Quantification of Waste Generation in the EU

A PPCA and regression analysis on prediction of recyclable waste

Filip Sandkvist

2016

Filip Sandkvist
Thesis for Master's degree in Applied Climate Change Strategies, 30 ECTS, Lund University
Supervisors: Johanna Alkan Olsson, Nina Reistad

Centre for Environmental and Climate Research
Lund University
Lund 2016

Abstract

In this study, data of the generation of recyclable wastes from the EU member states, and possible explaining factors for describing this generation, are examined through a combination of Probabilistic Principal Component Analysis (PPCA) and multivariate regression analysis. The purpose is to identify some of the biggest contributors to the generation of recyclable wastes, and, based on these contributors, find a linear function that describes the generation of different recyclable wastes, as well as assess the predictive power of this function. Initially, PPCA was used to reduce the number of datasets in order to include only the most important explaining factors. Later, multivariate regression analysis was used to define the coefficients of the waste-generation function. This function describes just above 86% of the total waste generation of recyclables, and an average of nearly 68% of the generation of the individual wastes. The generation of paper and cardboard, glass and plastic are well described by the function. The generation of rubber, textile and wood are less well described. This study points out GDP, primary energy consumption, LMP expenditure and low education level as important predictors of the waste generation of recyclable wastes. These four factors are also important to consider in the future, as they could help define areas of particular interest in the strive towards a sustainable society.

Keywords:
Correlation, multivariate, PCA, recyclable, regression, waste, avfall, korrelation, multivariat, återvinningsbar.

Contents

1. Introduction

1.1 Background

We live in a time of rapidly increasing resource flows. In 2012, the municipal solid waste in the world had doubled over just one decade, and was expected to triple within another ten years (The World Bank, 2012, referenced in Soltani, Sadiq & Hewage 2016, p.388). If this development is to stop, it is of great importance to identify the main causes and indicators of waste generation. And with increasing waste generation follow a demand for better techniques and strategies to discard as little material as possible. In this century, sustainable management will need to be a part of every step of the process in treating municipal solid waste, and a number of systems analysis techniques have been introduced over the course of the last decades to help handle municipal solid waste (Pires, Martinho & Chang, 2011, p.1034).

One example is found in the paper by Wang, Richardson, and Roddick (1996, p.235) that addresses the issues of solid waste management in municipalities, with respect to for example viability in terms of costs and effectivity of recycling strategies for reducing landfilling. Although possibly a useful tool for decision makers on a local scale, it appears as if detailed data from many measurements is required. In short, reaching simplicity (in the sense of building a model with few variables) could be an issue. Also, starting from a local scale, it could be difficult to capture the underlying patterns important for generalising the model to waste management on a larger scale.

There are examples of many similar studies with other approaches over the years. 45 different models on waste generation in European cities were reviewed by Beigl, Lebersorger and Salhofer (2008, p.202). Correlation and regression analyses were proposed as some of the better candidates (op. cit., p.212). Among these options there is among others an analysis of local data of every European city of more than half a million residents, conducted by Beigl et. al. (2004, p.2). The statistical technique used was partially multivariate (op. cit., p.3). Unfortunately, this study lacks a statistical technique to select data to be included in the final model. From the 65 cities fulfilling the criterion of more than half a million residents at the time of the analysis, Beigl et. al. were able to collect data on municipal solid waste from 31 cities, that is, less than half of the cities (op. cit., p.2). This is an indication that it was rather hard to find enough data with the proper quality on a local scale at that time. Lack of accurate data is also another reason to aim at a model that is as simple as possible so that the included variables, and measurements required to make predictions to support waste management, could be kept to a minimum.

Another approach is the time series analysis, such as the work done by Katasamaki, Willems and Diamadopoulos (1998). The focus in this paper is on predictions of municipal waste generation in a shorter time frame, that is, from day to day. The area under analysis covered 13 municipalites and about 1 000 000 citizens. Even though the time series analysis could make accurate predictions from day to day (op. cit., p.182), which could be very useful, the model parameters might not fit well when performing a similar analysis on the same scale somewhere else. And again, a generalisation to a larger scale could be difficult.

With the many analysis techniques already tested, there is still a need for stronger methodologies to help provide information for decision makers on the issue of waste treatment strategies locally throughout the EU (Pires, Martinho & Chang, 2011, p.1044). And, among other things, these new methodologies should consider economic, environmental and social aspects (ibid.). The interpretation of the big picture with regard to the studies this far considered, is that it appears as if there is still a need for a statistical model that

- is simple in terms of included variables, so that a sufficient number of measurements and data does not become an overwhelming task
- captures economical, social and environmental aspects of assessing waste generation
- is applicable on a local, national and an international scale

This study is an attempt to address these issues. The waste generation on national and European level seem to be well documented, and are collected in a database (European Commission, 2016 I). This data can, for example through Principal Component Analysis and regression, be used to make a simple model to find some of the main explaining factors for predicting waste generation. Keeping the model simple and using only data on national level from the EU member states, the model could be applicable in all of EU. With this initially wide scope, it might even be justified to use such a model to include also predictions of the complex local waste generation (although less accurate predictions could be expected). This could provide information about the availability of some important secondary materials, and allows for planning and a better use of such materials.

From another perspective, a model describing waste generation with a high predictive power could, with knowledge about the predicted behavior of the explaining factors, be a tool to make predictions about waste generation in the near future. Also, identifying the most important explaining factors of waste generation could give some guidance about how to decrease the genearion of wastes. Providing support within the topics just mentioned, it is of relevance for the good of the environment and the environmental science to look into the possibilities of making predictions of the waste generation of recyclable wastes.

1.2 Purpose

The purpose of this thesis is to provide a simple way of predicting the generation of the main recyclable wastes in the EU countries through PPCA and multivariate linear regression analysis in the software MATLAB. The predictive relationships will be expressed as a function, linking some of the most important explaining factors to the generation of recyclable wastes. Further, the purpose is to assess the model fit to the measured values as well as the predictive power of the function.

1.3 Problem formulations

- What are some of the most important factors for describing the generation of the main recyclable wastes in the European Union, and how do these factors, expressed as a function, affect the generation of the different wastes?
- What is the model fit, expressed as the correlation coefficient for the generation of each of the considered wastes, between the measured values and the values predicted by the function?
- What is the predictive power of the model, expressed as the coefficient of determination between the measured values and the values predicted by the function?

1.4 Scope

The scope of this thesis is limited to the 28 member states of the EU, and the data provided by the statistical office of the EU (2016a). The data considered in this thesis consists to one part of waste-generation data of recyclable waste from each member state within the categories glass; metal; paper and cardboard; plastic; rubber; textile; wood; and the total generation of recyclable waste. To the other part the data is limited to the potential explaining factors of the generation of these wastes from every member state, related to the categories production; information and technology; environmental focus; and standard of living. The data considered covers the years 2004, 2006, 2008, 2010 and 2012.

1.5 Disposition

First there is a section covering the theory behind the analysis tools used in this thesis, PPCA and multivariate linear regression analysis, as well as information about the data used in this thesis and how the data was pre-treated in order to improve the results. After the theory section, a method section will follow on how the process of selecting relevant data was conducted, and a walk-through of the entire process from raw data to results. The results are presented in the following section. The next section, discussion, include the main topics interpretion of results, verification of results, model utility, and last, a prognosis for the year 2020. Lastly, there are a few concluding remarks.

2. Theory

2.1 Principal Component Analysis

2.1.1 Principal Component Analysis – a brief introduction

Principal Component Analysis (PCA), was during its first appearance used as a method to reduce a set of correlated variables to a smaller set, ordered by decreasing variance (Izenman 2008, p.196). The n correlated data vectors, *y*, are reduced into a set of k orthogonal (uncorrelated) eigenvectors or principal axes (coordinate axes along the directions of the largest variance), *w* (Chen 2003, p.2). Another way to put this, is that the principal axes are the linear combinations of the original variables with the largest variance (Anderson 2003, p.459). The variance is relevant, since it describes the information content that a variable carries (Izenman 2008, p.196). Other benefits of the PCA technique is that it can decorrelate the original variables (op. cit., p.215), by rotating the coordinate system in such a way that the correlation of these variables becomes zero. By a dimensional reduction with PCA, a few variables can possibly be identified which contain the majority of the information from the original, larger, set of variables.

The principal axes w_j are found by solving the eigenvalue problem (Chen 2003, p.2)

$$S w_j = \lambda_j w_j \tag{1}$$

where the k principal axes have the corresponding eigenvalue λ, and

$$j \in 1,2,\ldots k$$

where the sample covariance matrix S is defined as

$$S = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^T}{n} \tag{2}$$

and

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{3}$$

with

$$i \in 1,2,\ldots n$$

A principal component (PC) x_i is defined as

$$x_i = W^T (y_i - \bar{y}) \tag{4}$$

where W is the matrix containing the k principal axes w_j

$$W = [w_1, w_2, \ldots w_k] \tag{5}$$

The intuition behind finding the first PCs could be illustrated by a number of people standing around a teapot, viewing it from different angles – some looking from the side and some from above. One could argue that it is possible to capture much of the visual information about the outside of the teapot, by placing only two of these people in an orthogonal position in relation to the other – one facing the side of the pot, and one looking from above. Finding the positions of these people that allows each person to see as much of the pot as possible from their respective angle in relation to the teapot (analogous to capturing as much of the information, or variance, as possible), would be analogous to finding the first two PCs as illustrated by Emmerich (2013, pp.15-16, referring to the picture in James X. Li, 2009).

2.1.3 Probabilistic Principal Component Analysis

The Probabilistic Principal Component Analysis (PPCA) is a version of the PCA that can handle data with missing values. Simply leaving out the missing values leads to the wrong solution (Ilin & Raiko 2010, p.1966). In the case of missing values, the PPCA algorithm in MATLAB reconstructs a value according to the algorithm (Tipping & Bishop 1999, p.612 refered to by The Mathworks Inc., 2016a)

$$y_i^T = W x_i^T + \mu + \varepsilon_i \tag{6}$$

where y_i are the vectors of observed data y at iteration i, related to the vectors x_i by the matrix W in the new space defined by the PCs w with the indices j, w_j (see eq. 1 and 4, p.6, and eq. 5). μ is a parameter vector that simplifies the solution by allowing a non-zero mean, and ε_i is the isotropic error term (op. cit., p.612). The error ε_i is a measure of the variability in a given data vector y_i (op. cit., p.613). ε is assumed to be a Gaussian distribution, and missing values in the data vectors y are predicted by taking values from a Gaussian probability distribution over an iterative procedure, reducing the reconstruction error (Chen 2003, pp.3-4).

2.1.4 PPCA – the springed rod analogy

Chen (2003, p.5) offers an intuitive analogy to the process of finding a PCs x by the iterative procedure mentioned in the previous subsection:

A rod is pinned in place at its origin, leaving it free to rotate. Finally, the rod will point along the direction of the PC. The rod is initially pointed in a direction (which is defined by the first guess of the PC x). With the rod fixed, a projection is made of every data point onto the rod (which corresponds to the predicted value of the data points in the data vectors y, using the guess of the PC x). The data points are attached to its origin by a spring, and the rod is released (which corresponds to comparing the data points in y with the value predicted by the guess of the PC x). From the new direction of the rod, every data point is projected onto the rod (corresponding to choosing a new guess for the PCs x, which makes the predictions of the data points of y, closer to the actual values of the data points of y). This process is repeated until the error of prediction in y (which can be thought of as the energy stored in the sping before release) is sufficiently small.

2.1.5 Explained variance

As previously mentioned, the principal axes w are the linear combinations of the original variables along the direction of the largest variance (Anderson 2003, p.459). The variance accounted for by a PC stands in direct proportion to the associated eigenvalue (see eq. 1, p.6). The explained variance in percent, v_l, accounted for by the first l of the k PCs is defined as (Jolliffe 2002, p.113)

$$v_l = 100 \frac{\sum_{j=1}^{l} \lambda_j}{\sum_{j=1}^{k} \lambda_j} \tag{7}$$

Put in other words, the explained variance by the first l PCs is the ratio of the associated eigenvalues of the first l components compared to the sum of all k eigenvalues.

2.1.6 Choosing the number of components

As a stopping criterion for PCA, it is of interest to consider only the most important PCs. This is a balance between keeping too many PCs, which could result in overfitting (see Prediction error, p.11), and keeping too few PCs with the result that they account for too little of the total variance (and thus, too little of the total information in the original set of variables). One way to get an idea about this balance for a given case is to make a scree plot, a graph of the PCs against the total explained variance (see eq. 7). A big decrease in explained variance from one PC to the next, along with a sufficiently high total explained variance for the PCs up to this point, is an indication that the number of PCs are sufficient. What could be considered a suffient explained variance depends on the situation, but as a rule of thumb, these figures are often in the region 70-90% (Jolliffe 2002, p.113). For this thesis, 80% was considered sufficient.

Another stopping criterion is Kaiser's rule (Kaiser 1960, p.146) that states that, as a rule of thumb, every PC with an associated eigenvalue λ (see eq. 1, p.6) above 1 should be kept, under the condition that each of the original variables are standardized. In this thesis the variables were standardized to standard scores (see Standard scores, p.12). Kaiser's rule stems from the fact that if an eigenvalue under the given constraint does not exceed 1, it does not contribute to more information than any of the original variables (Jolliffe 2002, p.114). This rule was later updated to the value 0.7 (Jolliffe 1972, p.170), and for this reason all PCs with an eigenvalue not greater than 0.7 were excluded in the analyses in this thesis.

When the number of PCs required to describe a sufficiently large proportion of the variance in a given dataset has been decided, the number of original variables to be kept can also be decided. The idea to include only one variable per PC is based on the claim that the effective dimensionality of some dataset often equals the minimum number of PCs required to describe this dataset (Joliffe 2002, p.137). The original variables are selected from loading with the PCs. The loading of a variable on a PC is defined as the correlation coefficient between the PC and the given variable (op. cit., p.72). Matching every PC with the original variable that has the highest loading on that PC is helpful in choosing original variables that are not correlated to one another (op. cit., p.138). In this thesis, one variable per PC has been chosen in every analysis, according the just mentioned procedure.

2.2 Multivariate linear regression analysis

2.2.1 Extending simple regression to multivariate regression

Multivariate linear regression is an extension of multiple linear regression, which, in its turn, is an extension of simple linear regression. The following is the equation for the simple linear regression model, on the form expressed by Hidalgo and Goodman (2013, p.39) but also including the indices i of each data point

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{8}$$

where x_i, y_i and ε_i is an independent (predictor) variable, a dependent (response) variable and the error term for the data point i, respectively. The term α is the intercept of the regression line with the y-axis, and β is the regression coefficient (the slope of the regression line). The method used in this thesis follows the method of least squares (Livingstone 2009, p.147), minimizing the the sum of the error terms ε_i. It could be illustrated as the fitting of the regression line through a collection of data points in such a way that the total distance between the regression line and every data point is minimized.

A multiple linear regression model is an extension of eq. 8 to include multiple independent variables, as expressed by Hidalgo and Goodman (2013, p.39), once again also including the indices i of each data point

$$y_i = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ik}\beta_k + \varepsilon_i \tag{9}$$

where k is the number of independent variables. Eq. 9 could be illustrated as the fitting of a regression plane with k dimensions, that is, as many dimensons as the number of independent variables, to the data (Livingstone 2009, p.155).

Multivariate linear regression is a generalisation of eq. 9 to not only include multiple independent variables, but also multiple dependent variables. Not only may the k independent variables be correlated to the l dependent variables, but the independent variables may also be correlated to each other, as well as the dependent variables may be correlated to one another (Izenman 2008, p.159).The multivariate linear regression model used in this thesis (more specifically called Multivariate Normal Regression), is on the form (The Mathworks Inc., 2016b)

$$Y_i = X_i \beta + E_i \tag{10}$$

where X_i is a vector containing all the values of data point i of the k independent variables (Jolliffe 2002, p.229)

$$X_i = [x_{i1}, x_{i2}, ..., x_{ik}] \tag{11}$$

and Y_i is a vector containing all the values of data point i of the l dependent variables (ibid.)

$$Y_i = [y_{i1}, y_{i2}, ..., y_{il}] \tag{12}$$

and E_i is a vector of l error terms associated with the values of the dependent variables of data point i (ibid.)

$$E_i = [\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{il}] \tag{13}$$

and β is a matrix with k rows and l columns of regression coefficients (ibid.)

$$\beta = \begin{bmatrix} \beta_{11}, \beta_{12}, \ldots, \beta_{1l} \\ \beta_{21}, \beta_{22}, \ldots, \beta_{2l} \\ \vdots, \vdots, \ldots, \vdots \\ \beta_{k1}, \beta_{k2}, \ldots, \beta_{kl} \end{bmatrix}$$

Multivariate regression has been described by Izenman (2008, p.163) not to be multivariate in the true sense, because the way to find the regression coefficients β is simply a series of multiple regressions, with one dependent variable y (see eq. 12) and all the independent variables x (see eq. 11, p.9) giving one new set of regression coefficients β until every y has been included once.

