Journal of Environmental Management 291 (2021) 112584

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: http://www.elsevier.com/locate/jenvman

Research article

Predictive modelling as a tool for effective municipal waste management policy at different territorial levels

Martin Rosecký [a,*], Radovan Šomplák [b], Jan Slavík [c], Jiří Kalina [d], Gabriela Bulková [e], Josef Bednář [a]

[a] Institute of Mathematics, Faculty of Mechanical Engineering, Brno University of Technology – VUT Brno, Technická 2896/2, 616 69, Brno, Czech Republic
[b] Institute of Process Engineering, Faculty of Mechanical Engineering, Brno University of Technology – VUT Brno, Technická 2896/2, 616 69, Brno, Czech Republic
[c] IEEP, Institute for Economic and Environmental Policy, Jan Evangelista Purkyně University, Moskevska 54, 400 96, Ústí nad Labem, Czech Republic
[d] Research Centre for Toxic Compounds in the Environment (RECETOX), Masaryk University, 625 00, Brno, Czech Republic
[e] Ministry of the Environment, Vršovická 65, 100 10, Praha 10, Czech Republic

ARTICLE INFO

Keywords:
Municipal waste generation
Territorial levels
Regression modelling
Machine learning
Socio-economic factors
Public policy

ABSTRACT

Nowadays, the European municipal waste management policy based on the circular economy paradigm demands the closing of material and financial loops at all territorial levels of public administration. The effective planning of treatment capacities (especially sorting plants, recycling, and energy recovery facilities) and municipal waste management policy requires an accurate prognosis of municipal waste generation, and therefore, the knowledge of behavioral, socio-economic, and demographic factors influencing the waste management (and recycling) behavior of households, and other municipal waste producers. To enable public bodies at different territorial levels to undertake an effective action resulting in circular economy we evaluated various factors influencing the generation of municipal waste fractions at regional, micro-regional and municipal level in the Czech Republic. Principal components were used as input for traditional models (multivariable linear regression, generalized linear model) as well as tree-based machine learning models (regression trees, random forest, gradient boosted regression trees). Study results suggest that the linear regression model usually offers a good trade-off between model accuracy and interpretability. When the most important goal of the prediction is supposed to be accuracy, the random forest is generally the best choice. The quality of developed models depends mostly on the chosen territorial level and municipal waste fraction. The performance of these models deteriorates significantly for lower territorial levels because of worse data quality and bigger variability. Only the age structure seems to be important across territorial levels and municipal waste fractions. Nevertheless, also other factors are of high significance in explaining the generation of municipal waste fractions at different territorial levels (e.g. number of economic subjects, expenditures, population density and the level of education). Therefore, there is not one single effective public policy dealing with circular economy strategy that fits all territorial levels. Public representatives should focus on policies effective at specific territorial level. However, performance of the models is poor for lower territorial levels (municipality and micro-regions). Thus, results for municipalities and micro-regions are weak and should be treated as such.

1. Introduction

The transition of the current waste management (WM) systems in Europe to the circular economy based on waste flows regulation has significant cost effects on WM system users (especially municipalities, but also households) (Tomić and Schneider, 2020). Therefore, there is a high demand for strategies that minimize socio-economic impacts on the European recycling society and for efficient solutions at the municipal level. Especially, strategies concentrated of waste prevention and minimization have a high scientific and practical attention nowadays.

There are lots of empirical studies based on sophisticated modelling approaches aimed at evaluating effects of various factors on municipal waste (MW) generation, or on separate collection and that enable to predict changes aroused by these factors (e.g. Lu et al., 2009; Dai et al., 2011; Ayvaz-Cavdaroglu et al., 2019). The knowledge of these factors is a necessary pre-requisite for the establishment of WM strategies at

* Corresponding author.
E-mail address: Martin.Rosecky@vutbr.cz (M. Rosecký).

https://doi.org/10.1016/j.jenvman.2021.112584
Received 21 November 2020; Received in revised form 24 March 2021; Accepted 8 April 2021
Available online 27 April 2021

Table 1
Overview of reviewed papers.

