviation values based on waste fraction compositions data set (SD) and variance (% Var); log-transformed (SD-log) and variance of log-transformed (% Var-log) absolute contribution of each waste fractions (SD-clr) to total variance, and the percentage distribution of the total variance (SD-clr) (n=779)

Waste fractions	Classical statistics				CoDa-technique	
	SD	%Var	SD-log	%Var-log	SD-clr	%Var-clr
Avoidablevegetablefoodwaste	10.76	17.52	3.49	12.55	1.1	13.16
Unavoidablevegetablefoodwaste	11.51	20.05	2.99	9.21	1.03	11.56
Avoidableanimal-derivedfoodwaste	6.88	7.16	5.68	33.24	1.51	24.73
Unavoidableanimal-derivedfoodwaste	3.12	1.47	4.46	20.5	1.32	18.84
Paperandboard	10.9	17.98	1.68	2.91	0.7	5.27
Metalpackaging	3.29	1.64	3.76	14.57	1.17	14.81
Plasticpackaging	4.26	2.75	2.04	4.29	0.72	5.53
Other	14.41	31.43	1.63	2.74	0.75	6.09
Totalvariance(totvar)	660.76	100.00	97.05	100.00	9.23	100.00

Table 5 Correlation matrix from Pearson correlation test and significance levels of raw data shown in Figure 2 (r: range:-1.00 to +1.00)

Waste fractions	AV[d]	UV[e]	AA[f]	UA[g]	Paper[h]	Metal [i]	Plastic [j]	Other	Datasets
Avoidable vegetable food waste	1.00	***	***	***	***	.	***	***	Percentage %
(AV)	1.00	***	***	**	***	.	***	***	kg/capita/week
Unavoidable vegetable food waste	-0.17	1.00	***	0.00	***	*	**	***	Percentage %
(UV)	0.23	1.00	***	***	***	*	*	***	kg/capita/week
Avoidable animal-derived food waste	0.16	-0.19	1.00	0.00	***	0.00	0.00	***	Percentage %
(AA)	0.46	0.14	1.00	***	***	0.00	**	***	kg/capita/week
Unavoidable animal-derived food	-0.12	0.02	0.00	1.00	.	0.00	0.00	**	Percentage %
Waste (UA)	0.11	0.17	0.14	1.00	*	*	.	*	kg/capita/week
Paper and board	-0.30	-0.16	-0.21	-0.06	1.00	*	0.00	***	Percentage %
	0.29	0.19	0.18	0.07	1.00	0.00	**	***	kg/capita/week
Metal packaging	-0.07	-0.08	-0.03	0.03	-0.09	1.00	0.00	0.00	Percentage %
	0.07	0.08	0.04	0.07	0.04	1.00	0.00	***	kg/capita/week
Plastic packaging	-0.13	-0.10	-0.05	0.03	-0.04	0.05	1.00	*	Percentage %
	0.13	0.09	0.10	0.06	0.11	0.04	1.00	***	kg/capita/week
Other	-0.38	-0.41	-0.27	-0.10	-0.26	-0.06	-0.08	1.00	Percentage %
	0.30	0.15	0.21	0.07	0.28	0.14	0.14	1.00	kg/capita/week

****** *Very high significance probability higher than 0.001*

**** *High significance probability between 0.001 and 0.01*

** *Significance probability between 0.01 and 0.05*

0.00 no significance-probability higher than 0.05

[a] *amount of waste (wet basis) per household per week*

[b] *amount of waste (wet basis) per person per week*

[c] *Composition of residual household waste on wet basis.*

[d] *Avoidable vegetable food waste*

[e] *Unavoidable vegetable food waste*

[f] *Avoidable animal-derived food waste*

[g] *Unavoidable animal-derived food waste*

[h] *Paper;* [i] *Metal packaging.;* [j] *Plastic packaging;* [k] *Other (see Table 1).*

Table 6: Variation array of waste fraction compositions computed using log-ratio transformation of the waste dataset shown in Figure 2