The proposed function to predict the generation of recyclable wastes in this thesis was determined by performing multivariate regression analysis with the generation of the different wastes as dependent variables (see eq. 12), and the explaining factors as independent variables (see eq. 11, p.9), such that the error (see eq. 13) was minimized. The resulting coefficients β together with the explaining factors of the defined generation of recyclable wastes constitutes the model (see eq. 18, p.29).

2.2.2 The correlation coefficient and the coefficient of determination

The correlation coefficient (r) is a measure of how much and in what direction two variables are related on a scale from -1 to 1 (Livingstone 2009, p.39). If the model fit is perfect such that the correlation between a variable x and a variable y is 1, an increase in x is certain to give an increase in y (ibid.). If the correlation is -1, an increase in x gives a decrease in y. The two preceding cases are perfect linear relationships. If there is no relationship between the changes in y and x, the correlation coefficient is 0. r has been chosen as a measure of the model fit in this thesis.

The coefficient of determination is a measure of how much of the variations in a set of data that could be explained by a given model – in other words how well the regression model fits the dataset (Livingstone 2009, p.158). This coefficient is the variation explained by the model divided by the total variation in the data, and is, for the simple linear regression model (see eq. 8, p.9), defined as (Härdle & Simar 2007, p.75)

$$r^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{14}$$

where r is the correlation coefficient, $\hat{y}$ is the value of the dependent variable predicted by the model with an error ε and $\bar{y}$ is the mean of y (see eq. 3, p.6). An r^2 close to 1 indicates that the model can explain almost all of the variations in the data, and an r^2 close to 0 that the model explains little of the variations in the data. r^2 has been chosen as the measure of the predictive power of the model, the waste-generation function, in this thesis.

2.3 Validation

2.3.1 Cross-validation

Models, based on for example partially PPCA as in this thesis, are built using training data. In order to test predictions from models, validation of the result is a way to test the predictive accuracy of the model when new data is presented. Sometimes data is scarce, and training data and validation data has to be drawn from the same source. Cross-validation can be used in this situation. The technique is based on a split of training data into a training set and a test (or validaton) set. The option k-fold cross-validation was used in this thesis, which means that the training set is split into k parts, where one set is used as a test set and the k-1 other parts are used as training set. The process is repeated k times with a new test set every time, so that every data point is used exactly ones as test data.

Different values of k were considered for this thesis. In the case where k equals the number of data points, so that for each iteration only one data point is used as test data, the technique is called leave-one-out cross-validation. This technique is nearly unbiased, but tend not to give the best solution for linear models (Kohavi 1995, p.1137). Since this thesis is based on the linear techniques PPCA and multivariate linear regression, leave-one-out cross-validation has not been considered an option. A low k, as 2- to 5-fold cross-validation, could also be disadvantageous (op. cit., p.1141). Kohavi (1995, p.1143) suggests 10-fold cross-validation, and Rodríguez, Pérez and Lozano (2010, p.575) suggests either 5- or 10-fold cross-validation. For these reasons, 10-fold cross-validation has been chosen for the analyses in this thesis.

2.3.2 Prediction error

One purpose of performing validation is to check for signs of overfitting, which means that the model has a low prediction error on training data, but when the model is tested on test data, the prediction error blows up. Overfitting can occur when the model is too complicated or has too many parameters (Izenman 2008, p.13). The estimator of the prediction error used for the analyses in this thesis is the mean squared error (*MSE*) of prediction (Camacho & Ferrer 2014, p.41), based on the difference between the model prediction $\hat{y}$ and the measured value y_i of the n data points (but normalized after the number of data points, and not the number of missing values as in Camacho's & Ferrer's equation):

$$MSE = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n} \tag{15}$$

2.3.3 Belsley collinearity diagnostics

The Belsley collinearity diagnostics is a way to identify collinearity (The Mathworks Inc., 2016c). Collinearity is a property of the data (Belsley, Kuh & Welsh 2005, p.86). Perfect collinearity indicates that one vector of data (in this thesis, the data of explaining factors or waste generation) is a linear combination of the other vectors, or for lower collinearity that a linear combination of the other vectors can form a line with a small angle to the first vector (op. cit., pp.85-86). Collinearity between variables can be described by the condition indices (The Mathworks Inc., 2016c). A high value indicates a strong collinearity (ibid.). In this thesis, the collinearity test is based on the condition indices.

2.4 Data pre-treatment

2.4.1 Standard scores

It is important that every variable has the same scale before PCA or PPCA, since the scale of the variables affect the result of the analysis (Härdle & Simar 2007, p.219). For example expressing data from one variable in terms of meters instead of kilometers could greatly affect the result of the analysis. In a similar way, the result is affected when different variables have different units. To avoid this problem, variables should be standardized before analysis. The method used in this thesis is a standardization to standard scores, z, such that the data points x with the index i, x_i, have the mean, $\bar{x}$, equal to 0 and standard deviation, σ, equal to 1 (The Mathworks Inc., 2016d):

$$z_i = \frac{x_i - \bar{x}}{\sigma} \tag{16}$$

where

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \tag{17}$$

Standardizing to unit standard deviation has the benefit that the standardized values are kept unitless, which was the reason this method was chosen in this thesis.

2.4.2 Outliers

Outliers, measurements with a value that is significantly different from other values in a dataset, can warp the results, as expressed by Izenman (2008, p.215) and Chen (2003, p.8). Although few, outliers can have a significant impact, since their values are so extreme. Outliers have been defined by the three sigma rule as all values deviating more that three standard deviations from the mean, accounting for a maximum of 5% of the total number of measurements (Pukelsheim 1994, p.88). Before PPCA in the analyses carried out in this thesis, all values of potential explaining factors with a deviation of more than three standard deviations from the mean have been removed from the datasets to improve PPCA analysis.

2.5 Data quality

This thesis was based on raw data from the time interval 2004-2012 provided by the statistical office of the European Commission, Eurostat (European Commission, 2016 I). The data is reported by the statistical authorities of the member states of the EU (European Commission, 2016 II). It could be argued that this data is of high quality. Eurostat has more than 800 employees, and national experts at hand (European Commission, 2016 III). There are reasons to believe that the information provided to Eurostat by the member states is reported by Eurostat with a high accuracy. Eurostat goes under the European Statistics Code of Practice (European Commission, 2016 IV). The Code of Practice consists of 15 principles aimed at assuring the quality of the provided statistics, and indicators to judge as to how well the principles are followed (ibid.). The database in which the information is reported (European Commission, 2016 I) is accessible to the general public, which gives it a chance to question the information. Although it is not certain if the general public has the knowledge to assess the quality of the data.

The data used in this thesis has to some extent been incomplete, in the sense that measured values have not been available from every EU country from every year and data category. Consider as an example the environmental protection expenditure in industry (see Appendix 1: Nr 2, p.45). This dataset has a number of missing values, some missing because of confidentiality issues. Others values are estimates or forecasts. There are also a few examples of breaks in the time series, which means that data from one year is not completely comparable with both previous and following years (possibly because of a change in statistical systems or definitions). It should be noted that the only distinction that has been considered for the raw data that this study is based on is that between missing and available values. No concern has been taken to the fact that some values are forecasts or for other reasons not actually measured values, which would have been a great undertaking to consider within the limited timeframe of this study.

3. Method

3.1 Data navigation

For this thesis, a selection was made among the more than 4600 data sheets from Eurostat's database (European Commission, 2016 V). The data sheets was initially accessed from the Eurostat Data Navigation Tree (European Commission, 2016 VI). From the sheets, specific raw data was chosen and downloaded (see Fig. 1-3 below) as described in Data selection 1 (see Data selection 1 - Initial data selection, pp.17-18).

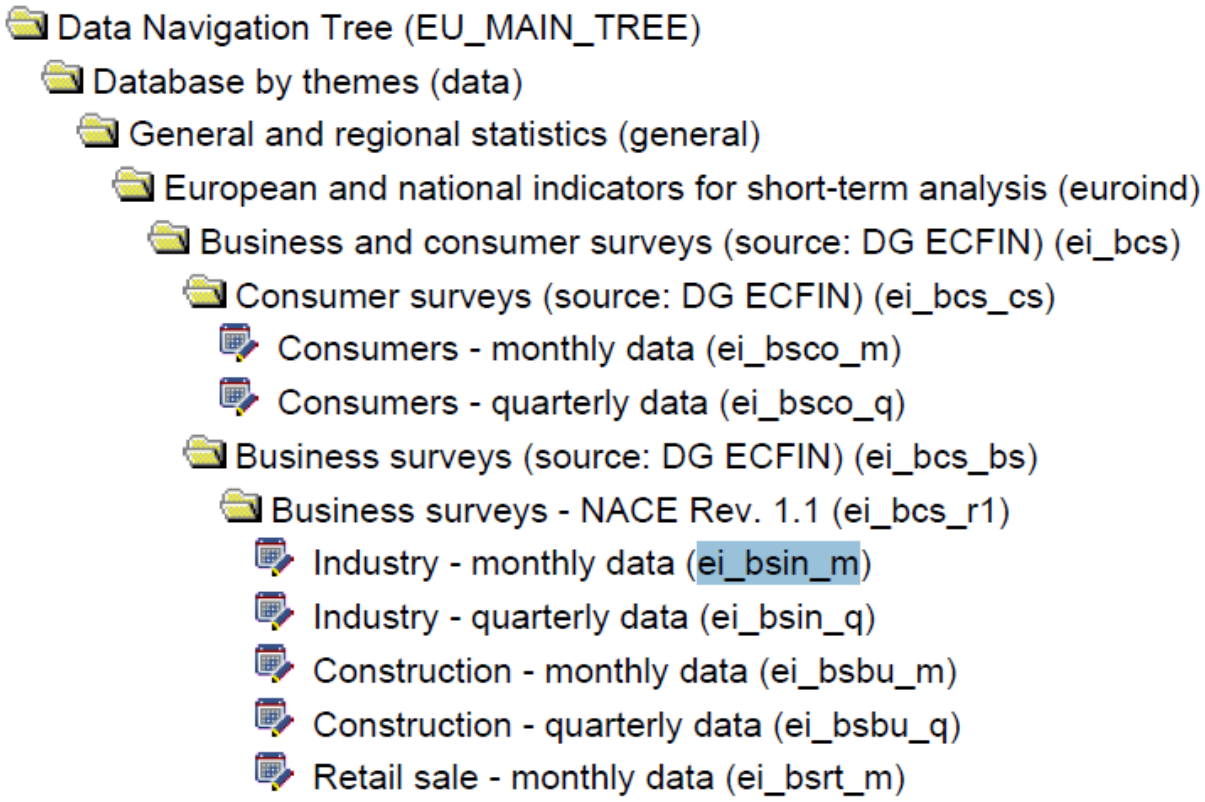

Fig. 1: Eurostat Data Navigation Tree

The Eurostat Data Navigation Tree (European Commission, 2016 VI). The highlighted text is a code specifying a specific data category, in this case monthly industrial data. All data in the tree is linked to Eurostats website, and was reached through the blue icon to the left.

Industry - monthly data

Last update: 12-10-2015

Table Customization show

TIME

Seasonal adjustment

Unadjusted data (i.e. neither seasonally adjusted nor calendar adjusted data)

TIME / GEO	2010M03	2010M04	2010M05
European Union (27 countries)	-1.1	11.6	16.2
Euro area (17 countries)	-2.3	12.1	14.5
Belgium	6.0	20.0	-2.0
Bulgaria	-21.6	-10.9	-6.5
Czech Republic	5.3	26.4	:
Denmark	-9.8	6.4	:
Germany (until 1990 former te	13.0	29.0	18.0
Estonia	-18.4	11.6	:
Ireland	:	:	:
Greece	-28.8	-18.3	-18.0
Spain	-20.0	-7.0	-3.0
France	4.9	25.0	:
Italy	-37.2	-29.2	:
Cyprus	-39.7	-27.5	:
Latvia	-18.1	7.9	:
Lithuania	-21.1	12.8	16.0
Luxembourg	22.0	29.0	:
Hungary	-13.5	-6.2	:
Malta	-7.2	1.9	:
Netherlands	-3.0	3.0	14.0
Austria	-2.2	23.1	27.9
Poland	-8.2	10.3	11.2
Portugal	-4.0	-1.0	:
Romania	-18.3	-12.7	:
Slovenia	13.3	37.3	:
Slovakia	23.0	38.0	51.0
Finland	12.0	21.0	32.0
Sweden	36.2	46.3	:
United Kingdom	0.0	2.7	23.2

Fig. 2: Data sheet from the Eurostat Data Navigation Tree

A data sheet on monthly industrial data. The symbol : corresponds to missing data values.

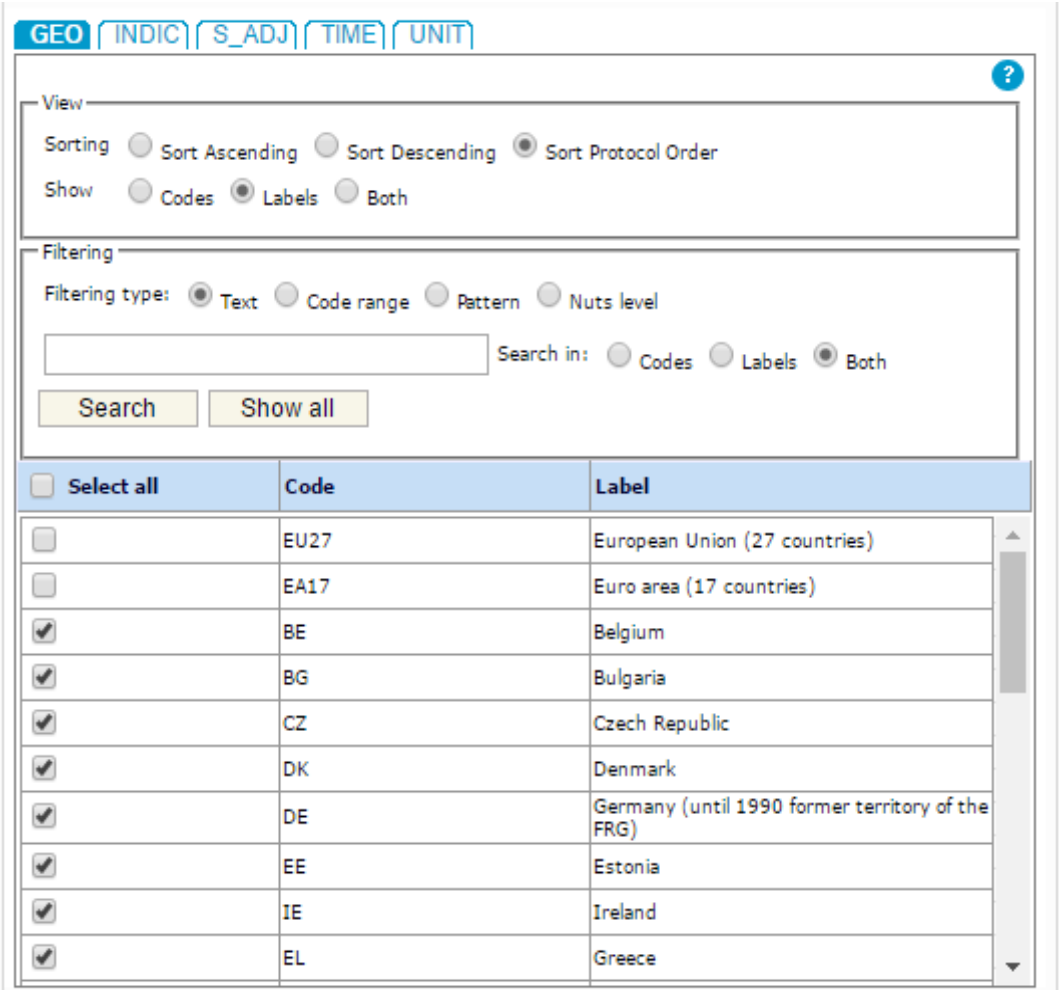

Fig. 3: The data selection menu

A menu allowed the choice to include only certain data in the data sheet before download. The choices included geographical location, time, unit of measure, data subcategories and data adjustments (such as calendar and season adjustments).

3.2 Data selection 1 - Initial data selection

The first data selection, Data selection 1, was conducted as illustrated in Fig. 4 below, initially from Eurostat's database.