Country	Level	Model	MMW	PAP	PLA	GLA	BIO	MET	SEP
Italy (Abrate and Ferraris, 2010)	M	GR	√						√
Spain (Prades et al., 2015)	M	LR		√	√	√	√		
USA (Kontokosta et al., 2018)	B	DT	√	√					√
France (Hatik and Gatina, 2017)	M	DR					√		
Malaysia (Jaafar et al., 2018)	M	DA		√	√	√	√		
Spain (Mateu-Sbert et al., 2013)	R	LR							√
Turkey (Keser et al., 2012)	R	LR	√						
Turkey (Ozkan et al., 2015)	CD	GR		√	√	√		√	
Ireland (Purcell and Magette, 2009)	CD	GIS					√		
Vietnam (Trang et al., 2017)	HH	LR		√	√	√	√		
Nepal (Dangi et al., 2011)	HH	DA		√	√	√	√	√	
Vietnam (Thanh et al., 2010)	HH	LR		√	√	√	√		

M – municipality, B – Building, R – region, CD – city district, HH – household; GR – general regression, LR – linear regression, DR – dimensionality reduction, DA – descriptive analysis, DT – decision tree based, GIS - Geographic information system.

different territorial levels by public bodies. As Lu et al. (2009), and Dai et al. (2011) confirmed mathematical models were developed as a tool for the optimization of municipal waste management (MWM) and planning. However, the explanatory value of these models is highly dependent on the accurate prediction of the waste generation (Jalili and Noori, 2008).

Based on the urban case studies (Beijing, or Istanbul) Dai et al. (2011), and Ayvaz-Cavdaroglu et al. (2019) used mathematical modelling to find optimal MW treatment for minimizing costs and reducing the environmental side effects. Benítez et al. (2008), or Lee et al. (2016) developed models evaluating the correlation between socio-economic factors (e.g. income, or level of education) and MW generation. Both studies provide recommendations for decision-makers how to improve MW treatment using the potential of the suitable infrastructure. Oliveira et al. (2018) developed a model to predict separate collection of packaging waste at the municipal level and concluded that especially inhabitants per bring-bank, relative accessibility to bring-banks, degree of urbanization, number of school years attended, and area affect the participation at the separate collection. Based on household data Alhassan et al. (2020), or Ling et al. (2021) evaluated socio-economic factors influencing the participation of residents in incentive programs of source separation.

As the public policies and WM strategies take place on different levels of the public administration that is equipped with various competences new approaches are needed to define priorities of the MWM policy settings. Our study fills the gap in mentioned modelling approaches by evaluating MW generation at three territorial levels - regional, micro-regional and municipal level. Furthermore, our aim is to identify influential factors that enable to predict MW (and its fractions) generation. The knowledge of influential factors enables the public administration to define a suitable strategy at every territorial level and implicitly, the MW treatment capacities.