Waste fractions						Variance ln(Xi/Xj)		
	AV[a]	UV[b]	AA[c]	UA[d]	Paper[e]	Metal[f]	Plastic[g]	Other[h]
AV[a]		2.53	3.11	3.83	2.10	3.09	2.15	2.18
UV[b]	0.16		4.21	3.00	1.52	2.93	1.77	1.83
AA[c]	-1.19	-1.35		5.14	3.54	4.49	3.43	3.62
UA[d]	-2.38	-2.54	-1.19		2.49	3.63	2.50	2.61
Paper[e]	0.59	0.43	1.78	2.97		2.08	0.77	0.64
Metal[f]	-2.23	-2.39	-1.04	0.15	-2.82		1.92	2.07
Plastic[g]	-0.81	-0.97	0.37	1.57	-1.40	1.41		0.80
Other[h]	0.96	0.81	2.15	3.34	0.37	3.19	1.78	
	Mean ln(Xj/Xi)						Total variance	

[a] *Avoidable vegetable food waste*

[b] *Unavoidable vegetable food waste*

[c] *Avoidable animal-derived food waste*

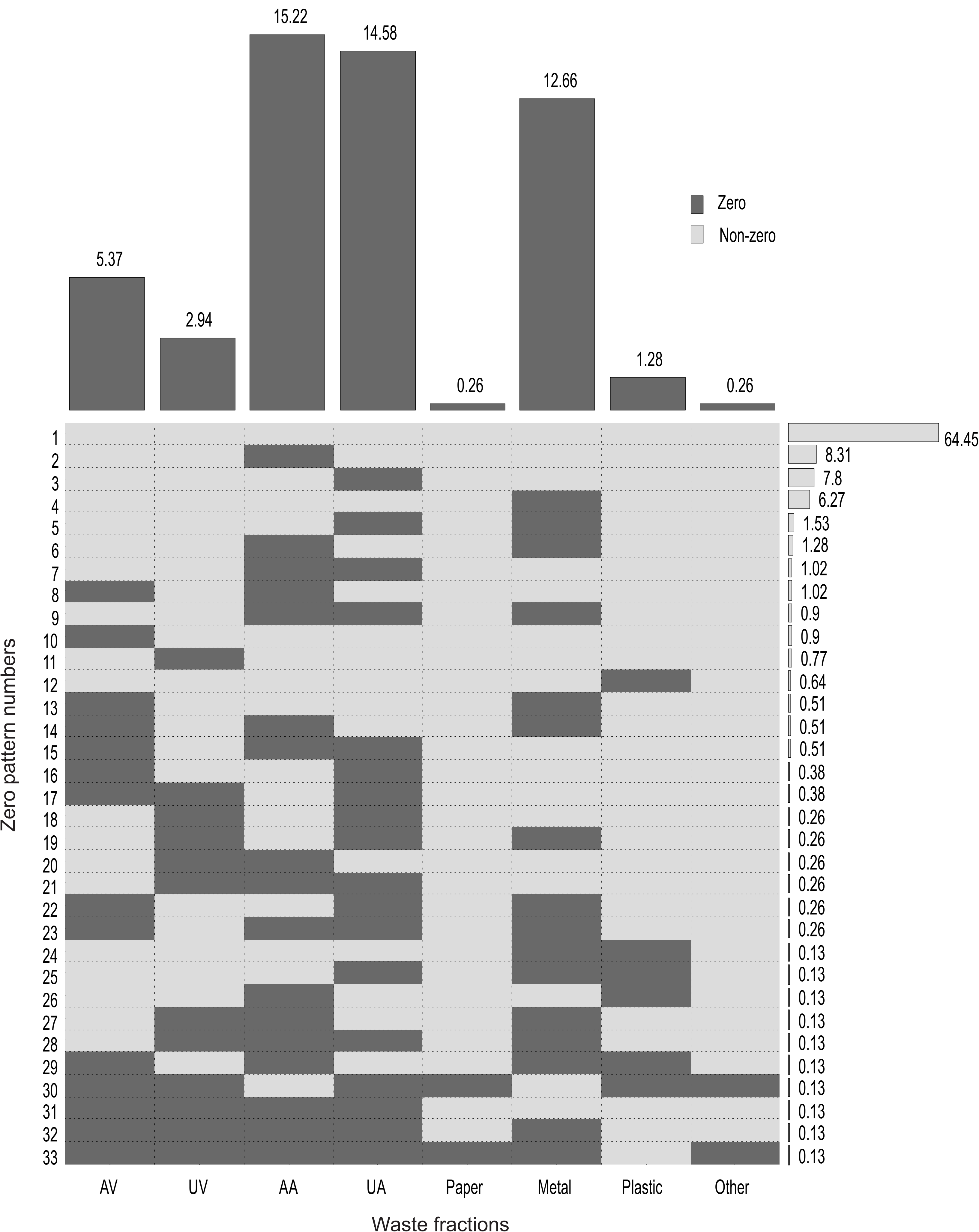
[d] *Unavoidable animal-derived food waste*