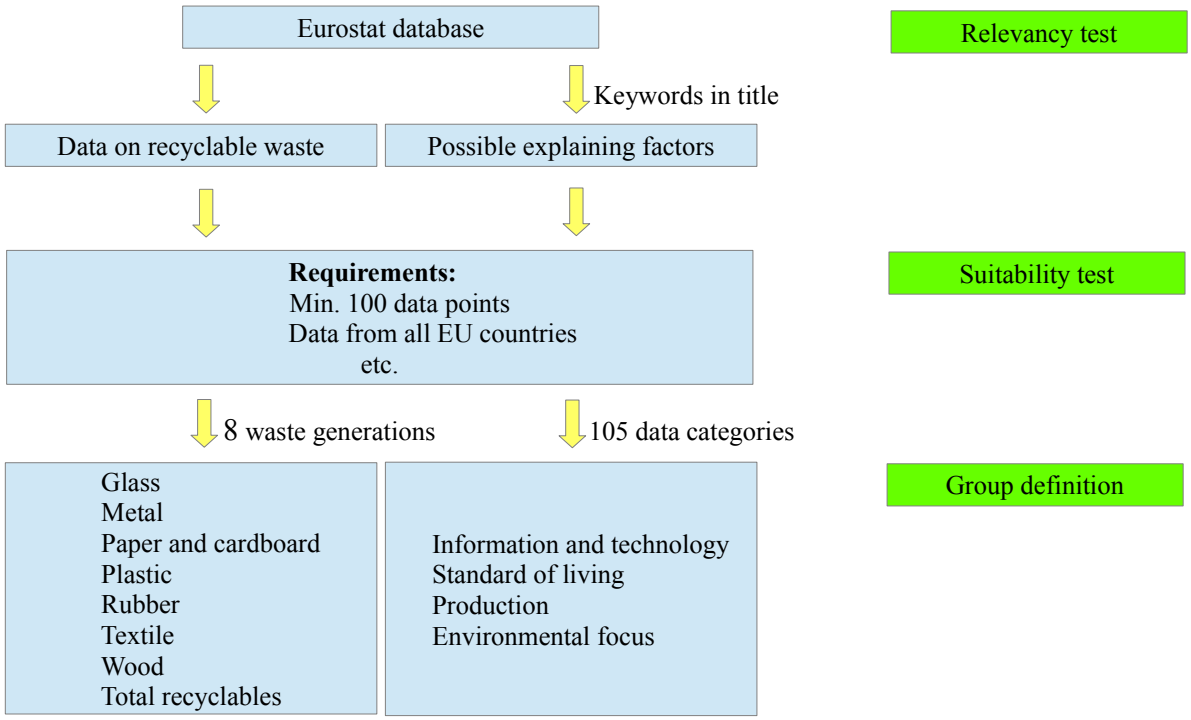

Fig. 4: Overview of Data selection 1

An illustration of Data selection 1.Green represents every phase, and the yellow arrows represent a selection from one step to the next.

First a relevancy test was made, so that only data on the generation of recyclable wastes (including glass; metal; paper and cardboard; plastic; rubber; textile; wood or the total generation of all these wastes) and factors that were expected to be related to the generation of recyclable waste, possible explaining factors, were chosen based on a set of requirements. All of these categories, both the data on recyclable waste and possible explaining factors were collected from Eurostat's database (European Commission, 2016 I). To pass through the relevancy test, the data category had to cover the generation of recyclable wastes, or some property that could be considered likely to affect the generation of these wastes judging from the name of the data category. Data was considered likely to affect the generation of recyclable waste if the title of the data category included words such as:

- Production, turnover, energy (or other words related to resource demanding processes, materially or economically)
- Environmental, sustainable (or words that in other ways relate to environmental or social awareness)
- Connectivity, media, Research and Development, education (or words that in other ways relate to media access or education level)
- Prices, income, employment (or other such words related to standard of living, either low or high)

In the next step a suitability test was made, to include only data that could be considered suitable for data analysis. To pass this step, the data had to:

- Be organised after year (not for example month or quarter)
- Match the years with available data for the generation of recyclable waste - 2004, 2006, 2008, 2010 and 2012 (if the data was not specifically on recyclable waste)
- Have in total, including all years, a minimum of approximately 100 data points (and in a few exceptions down to 60 data points if the data was considered to be of special importance)
- Include data from each of the 28 member states of the EU
- Include data from every member state organised by country (not for example regions or the EU as a whole)
- Not be apparently redundant (in the sense that categories describing similar properties were avoided)

All data categories that passed the suitability test was collected. The data selected from these categories are listed in Appendix 1-5, together with the specific selections that was made for each category to provide the data to be analysed. This data was then split into defined groups. The waste generation data group, to be used as dependent variables in the multivariate regression analyis (see Appendix 5, p.49), covered statistics on waste generation of recyclable wastes from all economic activities, defined by Eurostat's NACE classification system (European Commission, 2016 VII). The data also covered recyclable waste from households within the following categories, as defined in Eurostat's manual for the implementation of regulation (EC) No 2150/2002 on waste statistics (European Union 2013, pp.26-27):

- Glass (from for example production, sorting and recycling)
- Metal (non-hazardous ferrous, non-ferrous and mixed metals and alloys from for example industry, construction and separate collection)
- Paper and cardboard (mainly from paper and cardboard production, separate collection and mechanical treatment of waste and pulp)
- Plastic (from for example production, sorting, preparation processes and separate collections)
- Rubber (from end-of-life tyres)
- Textile (leather and textile from the leather and fur industry, textile industry, separate collection and mechanical treatment of waste)
- Wood (wood and waste bark mainly from wood processing, the demolition of buildings and the pulp and paper industry)
- Total generation of recyclable waste (which is the sum of the generation of the seven wastes above)

Besides from the waste-generation data, 105 data categories where kept for analysis that were considered possible explaining factors to describe the generation of recyclable wastes. These were divided into four classes of possible explaining factors for the waste generation of recyclable wastes. The classes were:

- Environmental focus (see Appendix 1, p.45)
- Information and technology (see Appendix 2, p.46)
- Production (see Appendix 3, pp.46-47)
- Standard of living (see Appendix 4, pp.48-49)

This grouping was done to prepare for PPCA (see Probabilistic Principal Component Analysis, p.7). Similar data was grouped together so that the later analysis would allow to reduce the data categories to a smaller number of categories, containing a majority of the information of the initial full set of categories.

3.3 Procedure – from raw data to results

3.3.1 Procedure overview

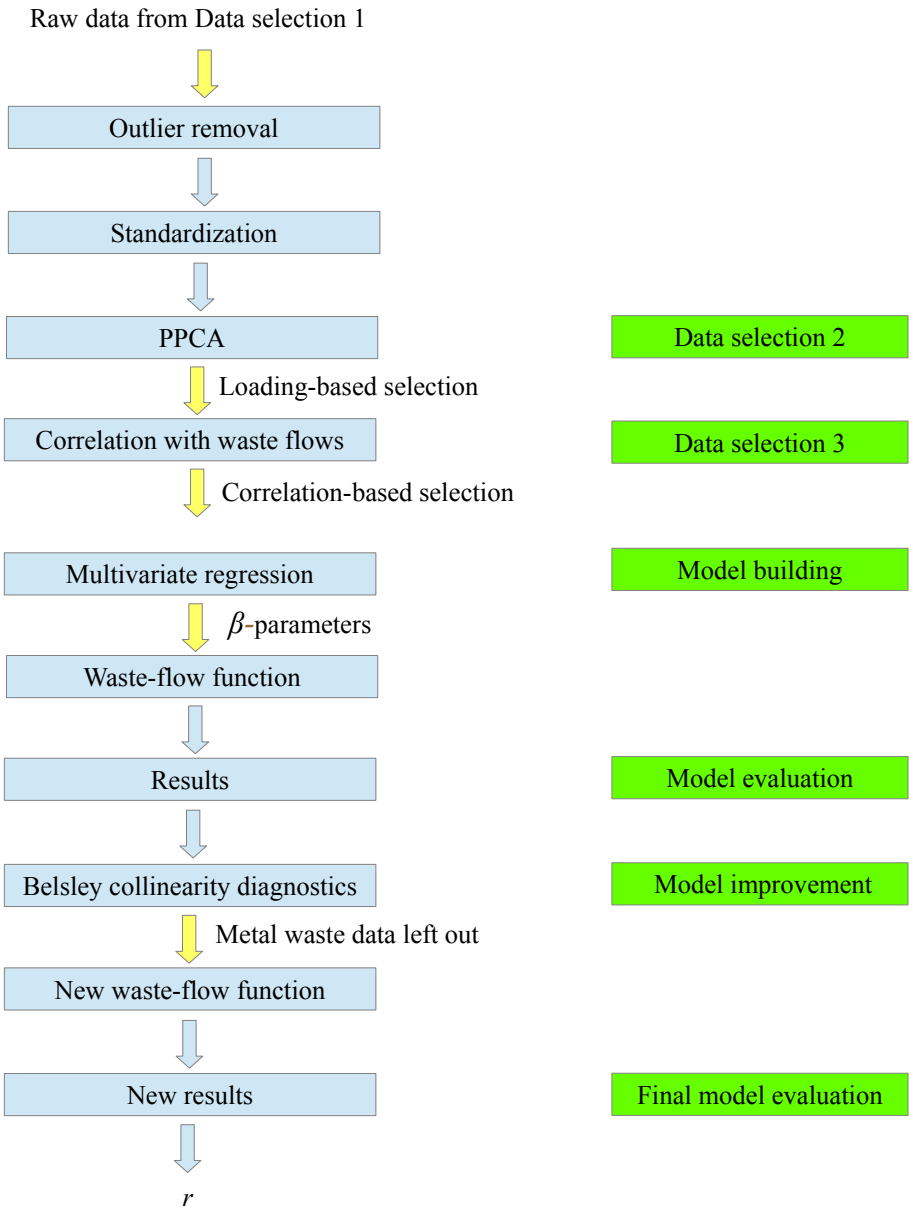

Fig. 5: Procedure overview

The picture depicts the entire process from raw data to results. Blue represents the technique used in every step of the different phases, represented by green. The yellow arrows represent a selection from one phase to be used in the next phase.

3.3.2 Data selection 2 – reduction through PPCA

The data selection through PPCA, Data selection 2, was conducted in three steps including PPCA and preparations for PPCA as as illustrated in Fig. 6 below.

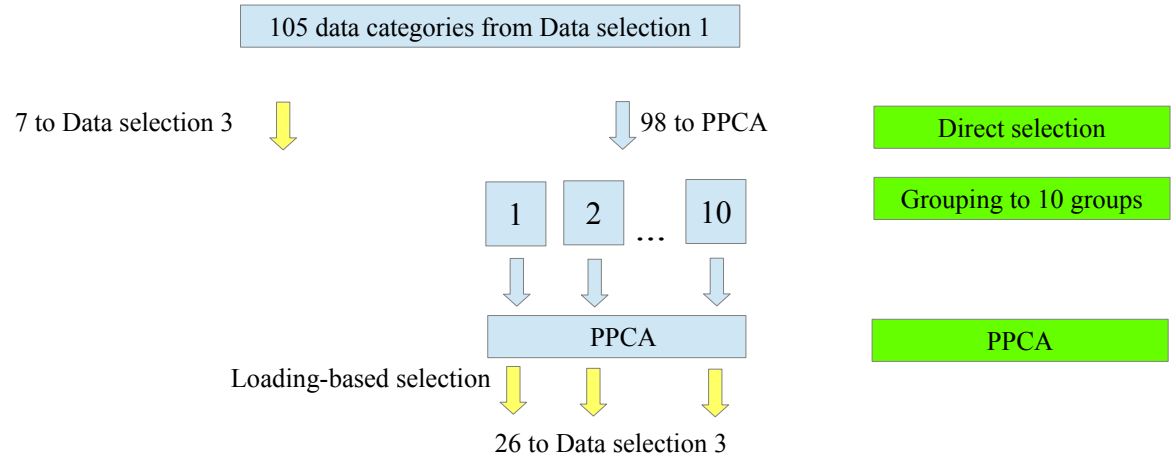

Fig. 6: PPCA

An illustration of the PPCA procedure. Green represents every phase of the procedure. Yellow arrows represent selections, and blue arrows represent steps between phases.

The purpose of Data selection 2 was to make a first reduction in the number of data categories to avoid overfitting (see Prediction error, p.11) in the multivariate regression analysis, while keeping as much of the information in the data as possible. First the data of every category was matched after country and year, so that every data point in one sheet corresponded to the same country and year in another data sheet. For example, when analysing household income and domestic material consumption at the same time, data was matched so that the value for the household income of Belgium in 2004, would correspond to the value of domestic material consumption in Belgium in 2004, and so forth, throughout the entire data sheets of the 105 data categories.

The 105 categories were sorted into groups of similar kind before PPCA (see Table 1, p.25), based on the four classes (such as productional factors, etc.). Seven data categories from the class productional factors were left out from the PPCA and sent directly to the next step of the selection process, because of a seemingly apparent importance of these variables for the generation of recyclable wastes, because they directly corresponded to the generation of materials within the studied waste categories. Three of these data categories were related to the generation of wood waste (see Appendix 3: Nr 38-40, p.47):

- Roundwood production
- Industrial roundwood import
- Total sawnwood import

and four were related to the industrial turnover of (see Appendix 3: Nr 58-61, p.47):

- Metal
- Paper
- Textile
- Wood

Before PPCA, outliers were first removed by sorting out values deviating more than three standard deviations (see Outliers, p.12). The remaining data was standardized to standard scores (see Standard scores, p.12) to avoid skewness when considering variables of different units. PPCA was performed on the 98 data categories one group of similar data at a time, keeping the variables with the highest loading (strongest relationship to the PCs) with a minimum demand of 80% explained variance for the reduced set of variables, neglecting eigenvalues not greater that 0.7 (see Choosing the number of components, p.8). 10-fold cross-validation (see Cross-validation, p.11) was performed to assure that predictions did not show considerably larger MSEs compared to the training data, which would have been a sign of overfitting.

Following the preceding procedure, many data categories could be removed with a small loss in information (see Explained variance, p.8), because a smaller number of data categories within each group seemed to well represent the groups as a whole. From the initial 105 categories, 33 remained and was kept for Data selection 3.

3.3.3 Data selection 3 – choosing explaining factors

The third data selection, Data selection 3, was conducted as illustrated in Fig. 7 below.

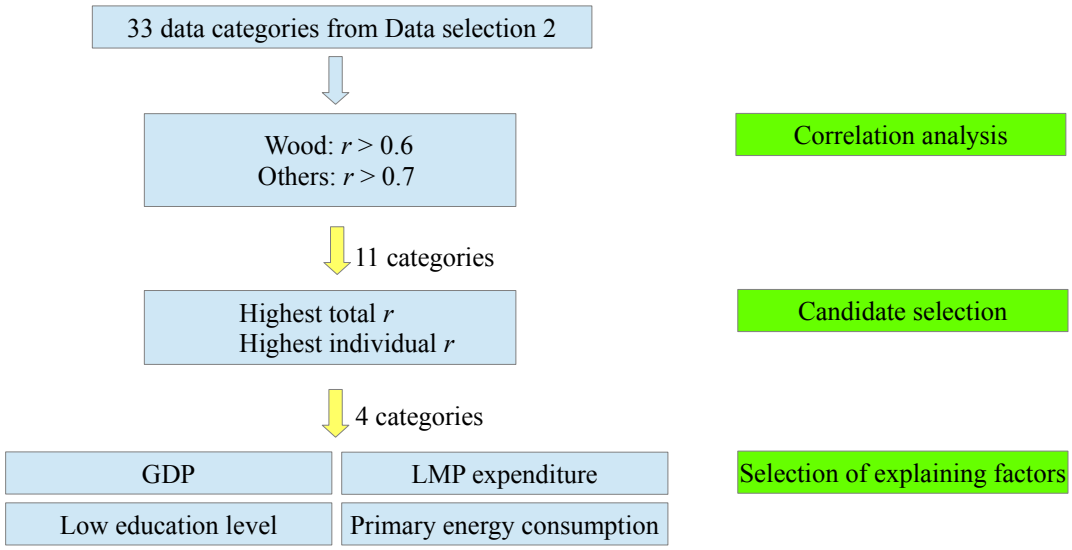

Fig. 7: Correlation analysis

The picture illustrates the process of selecting the four explaining factors through correlation analysis. Green represents phases, yellow arrows represent selections, and blue arrows represent steps between phases.

The purpose of Data selection 3 was to further reduce the 33 data categories to include only the strongest candidates. More precisely, the data categories with a correlation (see The correlation coefficient and the coefficient of determination, p.10) of more than 0.7 with any of the waste generations (except for wood for which the demand had to be lowered to 0.6) were initially kept . In total, eleven categories remained after this step (see Table 3, p.28). Only the categories with the strongest correlation to each waste generation, and the highest total correlation with every waste generation, were then considered as candidates for explaining factors. Four of the eleven categories were enough to include the categories with the highest correlation to every waste generation. These turned out to also have the highest total correlation, and were finally selected as the four explaining factors to build a model on. Two of these were from the class standard of living:

- Labour Market Policy expenditure, million Euro
- Gross domestic product at current market prices, Purchasing Power Standard

One belonged to the class informational and technological factors:

- Population, aged 15 to 74 years, with less than primary, primary and lower secondary education

One was from the class productional factors:

- Primary energy consumption, million tonnes of oil equivalent

3.3.4 Model building

From the four data categories from Data selection 3, the chosen explaining factors, with the strongest correlation to every waste generation, a multivariate regression analysis was made to calculate the regression coefficients, β (see Extending simple regression to multivariate regression, p.10). Every row of the β-matrix constituted the coefficients of each of the four explaining factors, with the generation of wastes as columns in the matrix. Every waste generation was associated with as many β-coefficients as the number of explaining factors, in this case four. For example, the predicted value of wood waste was calculated by multiplying every explaining factor by the respective β for wood waste. In this case the resulting value described the predicted amount of wood waste for given values of the explaining factors. In this way, the predicted values of the generation of wastes were calculated.