1.1. Predictive modelling in MWM

The quantity of MW is commonly predicted using social, economic, demographic and other variables. However, according to the results of existing reviews (Beigl et al., 2008; Kolekar et al., 2008; Cherian and Jacob, 2012), little attention has been given to predictive modelling of MW fractions. This problem is even more serious for higher territorial levels (municipalities, micro-regions, regions or their equivalents). The main reason is that such research usually obtains data via surveys, which is costly and time-consuming. Even in case the same pattern of behavior preserves from lower levels, collection of the same influencing factors is challenging and virtually impossible. Thus, it is not possible to carry out such research for a long time or large territory (e.g. region). Since Beigl et al. (2008) reviewed articles until 2005 and other reviews are not so comprehensive (they are not aimed at MW fractions and reviewed up to 20 papers), other research was done by authors of this paper. Time horizon of 2009–2019 was selected as a period of interest. The selection comprises only papers related to waste generation prediction, thus studies concerned with forecasting were not included. Main findings (see Table 1) could be summarised as follows. First, 50% of papers use territorial units smaller than municipalities. These concern city districts, buildings and households, special units closely connected to modelled waste (hospital, construction site, university, restaurant, etc.) were not included due to their irrelevance for this paper. It is remarkable that none of the papers deal with multiple levels of territorial breakdown. Second, a total of 34 (almost 3 per paper) waste fractions were modelled (marked by ticks in Table 1). About 40% of studies dealt with bio (BIO), paper (PAP), plastic (PLA) and glass (GLA) altogether. Other fractions important for this paper were not so common, which is surprising at least in case of MMW as it is the main contributor to MW generation. In some cases (Abrate and Ferraris, 2010; Prades et al., 2015), only total amount of separated waste (SEP) is subject to research and in case of Kontokosta et al. (2018), special "fraction" (MET + GLA + PLA) is collected. Third, the most common approach to analyse MW generation is (multiple) regression models (58%). No paper dealing with metal (MET) or bulky waste (BW) at the level of municipality (or higher) was found. Table 1 shows 12 papers dealing with at least one of the MW fractions relevant for this paper (MMW, PLA, PAP, GLA, MET, BIO, BW).

Prediction models are usually based on socio-economic, demographic and other factors so they are not suitable for forecasting. Forecast of all inputs will be needed for this purpose. Thus, prediction models can be used for estimation of missing or erroneous data as well as for scenario planning. According to above mentioned results, there is a lack of studies that compare different prediction modelling approaches of MW fractions generation for various levels of territorial breakdowns especially those defined by administrative definitions like regions, districts and cities or their equivalents. These are also essential for strategic planning of sorting policy and material recovery. The results are valuable for government advancements in MWM as well as for investors of WtE plants and sorting facilities.

1.2. Influential factors

The current scientific discussion about factors influencing the waste generation and recycling involves not only socio-demographics (Miafodzyeva and Brandt, 2013; Rybová et al., 2018) and socio-economics (Saphores and Nixon, 2014), but also technical and organizational variables describing the perceived convenience of the MWM system (e.g. distance to collection points, door-to-door collection, or availability of containers in public space), charging system (Miafodzyeva and Brandt, 2013; Slavík et al., 2020), or environmental values and psychological variables (Slavík et al., 2018). Especially the last group of factors gains increasing attention when socio-demographic and socio-economic factors are characterized by ambiguous results with low predictive power (Rybová et al., 2018).

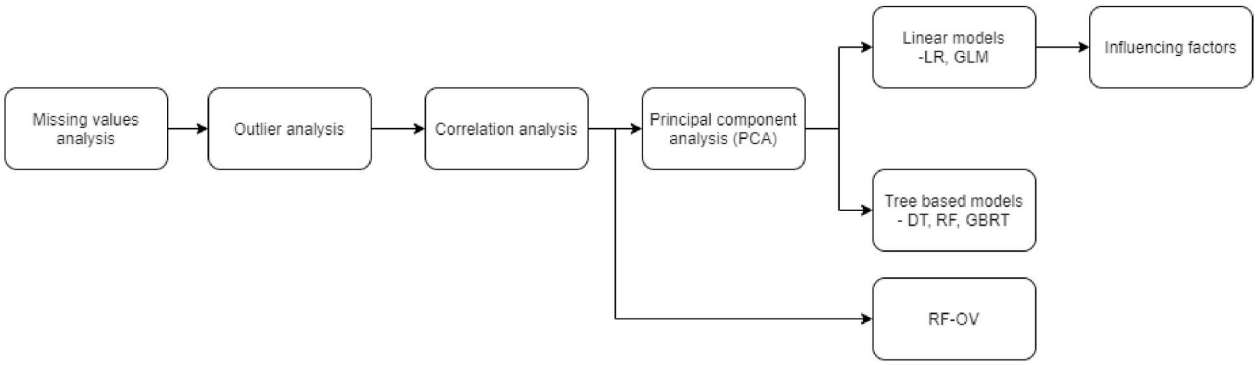

Fig. 1. Paper framework (LR – Linear regression, GLM – Generalized linear model, DT – Decision tree, RF – Random forest, GBRT – Gradient boosted regression tree).