[e] *Paper and board;* [f] *Metal packagin.;*

[g] *Plastic packaging;*

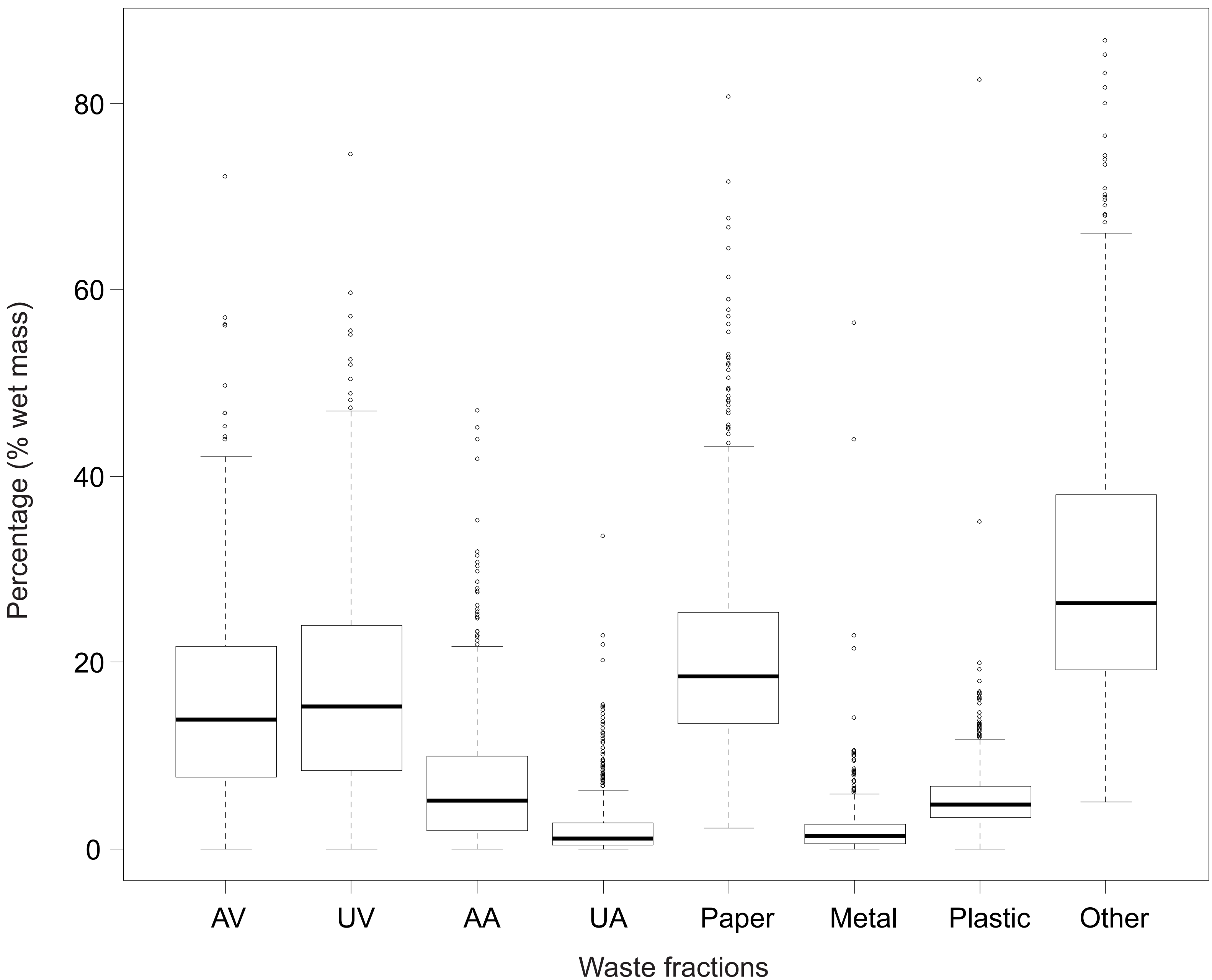
[h] *grouped waste fraction (see Table 1 for waste fractions)*