3.3.5 Model evaluation and model improvement

Judging from the high correlation between the four explaining factors and the generation of wastes (see Table 3, p.28), from which a high predicitive power could be expected, there was an apparent issue with the model. A collinearity test (See Belsley collinearity diagnostics, p.11) showed a very high collinearity in the data of the generation of wastes, from which a major contribution could be traced to the data of metal waste and to the total generation of recyclable waste. Since the total generation of recyclable waste was considered of bigger interest than the generation of metal waste, the metal waste data was removed from the model, and new β-coefficients were calculated. This resulted in a considerable improvement.

3.3.6 Final model evaluation

The final model, after excluding metal waste, included the generation of the following wastes in tonnes (see Appendix 5, p.49):

- Glass
- Paper and cardboard
- Plastic
- Rubber
- Textile
- Wood
- Total generation of recyclable waste

and the following explaining factors of the generation of these wastes:

- Population, aged 15 to 74 years, with less than primary, primary and lower secondary education (see Appendix 2: Nr 28, p.46)
- Primary energy consumption, million tonnes of oil equivalent (see Appendix 3: Nr 52, p.47)
- Labour Market Policy expenditure, million Euro (see Appendix 4: Nr 87, p.48)
- Gross domestic product at current market prices, Purchasing Power Standard (see Appendix 4: Nr 89, p.48)

The correlation coefficient and the coefficient of determination was calculated, and the measured values from the raw data were compared to the predicted values (see Fig. 19-25, pp.30-33).As a test of the validity of the calculated waste-generation function, the results from the multivariate regression analysis were tested with 10-fold cross-validation (see Fig. 26, p.34). To clarify, this cross-validation was not a test of the calculated β-coefficients (see Table 4, p.29), it was a test of the technique to use multivariate regression to calculate β-coefficients that are then used to predict the generation of recyclable wastes. In this validation, new β:s were calculated in each cycle (see Fig. 26, p.34). The results were considered satisfying, and the model was approved in the current state.

4. Results

4.1 Results from Data selection 2

In this subsection, the results from Data selection 2 are presented. First, the 98 data categories to be analysed in this step, were divided into ten groups of similar data as listed in Table 1 (see p.25). Ten different PPCAs were carried out, and the best candidates were chosen based on the loading with the individual PCs as described in the theory section (see Choosing the number of components, p.8). All PCs with an eigenvalue not exceeding 0.7 were excluded, based on the updated version of Kaiser's rule (see Choosing the number of components, p.8). The number of PCs was decided by the sum of the explained variance (see Explained variance, p.8) of the PCs. A large gap between the second last and the last PC, together with a sufficiently high total explained variance indicated that enough PCs were included. This limit was chosen to be 80%. As help to decide the number of PC was also scree plots, in order to assess the difference in explained variance of the PCs. An example of one of the scree plots that helped in choosing the number of PCs is presented in Fig. 8 below. A 10-fold cross-validation was made together with every PPCA, to test how accurately a PPCA with the chosen PCs from every PPCA in Data selection 2 could predict data (see Fig. 9-18, pp.26-27). The selection of values was performed randomly. The accuracy was judged by the difference in MSE of the training sets compared to the the MSE of the validation sets.

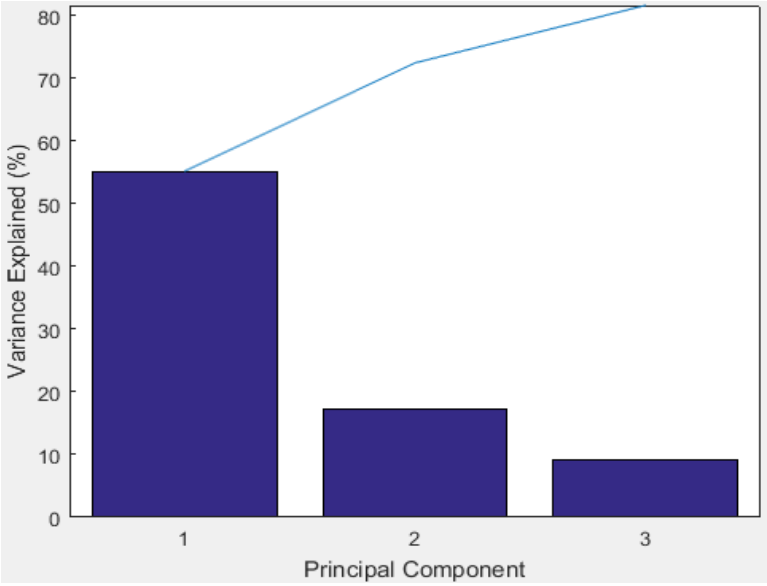

Fig. 8: Scree plot

A scree plot in MATLAB over the PCs and the variance they explain (dark-blue bars). The light-blue line represents the cumulative sum of the explained variance. The three PCs account for about 80% of the total variance. The third PC accounts for less than 10% of the total variance. Since succeeding PCs will contribute with less explained variance than the third, it could be considered justified to keep only three PCs.

Table 1: Loading-based data-category selection

Ten PPCAs were conducted on groups of data categories of similar kind. The group selections were based on the four different classes. The numbers 1-105 in the table identify the data categories, listed in Appendix 1-4. The data categories with the highest loading in every analysis were selected, with as many PCs as necessary to explain the majority of the data. One category were chosen per PC. The minimum requirement for the total explained variance, v_l, in every PPCA was 80%.

PPCA	Selection	Highest loading with PC1	Highest loading with PC2	Highest loading with PC3	Loading (PC1)	Loading (PC2)	Loading (PC3)	v_l (%)
PPCA 1: Environmental focus	1-5, 17-20	5) Organic crop area	1) Environmental expenditure	19) EMAS organisations	0.462	0.631	0.546	83.36
PPCA 2: Environmental focus	6-16	6) Inland renewable energy consumption	16) Final renewable energy consumption	13) Solar thermal energy production	0.402	0.523	0.567	81.56
PPCA3: Information and technology	21-34	30) BERD	27) Percentage of ICT on GDP	28) Low education level	0.406	0.496	0.542	85.17
PPCA 4: Production	35, 41, 42, 49, 51, 52, 55, 64	52) Primary energy consumption	-	-	0.363	-	-	94.80
PPCA 5: Production	36, 37, 48, 50, 53, 54, 56, 57, 62, 63, 65, 66	50) Primary energy production	-	-	0.423	-	-	80.55
PPCA 6: Production	43-47	47) Energy from non-renewable waste	45) Energy from solid fuels	45) Energy from solid fuels	0.572	0.729	0.599	91.48
PPCA 7: Standard of life	67, 68, 89-92, 97, 99, 104	68) Net annual earnings	89) GDP, million PPS	99) Purchasing power parities	0.415	0.453	0.947	86.05
PPCA 8: Standard of life	74-81, 105	75) Severe material deprivation	74) Persistant risk of poverty	76) Housing costs	0.384	0.788	0.940	92.20
PPCA 9: Standard of life	82-88	82) E-banking and e-commerce 85) Digital single market purchases	87) LMP expenditure, million Euro	86) Employment rate	0.450	0.705	0.906	90.29
PPCA 10: Standard of life	69-73, 93-96, 98, 100-103	102) Household income	70) Domestic material consumption	100) Corporation debt-to-income ratio	0.384	0.409	0.610	93.44

Table 2: Total explained variance and ratio of kept data categories

The total explained variance is a measure of how much of the information that was conserved from the 105 data categories from Data selection 1, by keeping only the data categories with the highest loadings to every PC in every analysis presented in Table 1. Selected data categories' are the ratios of kept categories from Data selection 2, compared to the initial 105 categories. The explained variance and the ratio of selected categories were compared both excluding and including the seven categories that were not analysed, but selected directly for Data selection 3.

Number of data initial categories	Total explained variance, v_l (%)	Selected data categories (%)
98 (excl. 38-40 and 58-61)	87.41	27.55
105 (incl. 38-40 and 58-61)	88.25	32.38

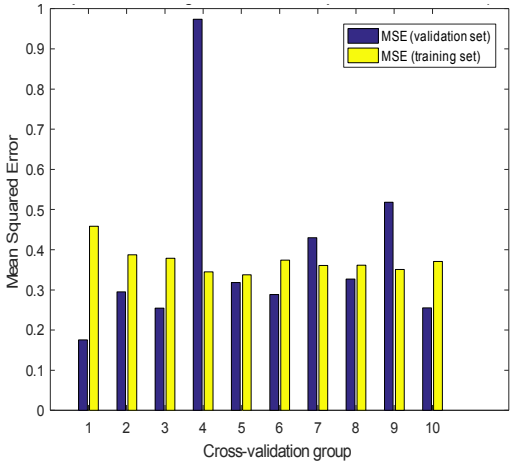

Fig. 9: PPCA 1 *Environmental focus.*

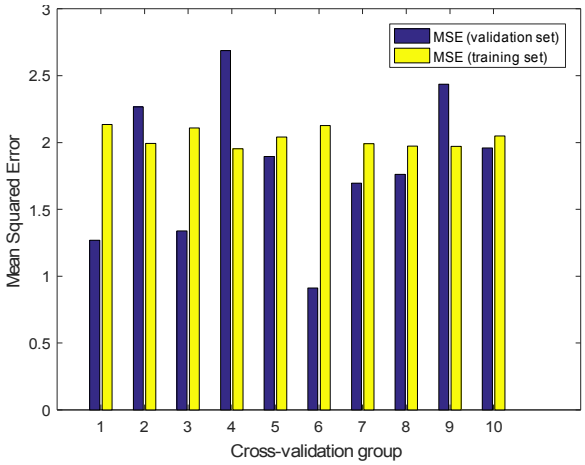

Fig. 10: PPCA 2 *Environmental focus.*

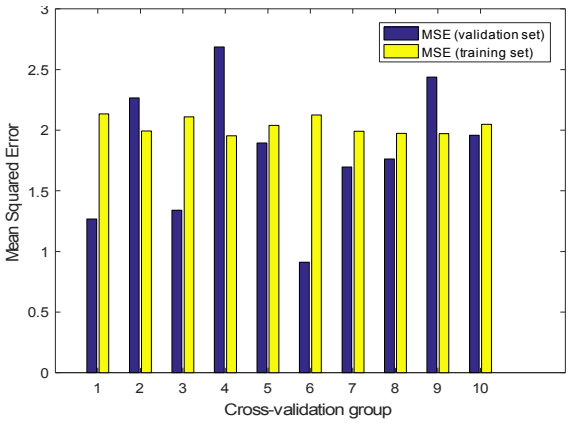

Fig. 11: PPCA3 *Information and technology.*

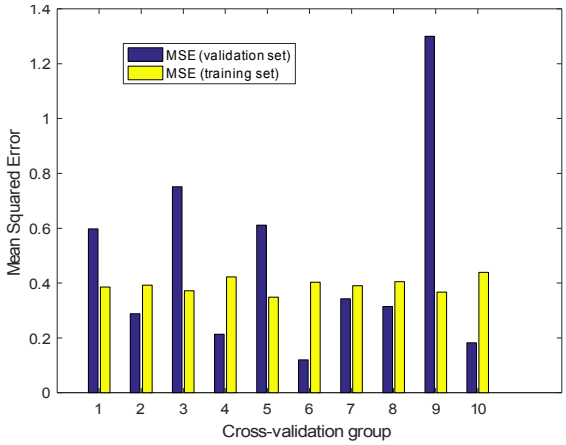

Fig. 12: PPCA 4 *Production.*

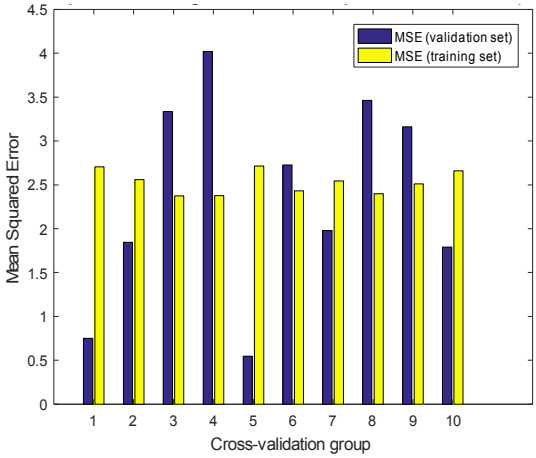

Fig. 13: PPCA 5 *Production.*

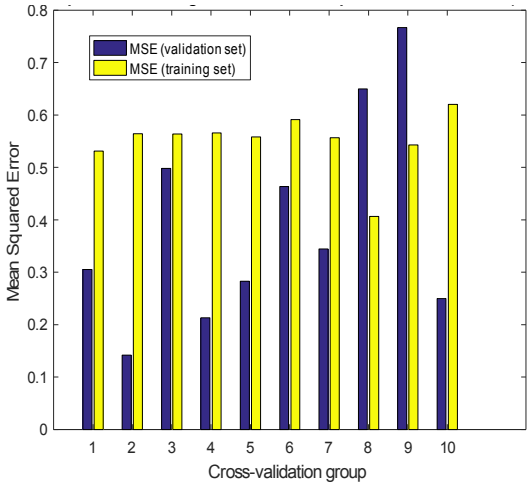

Fig. 14: PPCA 6 *Production.*

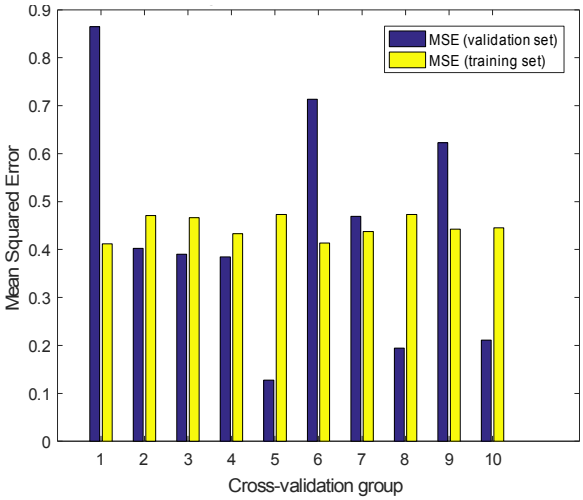

Fig. 15: PPCA 7 *Standard of life.*

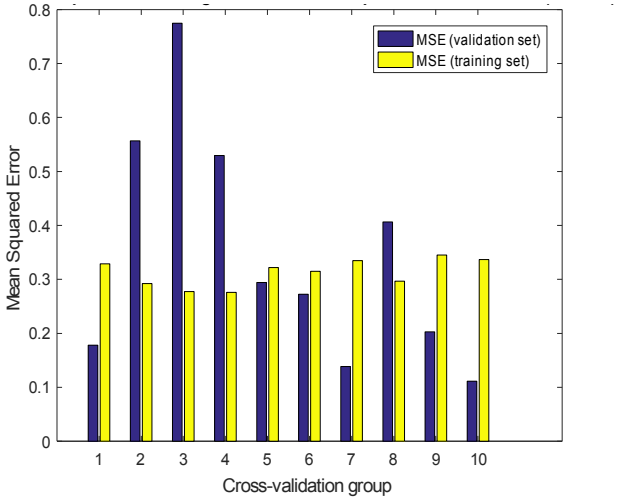

Fig. 16: PPCA 8 *Standard of life.*

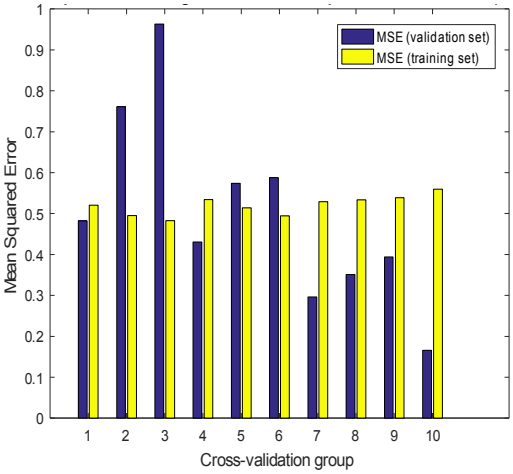

Fig. 17: PPCA 9 *Standard of life.*

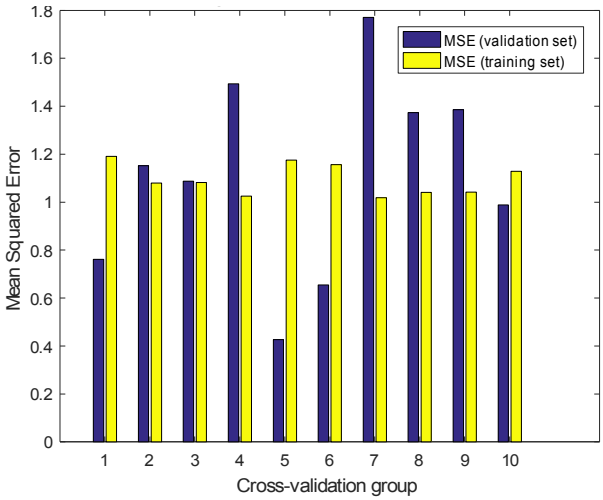

Fig. 18: PPCA 10 *Standard of life.*

4.2 Results from Data selection 3

Table 3: Correlation-based data-category selection

The correlation, r, between the generation of wastes and the best eleven candidates for explaining factors, was calculated. The numbers on the left-hand side in the table identify the data categories, listed in Appendix 1-4. The correlation was measured as correlation between an explaining factor and an individual waste generation, the sum of all correlations between an explaining factor and the generation of all wastes and the ratio of this sum compared to a perfect correlation with every generation of waste. The bolded data categories had the highest correlation with each generation of recyclable wastes. These four also happened to have the highest total correlation to the individual generation of all wastes together. For these two reasons, these four were chosen as explaining factors to include in the waste-generation function.