The research of socio-demographic factors includes especially age and gender. Because of the modest lifestyle (Sterner and Bartelings, 1999), higher willingness to participate at separate collection (Saphores et al., 2006; De Feo and Polito, 2015), higher level of environmental knowledge, or education (De Feo and De Gisi, 2010) older people generate less waste and recycle more. However, some authors argue that the relationships between age and the MW generation is very weak, or insignificant (Lebersorger and Beigl, 2011; Miliute-Plepiene et al., 2016).

When analyzing gender Daskalopoulos et al. (1998), or Talalaj and Walery (2015) concluded that because of a higher attention to their appearance, or consumption habits (catalogue shopping, home delivery services) higher share of women in the municipality leads to higher generation of waste. On the other hand, women are also more willing to use the municipal infrastructure for the waste separation and separate waste in the household (Saphores et al., 2006). However, also the relationship between gender and waste generation (or recycling) is ambiguous (Saphores and Nixon, 2014; Miliute-Plepiene et al., 2016).

The research of socio-economic variables influencing the MW generation and recycling involves the average household size (Beigl et al., 2008; Lebersorger and Beigl, 2011), income (Berglund, 2006; Gellynck et al., 2011; Gellynck et al., 2011, 2011), level of education (Sterner and Bartelings, 1999; Hage and Söderholm, 2008; Keser et al., 2012) and other (less obvious) factors (e.g. unemployment rate, share of people employed in agriculture, heating by solid fuels, or population density). Especially heating by solid fuels is important for WtE because it reduces collected waste quantities differently (not all MW fractions are suitable for burning). Solid fuel heating systems have a two-tier impact on waste generation. While Lebersorger and Beigl (2011) concluded that households equipped by solid fuel heating systems generate less waste (because of the by-burning of waste), Dennison et al. (1996) found higher generation of waste in those households induced by the generation of ash.

Keser et al. (2012) concluded that paved roads enable higher collection frequencies and better collection works and therefore, higher amounts of waste can be expected in rural areas. Lebersorger and Beigl (2011) found regions with higher share of people employed in agriculture as regions with lower waste generation. The explanation lies in the higher waste separation and recycling of organic matter. Furthermore, also population density influences waste generation. Johnstone and Labonne (2004) confirmed higher waste generation in regions with higher population density when the crucial infrastructure is missing competing with lack of space.

Beyond socio-demographic and socio-economic variables the MW generation is influenced also by behavioral (or psychological) variables and environmental values (intrinsic variables). Especially waste separation and recycling as alternatives to waste generation stay in the centre of attention. Higher separation rates lead to MMW reduction and change of its structure. Both MMW quantity and structure are crucial for WtE plants.

The last group of factors influencing the waste generation and recycling represent situational variables that reflect the perceived convenience of the MWM system. The better, convenient and user-friendly separation system the higher recycling (and motivation to reduce amount of waste generated). In this context, especially distance to the container (Struk, 2017), its localization (Mattson-Petersen and Berg, 2004; González-Torre and Adenso-Díaz, 2005), or space for waste storage at home (González-Torre and Adenso-Díaz, 2005) are highlighted.

Previous paragraphs suggest that the impacts of socio-economic and demographic variables are often quite weak and thus, reported results are questionable. So possible relationships should be judged not only by 'statistical' significance, but also by 'real world' significance (those are not the same). Moreover, it seems that it is not possible to generalize the results from one country to every other. This is particularly true in case of big differences in population well-being, current state of MWM, urbanization, and other aspects. According to our opinion, it can be misleading to analyse just household or municipality level and generalize these conclusions to upper territorial levels.