Figure 1



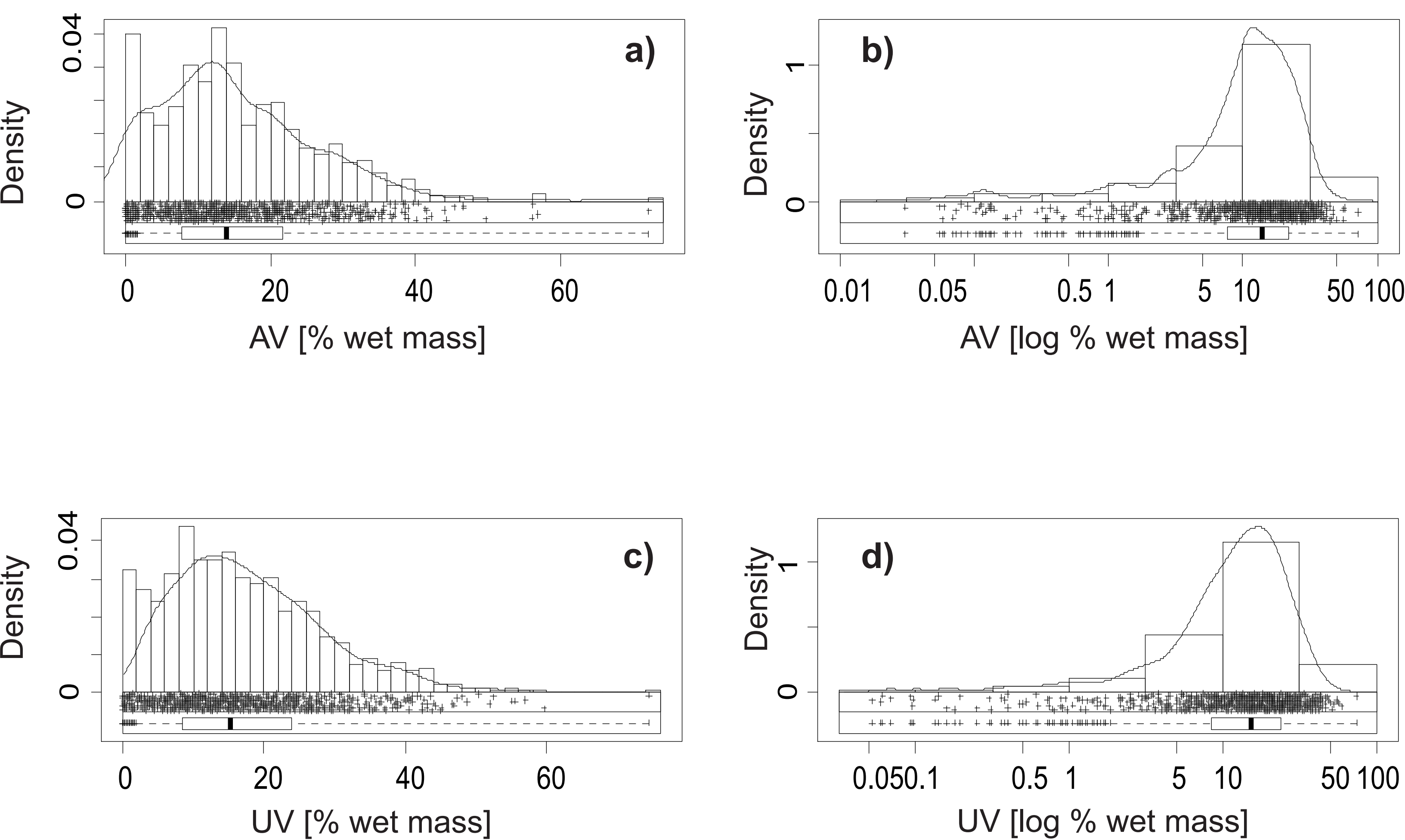
5.37
2.94
15.22
14.58
0.26
12.66
1.28
0.26
Zero
Non-zero
Zero pattern numbers
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
64.45
8.31
7.8
6.27
1.53
1.28
1.02
1.02
0.9
0.9
0.77
0.64
0.51
0.51
0.51
0.38
0.38
0.26
0.26
0.26
0.26
0.26
0.26
0.13
0.13
0.13
0.13
0.13
0.13
0.13
0.13
0.13
0.13
AV
UV
AA
UA
Paper
Metal
Plastic
Other
Waste fractions