Nr	Data category	Recyclables, total	Glass	Metal	Paper and cardboard	Plastic	Rubber	Textile	Wood	Total *r*	Total *r* (%)
5	Organic crop area	0.691	0.753	0.695	0.668	0.807	0.753	0.496	0.581	5.44	68.1
6	Inland renewable energy consumption	0.782	0.761	0.762	0.703	0.676	0.620	0.512	0.761	5.58	69.7
28	**Low education level**	0.785	0.924	0.818	0.830	**0.898**	**0.895**	0.692	0.638	6.48	**81.0**
38	Roundwood production	0.547	0.361	0.467	0.378	0.221	0.301	0.208	0.663	3.15	39.3
40	Sawnwood import	0.773	0.805	0.735	0.847	0.775	0.743	0.656	0.604	5.94	74.2
47	Energy from non-renewable waste	0.782	0.802	0.825	0.857	0.709	0.639	0.692	0.791	6.10	76.2
50	Primary energy production	0.826	0.748	0.838	0.769	0.562	0.661	0.587	0.742	5.73	71.7
52	**Primary energy consumption**	**0.930**	0.932	**0.952**	0.918	0.805	0.856	**0.733**	**0.801**	6.93	**86.6**
70	Domestic material consumption	0.800	0.843	0.808	0.778	0.801	0.832	0.610	0.720	6.19	77.4
87	**LMP expenditure, million Euro**	0.855	0.908	0.901	**0.927**	0.782	0.845	0.720	0.770	6.71	**83.9**
89	**GDP, million PPS**	0.915	**0.961**	0.938	0.917	0.874	0.875	0.731	0.756	6.97	**87.1**

4.3 The waste-generation function

Table 4: The *β*-coefficients

This table contains the calculated β-coefficients (without removing outliers before analysis) that relates the explaining factors to the generation of recyclable wastes, rounded to three significant figures.The coefficients were calculated by performing a multivariate regression analysis with the waste generations as the dependent variables (see eq. 12, p.9), and the explaining factors as independent variables (see eq. 11, p.9).

	Recyclables, total β_{cj}	**Glass** β_{gj}	**Paper and cardboard** β_{pj}	**Plastic** β_{lj}	**Rubber** β_{uj}	**Textile** β_{tj}	**Wood** β_{wj}
Low education level j=1	-634	7.95	-144	18.5	12.7	7.97	-77.6
Primary energy consumption j=2	107000	-1360	5590	-8151	-258	1460	93700
LMP expenditure, million Euro j=3	-337	-13.5	-49.0	-26.6	-2.68	-4.02	31.4
GDP, million PPS j=4	19.8	1.77	6.37	2.64	0.198	0.0511	-8.15

w_c = total recyclable waste, tonnes
w_g = glass waste, tonnes
w_p = paper and cardboard waste, tonnes
w_l = plastic waste, tonnes
w_u = rubber waste, tonnes
w_t = textile waste, tonnes
w_w = wood waste, tonnes

f_1 = low education level, thousands
f_2 = primary energy consumption, million TOE
f_3 = LMP expenditure, million Euro
f_4 = GDP, million PPS

$$w_i = \sum_{j=1}^{n} f_j \beta_{ij} = f_1 \beta_{i1} + f_2 \beta_{i2} + f_3 \beta_{i3} + f_4 \beta_{i4} \quad (18)$$

$$i \in c, g, p, l, u, t, w$$

$$j \in 1, 2, 3, 4$$

4.4 Final model evaluation

4.4.1 The correlation coefficients and the coefficients of determination

Table 5: Final model evaluation

The presented waste-generation function was tested against measured data. The correlation and the coefficient of determination between predicted values and measured values were calculated for every waste generation both with and without outliers. The predicted values were calculated using the estimated β-coefficients. The total amount of recyclable waste, 'Recyclables, total', is not entirely dependent on just the generation of the six wastes listed in the table, since this value also includes metal waste. The mean values of the generation of these wastes, 'Mean, all wastes', is the mean of the values for the generation of the six wastes listed below.

Outliers	Coefficients	Recyclables, total	Mean, all wastes	Glass	Paper and cardboard	Plastic	Rubber	Textile	Wood
With	r	0.928	0.824	0.956	0.918	0.931	0.754	0.669	0.659
	r^2	0.862	0.679	0.914	0.843	0.866	0.568	0.448	0.434

4.4.2 Graphical illustrations of model fit

The values predicted by the model using the estimated β-coefficients (see Table 4, p.29) were plotted against the measured values. Since the span between the biggest and the smallest values was great, the values were plotted on a logarithmic scale. Note that the logarithmic scale makes the error, the deviations from the regression line (in red), in the plots below appear smaller than they are.

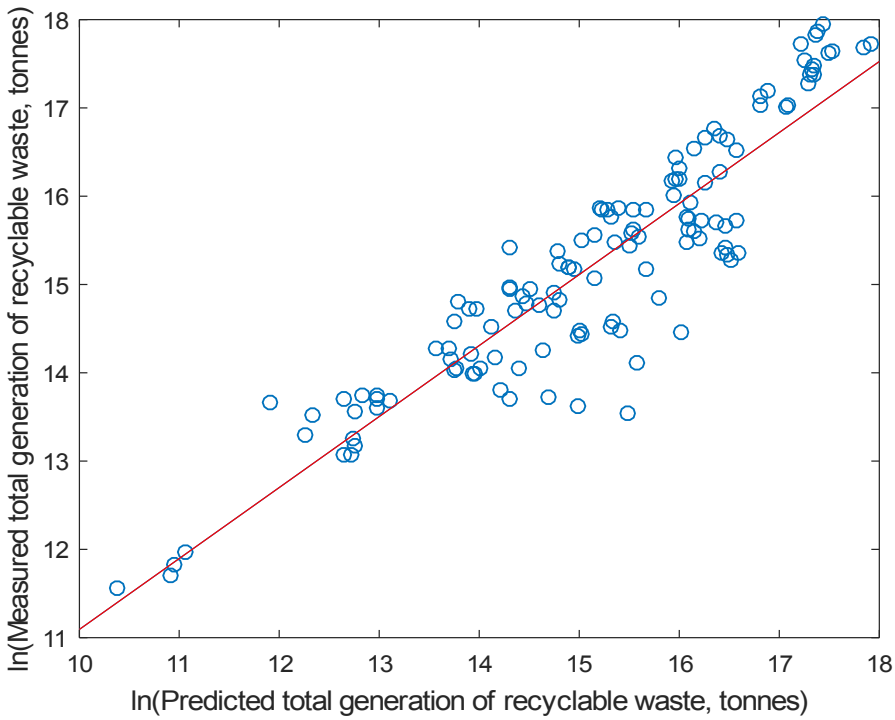

Fig. 19: Total generation of recyclable waste

Logarithmic plot of the total recyclable waste, with predicted values against measured values. No potential outliers have been removed from the data. r = 0.928, r^2 *= 0.862.*

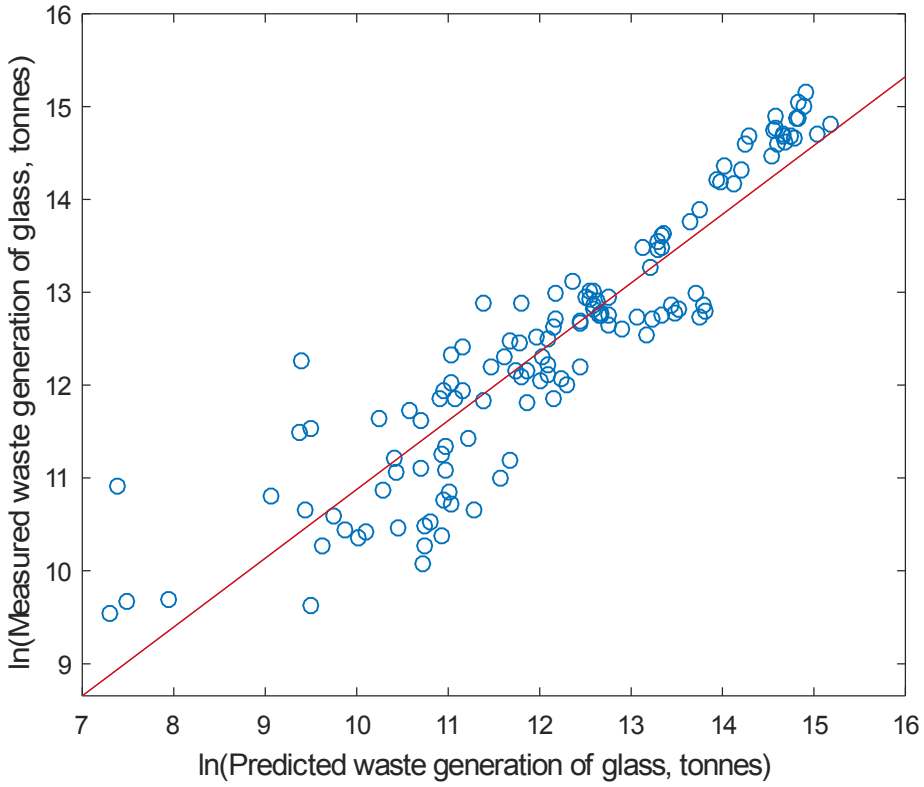

Fig. 20: Waste generation of glass

Logarithmic plot of the generation of glass, with predicted values against measured values. No potential outliers have been removed from the data. r = 0.956, r^2 *= 0.914.*

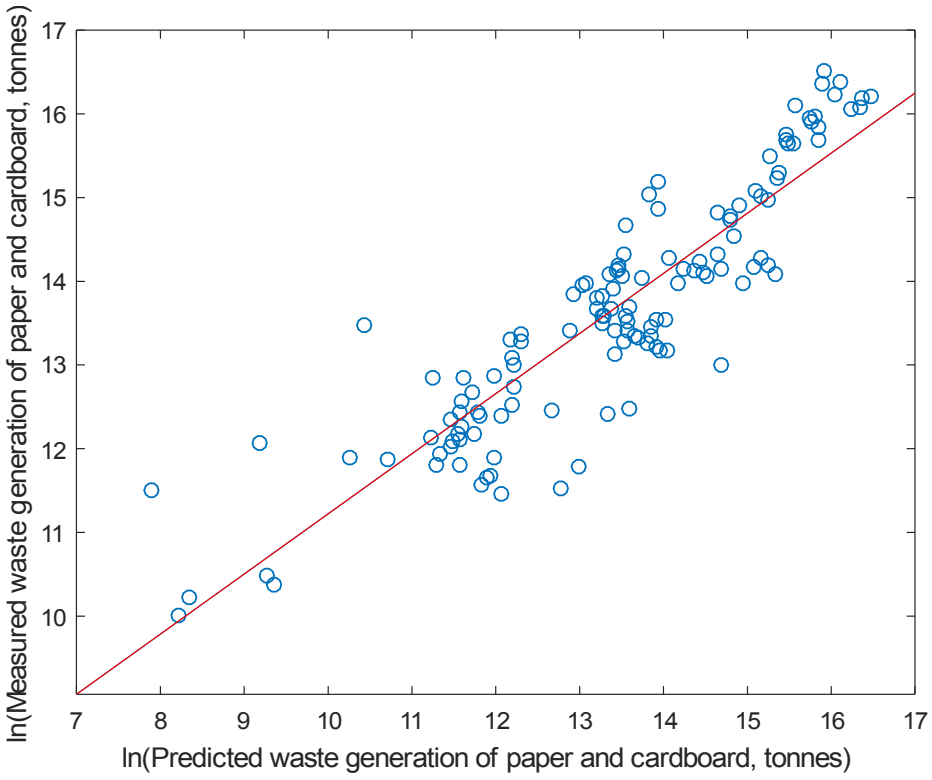

Fig. 21: Waste generation of paper and cardboard

Logarithmic plot of the generation of paper and cardboard, with predicted values against measured values. No potential outliers have been removed from the data. r = 0.918, r^2 *= 0.843.*

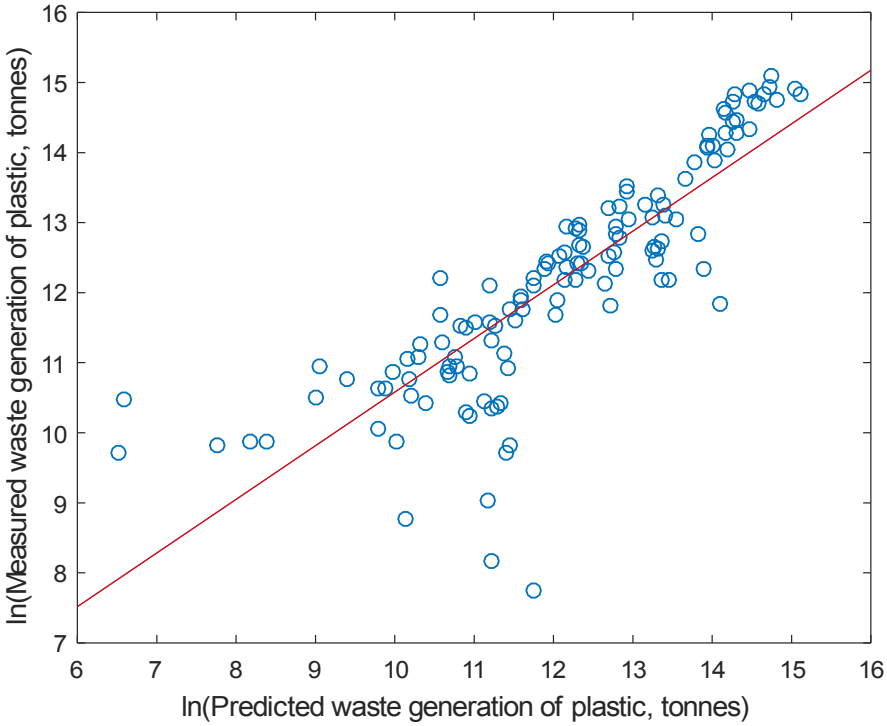

Fig. 22: Waste generation of plastic

Logarithmic plot of the generation of plastic, with predicted values against measured values. No potential outliers have been removed from the data. r = 0.931, r^2 *= 0.866.*

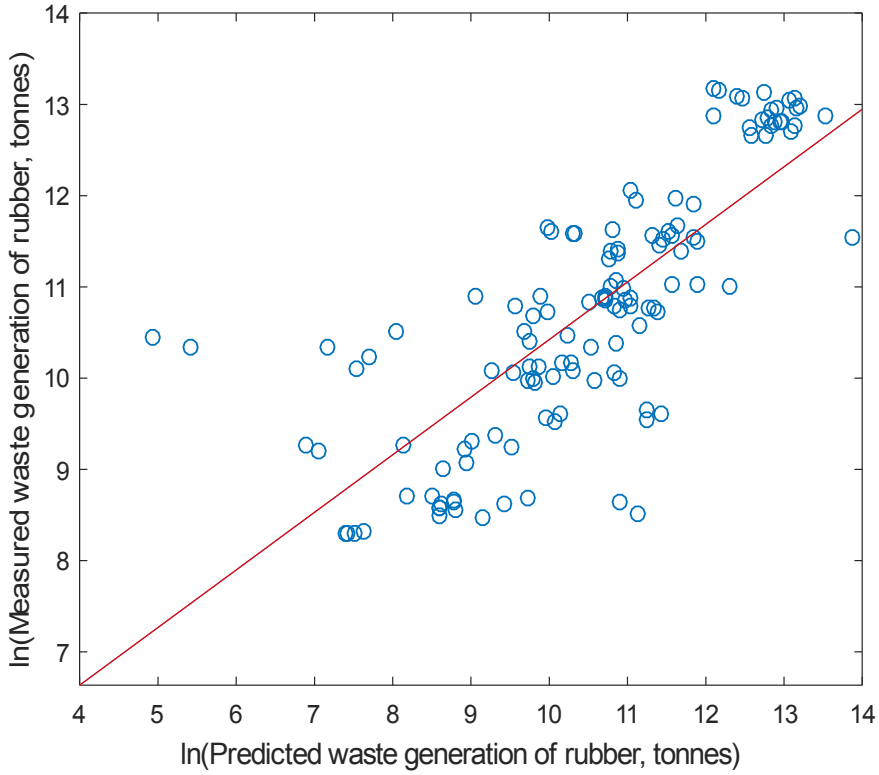

Fig. 23: Waste generation of rubber

Logarithmic plot of the generation of rubber, with predicted values against measured values. No potential outliers have been removed from the data. r = 0.754, r^2 *= 0.568.*