Unfortunately, analysis of influencing factors for levels of micro-regions and regions is quite rare in the literature (see Table 1). The only papers found are Keser et al. (2012), Mateu-Sbert et al. (2013) and Mazzanti et al. (2008). However, Keser et al. (2012) and Mazzanti et al. (2008) do not analyse separate collection fractions. Mateu-Sbert et al. (2013) analyses only the impact of tourism.

1.3. Novelty and contribution

This paper aims at exploration and comparison of predictability of main MW fractions for multiple territorial levels. Moreover, influencing factors for each MW fraction and territorial level were examined. The main research questions are:

- Which socio-economic or demographic factors are the most important for micro-regional and regional level?
- Are there any 'super predictors' which are important for multiple MW fractions on the same territorial level or for selected MW fraction across the levels?
- Is it possible to transfer qualitative results from one territorial level to another (e.g. Is it sufficient to analyse factors for municipality level and use them for higher levels or vice versa)?
- How much does the MW fractions predictability vary according to selected territorial level and MW fraction?

The reported analysis provides an example of a framework for predictive modelling and influencing factor identification for MW fractions with problematic MWM records. It was used for case study in the Czech Republic. Moreover it:

- Was done for multiple fractions of MW (MMW, PLA, PAP, GLA, BIO, MET, BW) on different territorial levels, which is not very common.

- Identifies important predictors for each 'MW fraction'-'territorial level' pair.
- Uses multiple modelling approaches.
- Enhances quality of scenario planning for material and energy recovery.

2. Materials and methods

Countries are usually divided by geographical, administrative or historical means into smaller parts like cantons, constituent states, regions, districts, micro-regions etc., which is problematic for intercountry comparison. Data for higher territorial levels are commonly obtained by aggregation from lower levels. Overall problem with data is the existence of various definitions, which are not always clearly defined, and presence of defective input.

At the beginning an analysis of missing values and outlier analysis was performed, followed by correlation analysis, principal component analysis (PCA). By applying PCA, mutually orthogonal principal components (PC) were created. PCs were then used as inputs for various models. Framework of the paper is summarised in Fig. 1. All data manipulation, preparation, modelling and visualisation were done using R programming environment (R Core Team, 2019).

2.1. Data

The Czech Republic is a country in Central Europe with 10.7 million inhabitants and an area of 78,866 km^2. Capital city of the Czech Republic is Prague (1.3 million inhabitants), which is also one of the 14 regions (population of 0.3–1.3 million). The regions consist of 206 micro-regions (8600 to 1.3 million inhabitants) and about 6250 municipalities (up to 1.3 million inhabitants).

Data sets were created for selected territorial units and the time horizon was defined by the availability of data from MWM (2009–2017 for all levels). MWM data from previous years were not included due to methodological changes in the MWM system. The regional-level data set consisted of 126 variables. Datasets containing about 50 (independent) variables were used at micro-regional and municipality level.

Same waste fractions and their definitions were used for all levels, namely MMW, PLA, PAP, GLA, BIO, MET and BW. Finally, MW, defined as the total sum of named fractions, was also included. This definition of MW represents about 93% of total MW in the Czech Republic. All the mentioned waste production variables (dependent variables) were converted to per capita rates. Similar operation was also done for independent variables (e.g. population in the age group of 0–14 was converted to percentage of total population, population to population density etc.). Such operation worsens the correlations between dependent and independent variables. Thus, model performance also gets worse. On the other hand, this helps to reveal and understand relationships in the data. It is well known that for established waste fraction generation rate, population is the most influencing factor. Total population also influences other variables (e.g. population in age groups, gross domestic product, etc.). By using these original data, their impact is masked by the main effect (size of the territorial unit).