Figure 2



Percentage (% wet mass)
0
20
40
60
80
AV
UV
AA
UA
Paper
Metal
Plastic
Other
Waste fractions

Figure 3



a)
Density
0
0.04
0
20
40
60
AV [% wet mass]
b)
Density
0
1
0.01
0.05
0.5
1
5
10
50
100
AV [log % wet mass]
c)
Density
0
0.04
0
20
40
60
UV [% wet mass]
d)
Density
0
1
0.05
0.1
0.5
1
5
10
50
100
UV [log % wet mass]

Figure 4

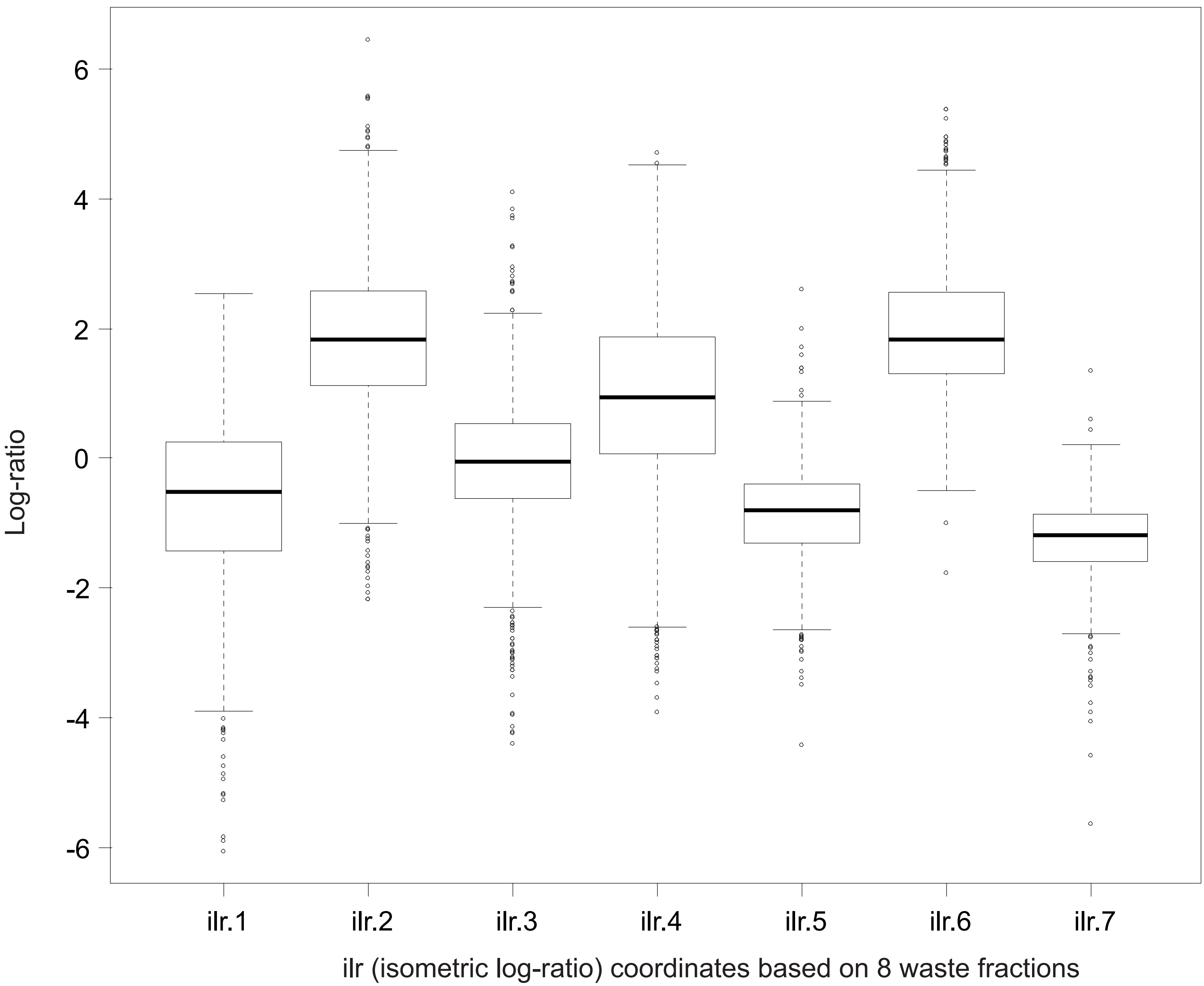

6
4
2
0
-2
-4
-6
Log-ratio
ilr.1
ilr.2
ilr.3
ilr.4
ilr.5
ilr.6
ilr.7
ilr (isometric log-ratio) coordinates based on 8 waste fractions