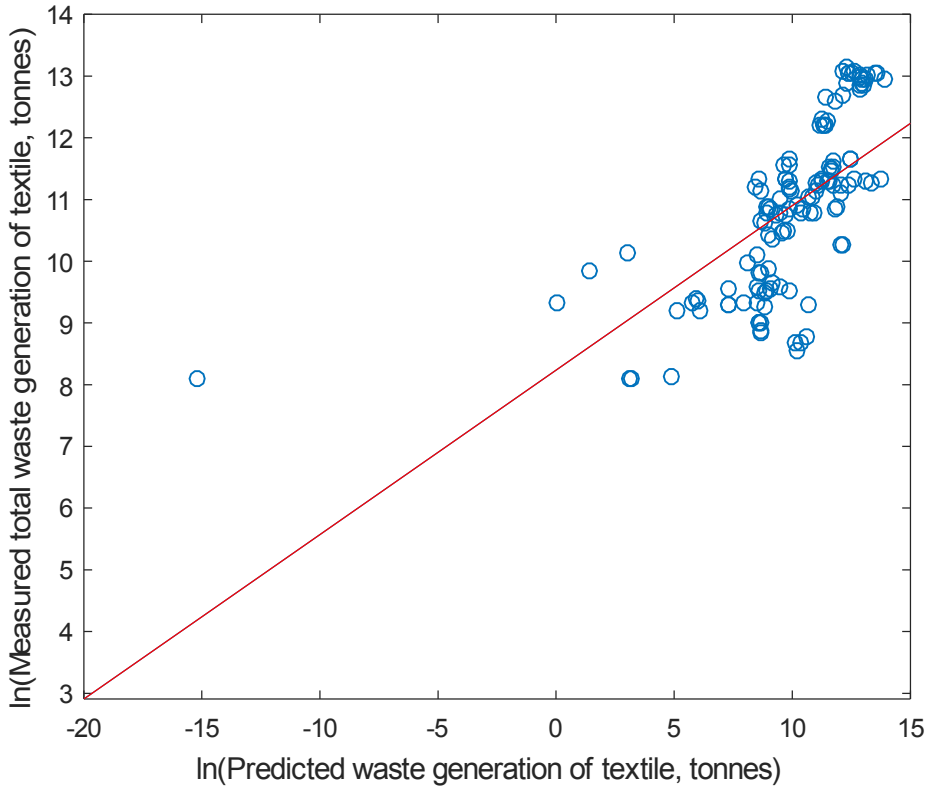

Fig. 24: Waste generation of textile

Logarithmic plot of the generation of textile, with predicted values against measured values. No potential outliers have been removed from the data. r = 0.669, r^2 *= 0.448.*

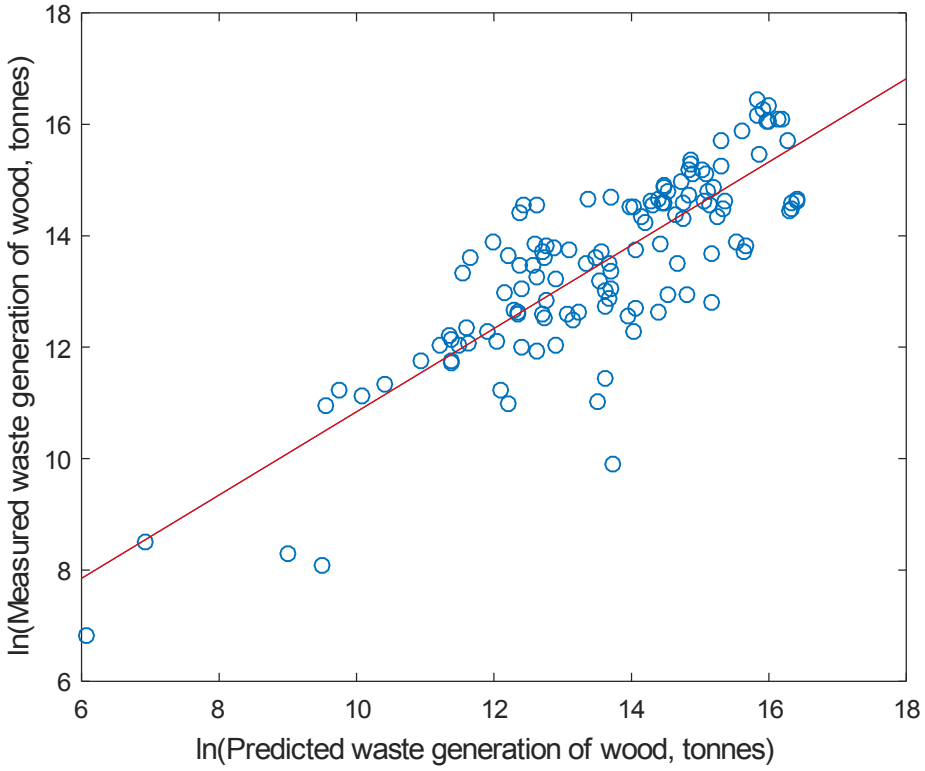

Fig. 25: Waste generation of wood

Logarithmic plot of the generation of wood, with predicted values against measured values. No potential outliers have been removed from the data. r = 0.659, r^2 *= 0.434.*

4.4.3 Cross-validation of results from multivariate regression

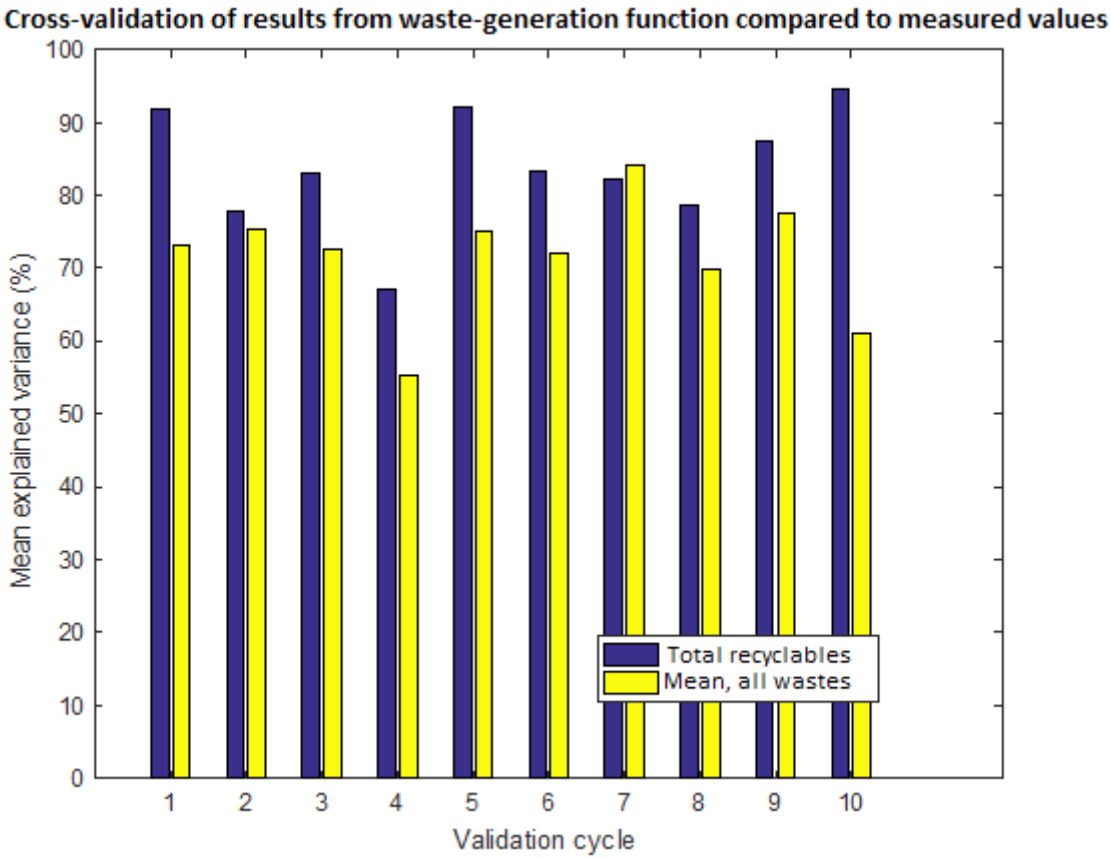

Fig. 26: Cross-validation test multivariate regression analysis

A 10-fold cross-validation of results from the multivariate regression analysis that resulted in the β-coefficients of the waste-generation function. The result is expressed as the mean explained variance of the measured values from the predicted values. Potential outliers have not been removed. The results in the picture show the prediction of the total generation of recyclable waste, Total recyclables', and the mean value of the generation of the six wastes glass; paper and cardboard; plastic; rubber; textile and wood, 'Mean, all waste generations'. In the cross-validation, new β:s are calculated in every cycle. The mean explained variance is based on r^2. r^2 *= 0.8 corresponds to a mean explained variance of 80%, etc.*

5. Discussion

5.1 Interpretation of results

The results from Data selection 2 shows a total explained variance of about 88%, including the seven data categories that was left out from the PPCA (see Table 2, p.25). This indicates that only a small part of the information from the original 105 data categories was lost. But, it would be sound to bring up that there were initially more than 4600 data categories to choose from in Eurostat's database (European Commission, 2016 V), and that it is highly unlikely that the 105 data categories chosen in Data selection 1 would cover all possible explaining factors for the generation of recyclable wastes. The cross-validation results from Data selection 2 (see Fig. 9-18 pp.26-27) show some cases of significantly higher prediction errors for validation data compared to training data. To mention the worst examples, see cycle 4 in PPCA 1, cycle 9 in PPCA 4, and cycle 3 in PPCA 8. There are also examples of the opposite, such as Cycle 6 in PPCA 2, cycle 5, PPCA 7 and Cycle 10, PPCA 9. The results of the cross-validation in every analysis are more stable for the training sets than for the validation sets, which is unsurprising. But the interpretation of the bigger picture are that the errors in the validation sets are not consistently much higher than for the training sets. Which suggests that the value of an explained variance of 88% is not very far from the truth.

The four candidates from Data selection 3 (see Table 3, p.28), low education level; primary energy consumption; LMP expenditure in millions Euro; and GDP in million PPS; are with a high probability the right choices of the eleven options. Together, they cover both the highest correlation with the generation of every considered waste. They are also the four candidates with the highest total correlation to the generation of all the considered wastes. There are other thinkable options for explaining factors from Data selection 3, such as domestic material consumption and energy from non-renewable waste, that could have improved the results of the final model. But for simplicity and to reduce the risk of a too high collinearity, these were left out.

The test of the waste-generation function's (see eq. 18, p.29) ability to predict measured values showed that the function is best at predicting the values of the total generation of recyclable wastes; glass; paper and cardboard; and plastic (see Table 5, p.30). With outliers, the results showed an explained variance in percent based on the coefficient of determination ($100r^2$) of 86.2, 91.4, 84.3 and 86.6%, respectively. The corresponding correlation coefficients r were 0.928, 0.956, 0.918, 0.931, respectively, which is considered a close model fit. The mean explained variance of the generation of the six considered wastes was 67.9% (r = 0.824). The predictions of rubber, textile and wood were consistently worse, as confirmed by the scatter plots (see Fig. 19-25, pp.30-33). The explained variance of rubber, textile and wood were, with outliers, 56.8, 44.8 and 43.4%, respectively. The cross-validation of the waste-generation function with outliers (see Fig. 26, p.34), and with ten different tests with new β-coefficients, was a test to confirm the concept of using multivariate regression to define the model with the four chosen explaining factors. The results of the predictions showed an explained variance in the range 67-95% for the total generation of recyclable waste, and an explained variance of 56-84% for the mean results of the generation of all six wastes. The interpretation of these figures, together with the relatively high predictive capacity of the waste-generation function for the total generation of recyclable wastes; glass; paper and cardboard; and plastic; is that a combination of PPCA and multivariate regression can deliver good results.

5.2 Verification of results

5.2.1 Quality of raw data

As discussed in subsection 2.5 (see p.13), the raw data that was used in this thesis was incomplete. As a rough estimate of 80% of the total values were available for the data used to find the explaining factors. As mentioned in subsection 2.5 some of the available values were forecasts, and there are examples of breaks in the time series. This has an impact on the results, although it is hard to say how great this impact is. Considering the high correlations (see Table 5, p.30) this impact is not considered to be high, at least not among the chosen explaining factors. The data of the explaining factors were almost entirely complete. The data on low education level included breaks in the time series, and the data on LMP expenditure included estimates. But if the data available for every individual dataset associated with the explaining factors would have been too incomplete, or include very deviating predicted values, it is unlikely that analyses based on these datasets would be highly correlated. The data of the recyclable wastes included only measured values or missing values, and no forecasts or breaks in time series. Only a small proportion of the data was missing from the datasets. The most part of the total missing values were found in the data on metal waste, which could explain the issue with collinearity discussed in subsection 3.3.5 (see Model evaluation and model improvement, p.22).

5.2.2 The difference between causality and correlation

In this study the concept correlation has a central role. It is important to point out the difference between causality and correlation. As an example, assume that a person has found a pattern between the turning of the steering wheel of a car, and the turning of the wheels. For some reason, the steering wheel and the wheels of the car move together in some direction. This represents correlation. If the person finds out that the steering wheel and the wheels are actually connected by the steering column so that one affects the other, causality has been proven. The results of this thesis are based on correlation, since the four explaining factors that were finally included in the model were chosen based on the correlation with the generation of the recyclable wastes. And for some reason about 86% of the variations in the generation of total recyclable waste could be explained by the amount of the popultion with a low education level, primary energy consumption, LMP expenditure and GDP. This is not proof that the explaining factors causes the generation of wastes. Perhaps they are not directly connected. But correlation prooves that the explaining factors for some reason, and to some extent, share patterns with the generation of the recyclable wastes. Even though causality has not been proven, the generation of recyclable wastes can be predicted by the movement of the explaining factors. The following subsections 5.2.3-5.2.6 contain arguments why the explaining factors and the generation of recyclable waste could be connected not only through correlation, but also through causality.

5.2.3 Low education level

Low education level (see Appendix 2 Nr: 28, p.46) is more precisely defined as the population aged 15 to 74 years with a highest attained education level of less than primary, primary and lower secondary education, expressed in thousands. To clarify, this group of people has not at least passed upper secondary school. Education level has earlier been known to affect the generation of residential solid waste in the Mexican city Mexicali. Benítez et. al. (2008, p.S10) showed that households with primary education produce more solid waste than households of higher education. Another study conducted in the Iranian city Ahvaz also showed a decrease in the generation of household solid waste with increasing education level (Monavari et. al. 2012, 1844). Chen (2010, p.451) showed that an increase in education level reduced the generation of municipal solid waste in Taiwan. The preceding cases suggests that the generation of recyclable waste in the EU would decrease with increasing education level, and the β-coefficients suggests this to be true for glass,

plastic, rubber and textile (see Table 4, p.29). But, the β-coefficients also imply that the total amount of recyclable waste; paper and cardboard; and wood increase heavily with increasing education level. These three relationships could be a mathematical construction, but they could also mean that education level plays different roles in the generation of different wastes. To conclude, there are reasons to believe that education level is related to the generation of waste, but this thesis does not show unambigously if the generation of recyclable waste either increases or decreases with an increasing education level.

5.2.4 Primary energy consumption

The primary energy consumption represents the total demand of energy for a country (European Commission, 2016 VIII). Primary energy consumption is in this case expressed in TOE, tonnes of oil equivalent (see Appendix 3 Nr: 52, p.47). One tonne of oil equivalent is the amount of energy available in one tonne of crude oil (European Commission, 2016 IX). The results of this thesis suggests that the total generation of recyclable waste increases with an increasing primary energy consumption (see Table 4, p.29). This claim have not been directly confirmed by previous studies. But it has been shown that the total primary energy supply per capita is strongly linked to domestic material consumption per capita (European Environment Agency 2012, p.12). A high material consumption is likely to be linked to a high generation of recyclable waste. Domestic material consumption was also among the top candidates for explaining the generation of recyclable wastes, although not as strongly correlated to the generation of recyclable wastes as primary energy consumption (see Table 3, p.28). It seems intuitively likely that a high energy demand within a country could also reflect a high presence of energy demanding processes, for example consumption, production and waste generation.