2.2. Outlier analysis

Prior to a further data analysis, it is advisable to identify outliers and extreme values. The identification and data modification took place in three steps:

1. Identification and omission of outliers at the lowest level of territorial breakdown (municipal level for data from MWM; for other factors, the smallest territorial unit is given by data availability).
2. Aggregation to higher territorial units, if the data from another source are not available, they are not used for aggregation (i.e. if the data are missing in the municipality, then the value for the given micro-region is calculated without the data from this municipality).
3. Identification and omission of outliers for higher territorial breakdown after the data aggregation.

Boxplots or z-scores are usually used for outlier identification. However, in this case, a more robust measure Q, proposed by Rousseeuw and Hubert (2011), was used, see equation (1) (x is variable of interest). The robust procedure was chosen with regard to the fact that it was necessary to assess about 150 variables (in total), which often lack expertise in the decision on outlying. Moreover, many variables were heavily skewed so traditional tools for outlier detection are not appropriate in this case. The value was marked as outlying (and removed) if its value of Q was greater than 3 (in absolute value). The exact threshold value should not have a major impact (outliers are typically significantly larger).

$$Q = \frac{\left(x_i - \underset{j=1,\ldots,n}{median}\left(x_j\right)\right)}{1.483 \cdot \underset{i=1,\ldots,n}{median}\left|x_i - \underset{j=1,\ldots,n}{median}\left(x_j\right)\right|} \quad (1)$$

The MWM values identified as outliers were removed for further work. Data at higher levels of territorial breakdown were then aggregated based on the data from the lower territorial unit. In case of outliers, neither MWM value nor population of the particular lower territorial level unit was included. Moreover, MW was defined as an aggregate of individual waste fractions. If the value at the municipal level was omitted because of outlying, it was replaced for the purpose of aggregation into MW. If a single point was omitted in the 2009–2017 time series, it was replaced by linear interpolation. If more points in the time series were missing, they are replaced by the average production of the relevant region in the given year.

2.3. Linear models

A linear regression (LR) analysis is a statistical method that is widely used to describe dependencies in data. Based on the literature review, the regression analysis is suitable for the use for MWM data. Generalized linear models (GLM) are not used as frequently in the MWM domain as the other mentioned approaches. However, they are a natural choice when the classical LR is not considered as an adequate tool. Based on the authors' previous experience, a gamma regression model (with an identical link function) was chosen. Main advantages of GLMs (when compared to LR) are greater flexibility, possibility to incorporate domain specific knowledge and deal with (some) nonlinear dependencies. A disadvantage of generalized models is that the global optimum is not guaranteed when estimating the coefficients and, in some cases, it may be necessary to supply appropriate starting values (Dobson, 1990). Special limitation of Gamma regression is that the dependent variable can only have positive values, so it is not possible to model the cases in which the production is zero.

Model selection was done by backward elimination with Akaike Information Criterion (AIC) for both LR and GLM models. Assumptions of Gauss-Markov theorem were tested by Anderson-Darling test (normality), t-test (zero mean) and Durbin-Watson test (independence).

2.4. Decision tree – based models

Although models of artificial neural network (ANN) and support vector machines (SVM) are used quite often to deal with similar issues (Kolekar et al., 2008), this analysis did not use them. The main reason is the doubts about their inadequacy for the chosen task, especially with regard to quantity and quality of data. Therefore, only models of decision tree (DT), random forest (RF) and gradient boosted regression trees (GBRT) were used. Tree-based models (especially RF and GBRT) are

quite popular in general, but they are rarely used in MWM. They are not even mentioned in reviews from the MWM area (Beigl et al., 2008; Kolekar et al., 2008; Cherian and Jacob, 2012). DT may not be of sufficient quality for the actual use, but it may serve as a source of information on the inner working of modelled processes.

DT models are quite popular in machine learning due to their capabilities to deal with complex nonlinear relationships and interpretability (Hastie et al., 2009). They are based on recursive partitioning of the feature space to subspaces. Individual models are then created for these subspaces. In each step of the partitioning process, the point performing the greatest reduction of the residuals is selected as a split point. Tree structure describing important variable relationships is generated by this process.