Figure 5

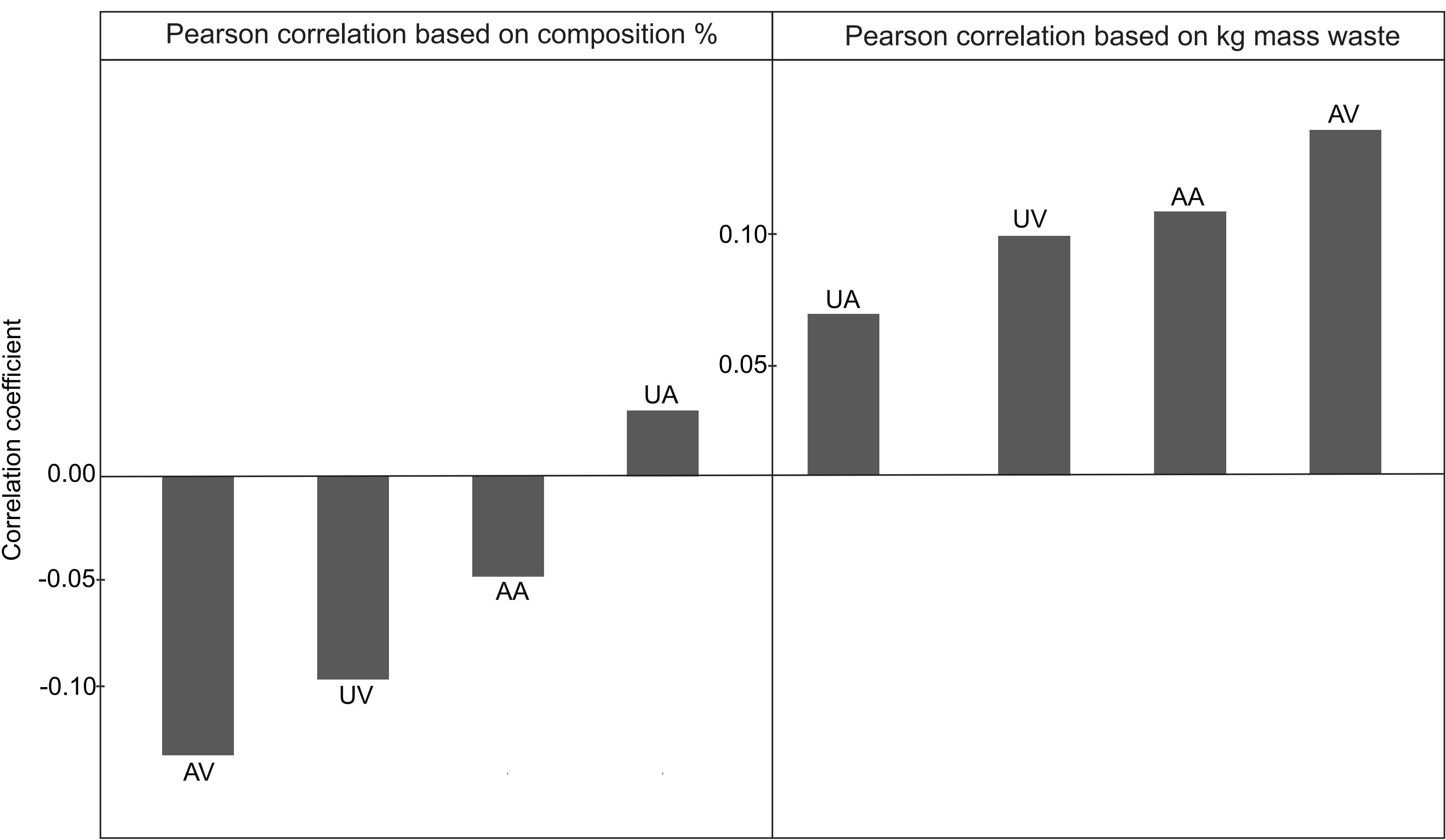

Pearson correlation based on composition %
Pearson correlation based on kg mass waste
Correlation coefficient
0.10
0.05
0.00
-0.05
-0.10
AV
UV
AA
UA
UA
UV
AA
AV

Figure capitations

Figure 1: Identification of zero value patterns in residual household waste dataset subdivided into eight waste fractions (see Table 1) and consisting of 779 observations (households). Vertical bars (in dark grey) represent percentage of count number of zero values for each waste fractions; Horizontal bars (light grey) indicate the percentage of count number of no zero value for each combination of eight waste fractions in the households-33 zero values patterns were observed.

Figure 2: Percentage distribution of the composition of residual household waste fractions on wet mass basis (see Table 1 for abbreviation).

Figure 3: Combined histogram and boxplot of raw (a) and log-transformed (b) avoidable vegetable food waste; and raw (c) and log-transformed (d) unavoidable vegetable food waste.

Figure 4: Boxplot showing the distribution of ilr coordinates (number of coordinates equals to number of waste fractions (D=8) minus 1)

Figure 5: Results of Pearson correlation test between plastic packaging and food waste fractions (AV, UV, AA, and UA), based on (i) percentage (%) and (ii) kg mass of waste fractions.

Statistical analysis of solid waste composition data: arithmetic mean, standard deviation and correlation coefficients

Maklawe Essonanawe Edjabou[1]*, Josep Antoni Martín-Fernández[2], Charlotte Scheutz[1], Thomas Fruergaard Astrup[1]

1) Department of Environmental Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

2) Dept. Computer Science, Applied Mathematics and Statistics, University of Girona, Campus Montilivi (P4), E-17071 Girona, Spain

Supplementary materials (SM)

Supplementary materials contain detailed food waste data used for calculations. SMs are divided into tables (Table SM) and figures (Figure SM).