5.2.5 LMP expenditure

LMP (labour market policy) expenditure (see Appendix 4 Nr: 87, p.48) is the public financial aid in the labour market aimed at creating equilibrium and efficient functioning in favour of disadvantaged groups, for example unemployed or people trying to enter the labour market (European Commission, 2016 X). The results of this thesis indicates a relationship between a high LMP expenditure expressed in million Euro, and a low total generation of recyclable waste (see Table 4, p.29). This could reflect the creation of jobs within waste management and environmental protection in the period 2000-2012 (European Commission, 2016 XI). According to OECD (2014, p.112) governments play an important role in green growth through economic measures, which include the creation of green jobs.

5.2.6 GDP

The GDP in million PPS, or gross domestic product in million purchasing power standards (see Appendix 4 Nr: 89, p.48), is a measure of the economic activity of a country (European Commission, 2016 XII). The conversion to PPS, an artificial monetary unit, removes the differences in price levels between countries in the EU (ibid.). The results indicate that an increase in GDP in PPS would increase the total generation of recyclable waste (see Table 4, p.29). GDP is likely to have part in waste generation, since it is linked directly to domestic material consumption. Although having a close relationship up until a few years ago, signs of decoupling between GDP and domestic material consumption have started to show in the recent years (European Commission, 2016 XIII). This conclusion can be drawn from the resource productivity, which would have been constant if the GDP and the domestic material consumption would have had the same relationship over time.

Also, a higher GDP in PPS would correspond to a higher ability to buy goods and services. This would lead to more capital to spend on for example articles of consumption, which would likely result in a higher generation of waste. Daskalopoulos, Badr and Probert (1998, p.157) supports the claim that the total consumer expenditure increases with increasing GDP, and also that the generation of municipal solid waste increases with increasing GDP (op. cit., p.160). The relationships between GDP and domestic material consumption, and between GDP and consumer expenditure, are probably only parts of the truth, since GDP in million PPS is more strongly correlated with every generation of recyclable waste compared to domestic material consumption (see Table 3, p.28).

5.2.7 Model time dependancy

The waste-generation function presented in this thesis does not take variations over time into account, since the relationships between the explaining factors and the generation of recyclable wastes have been assumed constant over time. The signs of decoupling between GDP and domestic material consumption could indicate that the importance of GPD in PPS as a predictor for the generation of recyclable waste might diminish in the years to come. A decoupling between total primary energy supply and domestic material consumption (European Environment Agency 2012, p.12) over time would also reduce the accuracy of the model. The relationships between recyclable waste generation and LMP expenditure, and recyclable waste generation and education level, could also be time dependant. For this reason, it might be necessary to modify the model over time.

5.3 Model utility

In the introduction, some requirements were discussed that should be fulfilled by a model in order to provide information on waste treatment strategies for decision makers in the EU. Such a model should be simple, consider economical; social; and environmental aspects, be applicable from local up to international scale, be able to make predicitons in the near future and be able to make suggestions on how to reduce waste.

The waste-generation function presented in this thesis has some limitations, but adress some of the most important issues. It could be considered simple, since it only requires four measurements of four quantities – low education level, primary energy consumption, LMP expenditure and GDP - in order to make predictions of the generation of recyclable wastes. The waste-generation function appears to make fairly accurate predictions of the total generation of recyclable waste, as well as predictions of the waste generation of glass; paper and cardboard; and plastic. Predictions could also be made, although not as accurately, of the generation of rubber, textile and wood. More data on these wastes, which is provided continuously by Eurostat, could improve these results. Hopefully rubber, textile and wood could be predicted by the waste-generation function within the next few years.

Data covering economic, social and environmental aspects were included in the initial stage, but most of this data was later sorted away due to an insufficient correlation with the considered generation of wastes. It would be valuable for the broader understanding of the subject to include more suitable data covering these aspects in future research.

The waste-generation function could be considered applicable on a national and international scale in Europe, although the latter would require a redefinition of the explaining factors (for example a total GDP in million PPS for a group of countries). It is uncertain if the waste-generation function could also make accurate predictions on a local scale, which would also for this case require a redefinition of the explaining factors to match local scale. Predictions on local scale are also likely to be less precise than on a national or international scale, because it is harder to make accurate predictions of quantities linked to smaller populations than of larger. More research is encouraged to test the accuracy of the waste-generation function on a local scale.

Predictions for the near future should be possible although time dependancy of the explaining factors could become an issue, which would require a modification of the waste-generation function. Two of the main sources of decreasing prediction accuracy in the future are though to be the decoupling of the GDP and the domestic material consumption, and the decoupling between the total primary energy supply and the domestic material consumption.

The waste-generation function has shown results also supported by other studies. An increase in primary energy consumption and GDP seem to increase the total generation of recyclable waste. An increase in LMP expenditure seems to decrease the total generation of recyclable waste (which might be related to the creation of green jobs, although this has not been confirmed by other studies). Previous studies point out a relationship between an increase in education level and a decrease in waste generation. This study suggests that the generation of some of the recyclable wastes decrease with increasing education level, but that the generation of other wastes increase with increasing education level. This is yet to be confirmed or rejected by future research. To sum up – in order to reduce the total generation of recyclable waste one should consider GDP, primary energy consumption and LMP expenditure.

The model predictions and the just mentioned advices to reduce the total generation of recyclable waste are the areas of the highest relevance for the environmental sciences that this study can contribute to.

5.4 A prognosis for the year 2020

In this section, a prognosis for recyclable waste for the year 2020 will be presented using the waste-generation function (eq. 18). The prognosis is based on data from 2012 and 2014, and projected or goal values of the explaining factors for the year 2020. The low education level, the early school leavers which had not at least passed higher secondary school, was 31.4% in the EU in 2012. This corresponded to about 116.4 million people. The EU has set a target to reduce the percentage of early school leavers to 10% by 2020 (European Commission 2016 XIV, p.1), which will be assumed that the EU can accomplish. The population of the EU in 2014 was 507 million (European Union 2015a, p.45), which is expected to rise with about 1% by 2020 (European Commission, 2016 XV). An early school leaver ratio of 10% by 2020 would based on the preceding information correspond to 51.2 million people. The primary energy consumption in the EU was 1584.0 million TOE in 2012. The EU has the goal to reduce this figure to 1 474 million TOE by 2020 (European Commission 2016 XVI, p.4), which will be assumed that they will. According to Gros and Alcidi (2013, p.59) the GDP in PPS in the EU is projected to grow at a constant rate to an approximate 40% increase in 2030 compared to 2010. This prediction will be assumed accurate. Projections for LMP expenditure have not been found, and these expenditures will be assumed constant during the period 2014-2020. Based on the previous assumptions and inserting the measured values for the generation of wastes from 2012 (see Appendix 5, p.49), the following predictions were reached for the year 2020 compared to 2012:

- +12.4% glass waste
- +38.9% paper and cardboard waste
- +21.2% plastic waste
- +24.4% total recyclable waste

As a comparison the predicted increase in total generation of recyclable waste in 2012-2020, with the expected low population increase in 2012-2020, show a similar development as the total material recycling per capita in the EU in 2005-2013 (European Union 2015b, p.91). If the predictions are accurate they do not necessarily reflect a pure increase in waste generation. They could also reflect a trend of hightened recycling rates.

6. Conclusions

In this thesis, PPCA and multivariate regression has been used to define a waste-generation function that could predict the waste generation of recyclable wastes on a national scale in the EU. PPCA was successfully used together with correlation analysis to to find four factors with a high correlation to the generation of recyclable wastes – low education level, LMP expenditure, GDP, and primary energy consumption. These four appeared important to consider in order to find strategies to reduce the generation of recyclable wastes in the EU in the future. The total waste generation appeared to increase with increasing primary energy consumption and GDP, and decrease with increasing LMP expenditure. The contribution from a low education level showed ambiguous results.

The waste-generation function could explain just above 86% of the total generation of recyclable waste (r = 0.928), and an average of nearly 68% of the generation of glass; paper and cardboard; plastic; rubber; textile and wood (r = 0.824). Metal waste was excluded from the model due to a high collinearity. The waste-generation function described the generation of the individual wastes of glass; paper and cardboard; and plastic to a higher degree, and the generation of rubber, textile and wood to a lower degree. More data could hopefully lead to better predictions of the waste generation of rubber, textile and wood in the near future.

The waste-generation function is likely to be time dependant because of a possibly emerging decoupling between material consumption and GDP, and between material consumption and primary energy supply. Nevertheless, the model prediction for 2020, a 24.4% increase in the total generation of recyclable waste in the EU, appeared to be in line with the development of the generation of municipal recyclable waste in the recent years.

We live in a time with increasing resource flows, and it is important that we do our best to make the most of our limited resources. One part of this is to learn how to predict the quantities of available recyclable resources, so that they can be used in an efficient manner. Hopefully, this thesis will constitute a small contribution in this struggle.

7. Acknowledgements

I would like to thank my supervisors Johanna Alkan Olsson and Nina Reistad for giving discussions, good advice and for their belief in my ideas.

8. References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. 3. ed. New Jersey: John Wiley & Sons, Inc.

Beigl, P., Lebersorger, S. & Salhofer, S. (2008). Modelling municipal solid waste generation: A review. *Waste Management* 28(1): 200–214. DOI: 10.1016/j.wasman.2006.12.011.

Beigl, P., Wassermann, G., Schneider, F. & Salhofer, S. (2004). Forecasting Municipal Solid Waste Generation in Major European Cities. *In iEMSs 2004 International Congress: Complexity and Integrated Resources Management.*

Belsley, D. A., Kuh, E. & R. E. Welsh. (2005). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New Jersey: John Wiley & Sons, Inc. E-book. DOI: 10.1002/0471725153.fmatter.

Benítez, S. O., Lozano-Olvera, G., Morelos R. A. &Vega, C. A. d. (2008). Mathematical modeling to predict residential solid waste generation. *Waste Management* 28: S7-S13. DOI: 10.1016/j.wasman.2008.03.020.

Camacho, J. & Ferrer, A. (2014). Cross-validation in PCA models with the element-wise k -fold (ekf) algorithm: Practical aspects. *Chemometrics and Intelligent Laboratory Systems* 131: 37–50.
DOI: 10.1016/j.chemolab.2013.12.003.

Chen, C. C. (2010). Spatial inequality in municipal solid waste disposal across regions in developing countries. *International Journal of Environmental Science and Technology* 7(3): 447-456. DOI: 10.1007/BF03326154.

Chen, H. (2003). *Principle Component Analysis With Missing Data and Outliers*. New Jersey: Rutgers University, Department of Electrical and Computer Engineering.
http://www.nec-labs.com/~haifeng/mypubs/tutorialrpca.pdf (Accessed: 2016-05-10).

Daskalopoulos, E., Badr, O. & Probert, S.D. (1998). Municipal solid waste: a prediction methodology for the generation rate and composition in the European Union countries and the United States of America. *Resources, Conservation and Recycling* 24(2): 155-166. DOI: 10.1016/S0921-3449(98)00032-9.

Emmerich, Greg. (2013). *Demystifying Big Data: Skytree Brings Machine Learning to the Masses.* Thesis for M. Sc., University of Wisconsin-Madison.

European Commission. (2016 I). Database. *eurostat – Your key to European statistics*.
http://ec.europa.eu/eurostat/data/database (Accessed: 2016-02-16).

European Commission. (2016 II). What we do. *eurostat – Your key to European statistics*.
http://ec.europa.eu/eurostat/about/overview/what-we-do (Accessed: 2016-04-21).

European Commission. (2016 III). Who does what. *eurostat – Your key to European statistics*.
http://ec.europa.eu/eurostat/about/overview/who-does-what (Accessed: 2016-04-21).

European Commission. (2016 IV). Overview. *eurostat – Your key to European statistics*.
http://ec.europa.eu/eurostat/web/quality/overview (Accessed: 2016-04-21).

European Commission. (2016 V). Database. *eurostat – Your key to European statistics*.
http://ec.europa.eu/eurostat/help/first-visit/database (Accessed: 2016-04-21).

European Commission. (2016 VI). Eurostat Data Navigation Tree. *eurostat – Your key to European statistics*.
http://ec.europa.eu/dgs/eurostat/contingency/table_of_contents_en.pdf (Accessed: 2016-04-21).

European Commission. (2016 VII). NACE background. *eurostat – Your key to European statistics*.
http://ec.europa.eu/eurostat/statistics-explained/index.php/NACE_background#Structure_and_coding_of_NACE
(Accessed: 2016-04-24).

European Commission. (2016 VIII). Glossary:Primary energy consumption. *eurostat – Your key to European statistics*. http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Primary_energy_consumption (Accessed: 2016-05-12).

European Commission. (2016 IX). Glossary:Tonnes of oil equivalent (toe). *eurostat – Your key to European statistics*. http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Tonnes_of_oil_equivalent_(toe) (Accessed: 2016-05-12).

European Commission. (2016 X). Labour market policy. *eurostat – Your key to European statistics*. http://ec.europa.eu/eurostat/web/labour-market/labour-market-policy (Accessed: 2016-05-16).

European Commission. (2016 XI).
Environmental economy – employment and growth. *eurostat – Your key to European statistics*. http://ec.europa.eu/eurostat/statistics-explained/index.php/Environmental_economy_-_employment_and_growth (Accessed: 2016-05-16).

European Commission. (2016 XII). GDP per capita in PPS. *eurostat – Your key to European statistics*. http://ec.europa.eu/eurostat/web/products-datasets/-/tec00114 (Accessed: 2016-05-17).

European Commission. (2016 XIII). Resource productivity statistics. *eurostat – Your key to European statistics*. http://ec.europa.eu/eurostat/statistics-explained/index.php/Resource_productivity_statistics (Accessed: 2016-05-17).

European Commission. (2016 XIV). Europe 2020 Target: Early leavers from education and training. *eurostat – Your key to European statistics*. http://ec.europa.eu/europe2020/pdf/themes/29_early_school_leaving.pdf (Accessed: 2016-05-18).

European Commission. (2016 XV). People in the EU – population projections. *eurostat – Your key to European statistics*. http://ec.europa.eu/eurostat/statistics-explained/index.php/People_in_the_EU_%E2%80%93_population_projections#Europop2013_.E2.80.94_population_projections (Accessed: 2016-05-18).

European Commission. (2016 XVI). Europe 2020 Targets: Climate change and energy. *eurostat – Your key to European statistics*. http://ec.europa.eu/europe2020/pdf/themes/16_energy_and_ghg.pdf (Accessed: 2016-05-18).

European Environment Agency. (2012). *The European Environment: State and Outlook 2010. Material Resources and Waste – 2012 update.* Copenhagen: European Environment Agency.

European Union. (2013). *Manual on waste statistics - A handbook for data collection on waste generation and treatment - 2013 edition*. DOI: 10.2785/4198.

European Union. (2015a). *Eurostat Regional Yearbook 2015*. E-book. http://ec.europa.eu/eurostat/web/products-statistical-books/-/KS-HA-15-001 (Accessed: 2016-05-18).

European Union. (2015b). *Sustainable developement in the European Union – 2015 monitoring report of the EU Sustainable Development Strategy.* E-book. http://ec.europa.eu/eurostat/documents/3217494/6975281/KS-GT-15-001-EN-N.pdf/5a20c781-e6e4-4695-b33d-9f502a30383f (Accessed: 2016-05-19).

Gros, D. & Alcidi, C. (2013). *The Global Economy in 2030: Trends and Strategies for Europe*. Brussels: Centre for European Policy Studies. E-book. http://europa.eu/espas/pdf/espas-report-economy.pdf (Accessed: 2016-05-18).

Hidalgo, B. & Goodman, M. (2013). Multivariate or Multivariable Regression? *American Journal of Public Health* 103(1): 39–40. DOI: 10.2105/AJPH.2012.300897

Härdle, W. & Simar, L. (2007). *Applied Multivariate Statistical Analysis*. 2. ed. Berlin: Springer-Verlag Berlin Heidelberg.

Ilin, A. & Raiko, T. (2010). Practical Approaches to Principal Component Analysis in the Presence of Missing Values. *Journal of Machine Learning Research* 11: 1957–2000. http://www.jmlr.org/papers/volume11/ilin10a/ilin10a.pdf (Accessed: 2016-05-10).

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning*. New York: Springer Publishing Company, Inc.

Jolliffe, I. T. (1972). Discarding Variables in a Principal Component Analysis. I: Artificial data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 21 (2): 160-173. http://eds.b.ebscohost.com.ludwig.lub.lu.se/eds/pdfviewer/pdfviewer?vid=2&sid=a2ec5fda-1273-4ef4-ae16-f67e6bbcb5b1@sessionmgr106&hid=111&preview=false (Accessed: 2016-05-10).

Jolliffe, I. T. (2002). *Principal Component Analysis*. 2. ed. New York: Springer-Verlag New York, Inc.

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational & Psychological Measurement* 20 (1): 141–151. DOI: 10.1177/001316446002000116.

Katasamaki, A., Willems, S., Diamadopoulos, E. (1998). Time series analysis of municipal solid waste generation rates. *Journal of Environmental Engineering* 124 (2): 178–183. DOI: 10.1061/(ASCE)0733-9372(1998)124:2(178).

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* 2 (12):1137-1143. http://web.cs.iastate.edu/~jtian/cs573/Papers/Kohavi-IJCAI-95.pdf (Accessed: 2016-05-10).

Livingstone, D. (2009). *A Practical Guide to Scientific Data Analysis*. Chichester, UK: Wiley.

Monavari, S. M., Omrani, G. A., Raof, F. F., Karbassi, A. (2012). The effects of socioeconomic parameters on household solid-waste generation and composition in developing countries (a case study: Ahvaz, Iran). *Environmental Monitoring and Assessment* 184(4): 1841-1846. DOI: 10.1007/s10661-011-2082-y.

OECD. (2014). *Green Growth Indicators 2014.* OECD Green Growth Studies. OECD publishing. DOI: 10.1787/9789264202030-en.

Pires, A., Martinho, G. & Chang, N. (2011). Solid waste management in European countries : A review of systems analysis techniques. *Journal of Environmental Management* 92(4): 1033–1050. DOI: 10.1016/j.jenvman.2010.11.024.

Pukelsheim, F. (1994). The Three Sigma Rule. *The American Statistician* 48 (2): 88-91. DOI: 10.1080/00031305.1994.10476030.

Rodríguez J. D., Pérez, A. & Lozano, J.A. (2010). Sensitivity Analysis of *k*-Fold Cross Validation in Prediction Error Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (3): 569–575. DOI: 10.1109/TPAMI.2009.187.

Soltani, A., Sadiq, R., & Hewage, K. (2016). Selecting sustainable waste-to-energy technologies for municipal solid waste treatment : a game theory approach for group decision making. *Journal of Cleaner Production* 113: 388–399. DOI: 10.1016/j.jclepro.2015.12.041.

The Mathworks Inc. (2016a). ppca. *Mathworks – Makers of MATLAB and Simulink.* http://se.mathworks.com/help/stats/ppca.html?searchHighlight=ppca (Accessed: 2016-04-14).

The Mathworks Inc. (2016b). mvregress. *Mathworks – Makers of MATLAB and Simulink.* http://se.mathworks.com/help/stats/mvregress.html (Accessed: 2016-04-19).

The Mathworks Inc. (2016c). collintest. *Mathworks – Makers of MATLAB and Simulink.* http://se.mathworks.com/help/econ/collintest.html?refresh=true (Accessed: 2016-05-06).

The Mathworks Inc. (2016d). zscore. *Mathworks – Makers of MATLAB and Simulink.* http://se.mathworks.com/help/stats/zscore.html?refresh=true (Accessed: 2016-04-19).

Tipping, M. E. & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B* 61: 611-622. DOI: 10.1111/1467-9868.00196.

Wang, F. S., Richardson, A. J. & Roddick, F. A. (1996). SWIM – A computer model for solid waste integrated management. *Computers, Environment and Urban Systems* 20(4): 233-246. DOI: 10.1016/s0198-9715(96)00019-

Appendices

Appendix 1 – Environmental focus factors, sorted by code and subgroup

Bolded *= passed selection 2*

Nr	Eurostat code	Description	Subgroup	Unit
1	env_ac_exp2	Environmental protection expenditure in Europe - EUR per capita and % of GDP	General government	%
2	**env_ac_exp2**	**Environmental protection expenditure in Europe - EUR per capita and % of GDP**	**Industry (except construction, sewage, waste management and remediation activities)**	**%**
3	env_ac_tax	Environmental tax revenues	Total environmental taxes, percentage of total revenues from taxes and social contributions (including imputed social contributions)	%
4	food_act2	Number of certified registered organic operators by type of operators	Registered operators at the end of the year	-
5	**food_in_porg1**	**Certified organic crop area by crops products**	**Total organic crop area**	**Ha**
6	**nrg_107a**	**Supply, transformation, consumption - renewable energies - annual data**	**Gross inland consumption**	**TJ**
7	nrg_107a	Supply, transformation, consumption - renewable energies - annual data	Primary production of biogas	TJ
8	nrg_107a	Supply, transformation, consumption - renewable energies - annual data	Primary production of biomass and renewable wastes	TJ
9	nrg_107a	Supply, transformation, consumption - renewable energies - annual data	Primary production of hydro power	TJ
10	nrg_107a	Supply, transformation, consumption - renewable energies - annual data	Primary production of municipal waste (renewable)	TJ
11	nrg_107a	Supply, transformation, consumption - renewable energies - annual data	Primary production of renewable energies	TJ
12	nrg_107a	Supply, transformation, consumption - renewable energies - annual data	Primary production of solar photovoltaic	TJ
13	**nrg_107a**	**Supply, transformation, consumption - renewable energies - annual data**	**Primary production of solar thermal energy**	**TJ**
14	nrg_107a	Supply, transformation, consumption - renewable energies - annual data	Primary production of solid biofuels (excluding charcoal)	TJ
15	nrg_107a	Supply, transformation, consumption - renewable energies - annual data	Primary production of wind power	TJ
16	**nrg_ind_335a**	**Share of energy from renewable sources**	**Share of gross final energy consumption**	**%**
17	t2020_rt120	Recycling rate of municipal waste	-	%
18	tsdcc350	Combined heat and power generation	Percentage of gross electricity generation	%
19	**tsdpc410**	**Organisations and sites with EMAS (Eco-Management and Audit Scheme) registration**	**EMAS organisations**	**-**
20	tsdpc410	Organisations and sites with EMAS (Eco-Management and Audit Scheme) registration	Sites	-

Appendix 2 – Informational and technological factors, sorted by code and subgroup

Bolded = passed selection 2

Nr	Eurostat code	Description	Subgroup	Unit
21	edat_lfs_9903	Population, aged 15 to 74 years, by educational attainment level, sex and age	Less than primary, primary and lower secondary education (levels 0-2)	%
22	hrst_st_ncat	Human resources in science and technology (HRST) by sub-groups, sex and age	Persons employed in science and technology, from 25-64 years. Total	%
23	htec_si_exp4	High-tech exports - Exports of high technology products as a share of total exports (from 2007, SITC Rev. 4)	Percentage of total	%
24	htec_trd_tot4	Total high-tech trade in million euro and as a percentage of total (from 2007, SITC Rev. 4)	Exports. All countries of the world	%
25	htec_sti_exp	Business enterprise R&D expenditure in high-tech sectors - NACE Rev. 1.1	Total - all NACE activities	Million €
26	htec_sti_exp2	Business enterprise R&D expenditure in high-tech sectors - NACE Rev. 2	Total - All NACE activities	Million €
27	**isoc_bde15ag**	**Percentage of the Information and Communication Technology sector on GDP**	**ICT. Total**	**%**
28	**lfsa_pgaed**	**Population, aged 15 to 74 years, by sex, age and highest level of education attained**	**Less than primary, primary and lower secondary education (levels 0-2)**	**1000**
29	lfsa_pgaied	Population, aged 15 to 74 years, by sex, age and participation in education or training (last 4 weeks)	Total	1000
30	**rd_e_berdindr2**	**Business enterprise R&D expenditure (BERD) by economic activity (NACE Rev. 2) (rd_e_berdindr2)**	**Percentage of GDP**	**%**
31	rd_e_gerdtot	Total intramural R&D expenditure (GERD) by sectors of performance	All sectors	%
32	tps00052	Shool expectancy	-	Years
33	trng_lfse_01	Participation rate in education and training (last 4 weeks) by sex and age	From 25 to 64 years. Total	%
34	tsdec320	Total R&D expenditure	Percentage av GDP	%

Appendix 3 – Productional factors, sorted by code and subgroup

Bolded = passed selection 2

Nr	Eurostat code	Description	Subgroup	Unit
35	apro_mt_lscatl	Cattle population - annual data	Live bovine animals. December	1000
36	bd_9ac_l_form_r2	Business demography by legal form (from 2004 onwards, NACE Rev. 2)	Population of active enterprises in t. Total. Business economy except activities of holding companies	-

37	env_waselvt	End-of-life vehicles: Reuse, recycling and recovery, Totals	Waste generated	t
38	**for_basic**	**Roundwood, fuelwood and other basic products**	**Roundwood. Total**	**1000 m^3**
39	**for_irspec**	**Industrial roundwood by species**	**Import. Total**	**1000 m^3**
40	**for_swspec**	**Sawnwood trade by species**	**Import. Total**	**1000 m^3**
41	nama_10_gdp	GDP and main components (output, expenditure and income)	Final consumption expenditure of general government. Current prices, million PPS	Million PPS
42	**nrg_100a**	**Simplified energy balances - annual data**	**Gross inland consumption**	**TJ**
43	nrg_100a	Simplified energy balances - annual data	Primary production of gas	TJ
44	nrg_100a	Simplified energy balances - annual data	Primary production of nuclear heat	TJ
45	nrg_100a	Simplified energy balances - annual data	Primary production of solid fuels	TJ
46	nrg_100a	Simplified energy balances - annual data	Primary production of total petroleum products	TJ
47	**nrg_100a**	**Simplified energy balances - annual data**	**Waste (non-renewable)**	**TJ**
48	nrg_104a	Supply, transformation - nuclear energy - annual data	Nuclear heat. Gross inland consumption	TJ
49	nrg_105a	Supply, transformation, consumption - electricity - annual data	Electrical energy available for final consumption	GWh
50	**nrg_109a**	**Primary production - all products - annual data**	**All products**	**TJ**
51	nrg_ind_334a	Energy saving - annual data	Final energy consumption	Million TOE
52	**nrg_ind_334a**	**Energy saving - annual data**	**Primary energy consumption**	**Million TOE**
53	rail_go_typeall	Railway transport - Goods transported, by type of transport	Total transport	1000 t
54	road_eqs_carhab	Passenger cars per 1 000 inhabitants	-	/1000
55	road_go_ta_tott	Summary of annual road freight transport by type of operation and type of transport	Loaded national transport. Total	1000 t
56	sts_copr_a	Production in construction - annual data	Volume index, construction. Calendar adjusted	2010=100
57	sts_inlb_a	Labour input in industry - annual data	Industry and construction (except sewerage, waste management and remediation activities). Calendar adjusted	Hours
58	**sts_intv_a**	**Turnover in industry, total - annual data**	**Manufacture of basic metals. Calendar adjusted**	**2010=100**
59	**sts_intv_a**	**Turnover in industry, total - annual data**	**Manufacture of paper and paper products. Calendar adjusted**	**2010=100**
60	**sts_intv_a**	**Turnover in industry, total - annual data**	**Manufacture of textiles. Calendar adjusted**	**2010=100**
61	**sts_intv_a**	**Turnover in industry, total - annual data**	**Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials. Calendar adjusted**	**2010=100**
62	sts_intvd_a	Turnover in industry, domestic market - annual data	Manufacturing. Calendar adjusted	2010=100
63	sts_trtu_a	Turnover and volume of sales in wholesale and retail trade - annual data	Wholesale trade, except of motor vehicles and motorcycles. Total index of turnover. Calendar adjusted	2010=100
64	tps00001	Population on 1 January	Total	-
65	tps00003	Population density	-	km^{-2}
66	tran_hv_frtra	Volume of freight transport relative to GDP	-	2000=100

Appendix 4 – Standard of living factors, sorted by code and subgroup

Bolded = passed selection 2

Nr	Eurostat code	Description	Subgroup	Unit
67	apri_pi10_ina	Price indices of the means of agricultural production, input - annual data	Goods and services currently consumed in agriculture (Input 1). Real index	2010=100
68	**earn_nt_net**	**Annual net earnings**	**Two-earner married couple, one at 100%, the other at 100% of Average Worker, with two children**	**PPS**
69	env_ac_mfa	Material flow accounts	Domestic material consumption. Total. Tonnes per capita	t/capita
70	**env_ac_mfa**	**Material flow accounts**	**Domestic material consumption. Total. Thousand tonnes**	**1 000 t**
71	env_ac_rp	Resource productivity	Euro per kilogram	€/kg
72	env_ac_rp	Resource productivity	PPS per kilogram	PPS/kg
73	env_waselee	Waste Electrical and Electronic Equipment (WEEE)	Total waste. Waste collected from households. Kilograms per capita	kg/capita
74	**ilc_li21**	**Persistent at-risk-of-poverty rate by sex and age (source: SILC)**	**Cut-off point: 60% of median equivalised income. Total**	**%**
75	**ilc_mddd11**	**Severe material deprivation rate by age and sex**	**Percentage of total population**	**%**
76	**ilc_mded03**	**Total housing costs in PPS (source: SILC)**	**Total**	**PPS**
77	ilc_mdes03	Inability to afford a meal with meat, chicken, fish (or vegetarian equivalent) every second day (source: SILC)	Household type – total. Income group – total. Percentage of total population	%
78	ilc_mdes04	Inability to face unexpected financial expenses (source: SILC)	Household type – total. Income group – total. Percentage of total population	%
79	ilc_peps01	People at risk of poverty or social exclusion by age and sex	Age – total. Sex – total. Percentage of total population	%
80	ilc_sip8	Material deprivation rate - Economic strain and durables dimension (source: SILC)	Age – total. Sex – total. Number of items – 3 or more. Percentage of total population	%
81	ilc_sis4	Mean number of deprivation items among the deprived - Economic strain and durables dimension (source: SILC)	Age – total. Sex – total. Average	#
82	**isoc_bde15cbc**	**E-banking and e-commerce**	**Last online purchase: in the 12 months. All individuals**	**%**
83	isoc_bde15cua	Internet use and activities	All Individuals. Frequency of Internet access: once a week (including every day). Percentage of individuals	%
84	isoc_bdek_di	Digital inclusion - individuals	Frequency of Internet access: once a week (including every day). All individuals. Percentage of individuals.	%
85	**isoc_bdek_smi**	**Digital single market - promoting e-commerce for individuals**	**Last online purchase: in the 12 months. All individuals**	**%**
86	**lfsa_ergan**	**Employment rates by sex, age and nationality**	**From 15 to 64 years**	**%**
87	**lmp_expsumm**	**Labour Market Policy expenditure by type of action - summary tables (source: DG EMPL)**	**LMP expenditure - total (categories 1-9). Expenditure type - total. Million euro**	**Million €**
88	lmp_expsumm	Labour Market Policy expenditure by type of action - summary tables (source: DG EMPL)	LMP expenditure - total (categories 1-9). Expenditure type - total. Percentage of GDP	%
89	**nama_10_gdp**	**GDP and main components (output, expenditure and income)**	**Gross domestic product at market prices. Current prices, million PPS**	**Million PPS**
90	nama_10_gdp	GDP and main components (output, expenditure and income)	Gross domestic product at market prices. Million euro	Million €
91	nama_10_pc	Main GDP aggregates per capita	Gross domestic product at market prices. Current prices, PPS per capita	PPS/capita
92	nama_aux_gph	GDP per capita - annual data	Real GDP per capita	€/capita

93	nama_co3_c	Final consumption expenditure of households by consumption purpose - COICOP 3 digit - aggregates at current prices	COICOP – total. Euro per inhabitant	€/capita
94	nama_co3_c	Final consumption expenditure of households by consumption purpose - COICOP 3 digit - aggregates at current prices	COICOP – total. Million euro	Million €
95	nama_fcs_c	Final consumption aggregates - Current prices	Final consumption expenditure. Euro per inhabitant	€/capita
96	nama_fcs_c	Final consumption aggregates - Current prices	Final consumption expenditure. Million euro	Million €
97	nama_gdp_c	GDP and main components - Current prices	Gross domestic product at market prices. PPS per inhabitant	PPS/capita
98	prc_hicp_aind	Harmonised index of consumer prices (HICP) - annual data (average index and rate of change)	All-items HICP. Annual average index	2015 = 100
99	**prc_ppp_ind**	**Purchasing power parities (PPPs), price level indices and real expenditures for ESA2010 aggregates**	**Purchasing power parities**	**EU28 = 1**
100	**tec00102**	**Net debt-to-income ratio, after taxes, of non-financial corporations**	**-**	**%**
101	tec00104	Gross debt-to-income ratio of households	-	%
102	**tec00113**	**Real adjusted gross disposable income of households per capita**	**-**	**PPS**
103	tsdpc310	Electricity consumption by households	1000 tonnes of oil equivalent	1000 TOE
104	urt_e3gdp	Gross domestic product (GDP), market prices	Intermediate regions	Million €
105	yth_incl_070	Severe material deprivation of young people by sex and age	Percentage of total population. Sex – total. Age – from 15 to 29 years.	%

Appendix 5 – Generation of recyclable wastes, sorted by description

Eurostat code	Description	Subgroup	Unit
env_wasgen	Glass wastes	All NACE activities plus households. Total	t
env_wasgen	Metallic wastes (W061+W062+W063)	All NACE activities plus households. Total	t
env_wasgen	Paper and cardboard wastes	All NACE activities plus households. Total	t
env_wasgen	Plastic wastes	All NACE activities plus households. Total	t
env_wasgen	Recyclable wastes (subtotal, W06+W07 except W077)	All NACE activities plus households. Total	t
env_wasgen	Rubber wastes	All NACE activities plus households. Total	t
env_wasgen	Textile wastes	All NACE activities plus households. Total	t
env_wasgen	Wood wastes	All NACE activities plus households. Total	t