Supplementary materials (SM) –Tables

SM Table 1 Correlation matrix from Pearson` correlation test and significance levels of **log-transformed** data(r: range:-1.00 to +1.00)

	AV[d]	UV[e]	AA[f]	UA[g]	Paper[h]	Metal[i]	Plastic[j]	Other	Datasets
Avoidable vegetable food waste (AV)	1	*	***	0	***	.	0	***	Percentage %
	1	***	***	***	***	***	***	***	kg/capita/week
Unavoidable vegetable food waste (UV)	0.08	1	0	***	0	0	0	***	Percentage %
	0.41	1	***	***	***	***	***	***	kg/capita/week
Avoidable animal-derived food waste (AA)	0.34	0	1	0	***	.	0	***	Percentage %
	0.53	0.27	1	***	***	***	***	***	kg/capita/week
Unavoidable animal-derived food Waste (UA)	-0.01	0.13	0.02	1	0	*	**	**	Percentage %
	0.23	0.29	0.2	1	***	***	***	***	kg/capita/week
Paper and board	-0.21	-0.05	-0.14	0.01	1	0	0	***	Percentage %
	0.41	0.38	0.31	0.22	1	***	***	***	kg/capita/week
Metal packaging	0.07	0.01	0.06	0.09	-0.05	1	***	.	Percentage %
	0.34	0.24	0.27	0.21	0.28	1	***	***	kg/capita/week
Plastic packaging	-0.04	-0.04	0.04	0.11	0.02	0.18	1	*	Percentage %
	0.4	0.29	0.36	0.25	0.38	0.38	1	***	kg/capita/week
Other	-0.31	-0.37	-0.22	-0.1	-0.27	-0.06	-0.08	1	Percentage %
	0.38	0.23	0.29	0.18	0.43	0.3	0.38	1	kg/capita/week

[***]Very high significance probability higher than 0.001

[**]High significance probability between 0.001 and 0.01

[*]Significance probability between 0.01 and 0.05

() no significance-probability higher than 0.05

[a] amount of waste (wet basis) per household per week

[b] amount of waste (wet basis) per person per week

[c] Composition of residual household waste on wet basis.

[d]Avoidable vegetable food waste

[e]Unavoidable vegetable food waste

[f]Avoidable animal-derived food waste

[g]Unavoidable animal-derived food waste

[h]Paper; [i]Metal packaging.; [j]Plastic packaging; [k]Other (see Table 1).

SM Table 2 Summary of waste fraction generation rates data set **before** zero values replacement

	min	max	mean	Standard deviation
Avoidable vegetable food waste (AV)	0.000	12.435	1.760	1.654
Unavoidable vegetable food waste (UV)	0.000	21.750	1.687	1.457
Avoidable animal-derived food waste (AA)	0.000	9.314	0.755	0.891
Unavoidable animal-derived food Waste (UA)	0.000	5.450	0.210	0.344
Paper and board	0.050	14.519	2.042	1.616
Metal packaging	0.000	13.415	0.213	0.556
Plastic packaging	0.000	19.415	0.524	0.753
Other	0.194	25.747	3.063	2.583

SM Table 3 Summary of waste fraction generation rates data set **after** zero values replacement

	min	max	mean	Standard deviation
Avoidable vegetable food waste (AV)	0.006	12.435	1.760	1.653
Unavoidable vegetable food waste (UV)	0.006	21.750	1.687	1.457
Avoidable animal-derived food waste (AA)	0.003	9.314	0.756	0.891
Unavoidable animal-derived food Waste (UA)	0.002	5.450	0.210	0.344
Paper and board	0.050	14.519	2.042	1.616
Metal packaging	0.002	13.415	0.213	0.556
Plastic packaging	0.007	19.415	0.524	0.753
Other	0.194	25.747	3.063	2.583

SM Figure 1: Comparison of waste data sets before and after zero values replacement