GBRT model is an extension of DT model which incorporates so-called boosting (Friedman, 2001). GBRT does not try to find the best possible model (as DT model) but creates many weak models (trees). DTs are generated sequentially based on residuals of the previous tree. So, the generating process is basically a gradient algorithm which improves the resulting prediction by adding another tree to minimize loss function in each step. The resulting model is a linear combination of all created trees (more accurate models have higher influence).

RF is another tree-based model (Breiman, 2001). Similarly, to GBRT, RF uses many weak learners and combines their results. Randomness is incorporated during creation of this model in two ways. First, only a randomly selected subset of observations is used for a given tree (so called bagging). Second, for this randomly selected subset, a random subset of features (independent variables) is used to create DT. So different features are used for different trees. Such an approach increases the uniqueness of individual trees across the forest. Bagging meta-algorithm increases stability and precision of the model and decreases probability of overfitting.

2.5. Performance measures

Based on the research, Root mean square error (*RMSE*), Mean absolute error (*MAE*) and Mean absolute percentage error (*MAPE*) were selected as measures for model quality assessment and comparison. General definition of R^2 is not included; it was used just for LR and GBRT models. In other cases (GLM, DT, RF), model specific definitions for given models were used. For GLM, multiple definitions of "pseudo" coefficient of determination are available, deviance residuals version was used, see Dobson (1990) for details. DT algorithm uses relative error and RFs coefficient of determination is defined via mean squared error (MSE). Thus, R^2 values are not fully comparable.

All definitions are taken as in eqs. (2)–(8), where n is number of observations, Y_i actual value and $\widehat{Y}_i$ predicted value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{n}} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{Y}_i\right| \tag{3}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \widehat{Y}_i}{Y_i}\right| \tag{4}$$

$$R^2_{GLM} = 1 - \frac{ResidualDeviance}{NullDeviance} \tag{5}$$

$$R^2_{DT} = 1 - \frac{Y_i - \widehat{Y}_i}{Y_i} \tag{6}$$

$$R^2_{RF} = 1 - \frac{MSE}{Var(Y)} \tag{7}$$

Table 2
Overview of outlier identification in MWM data.

Upper bound [t/capita/year]	MMW	PLA	PAP	GLA	MET	BIO	BW
Municipality	0.447	0.032	0.037	0.032	0.031	0.184	0.062
Micro-region	0.318	0.019	0.030	0.018	0.031	0.080	0.057
Region	0.248	0.017	0.022	0.014	0.023	0.059	0.04
Identified [%]							
Municipality	5.3	2.6	4.9	3.3	22.9	3.9	8.1
Micro-region	2.4	0.4	0.8	1.4	3.2	2.4	0
Region	3.2	0	0	0	0	0	2.4

$$MSE = \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{n} \tag{8}$$

All modelling approaches were compared via cross validation (CV). Dataset was randomly split to training (75% of observations) and testing (25% of observations) datasets five times. For each of these training sets, models were created and performance on testing datasets is reported.

3. Results and discussion

Missing values analysis was carried out for each territorial level. At the regional level, about 8% of data was missing, for micro-regions and municipalities, the number was higher (20% and 26% respectively). Missing data with known cause was replaced by a suitable estimate (where applicable). Proportions of missing values point to the fact that data quality decreases for lower territorial levels. The same phenomena can be expected also for created models.

Waste generation variables (i.e. dependent variables) contained a significant number of missing values. There are cases of municipalities where the generation of MMW is not reported and, at the same time, positive generation of some other MW components is registered. Such a situation was handled as an error. In the case of a missing value in MMW, the MW value was also removed, since the MW is artificially defined as the sum of the respective fractions. MMW forms a large part of MW; therefore, in the absence of information on MMW generation, the value of the MW generation was not used either. For other fractions of waste, it was assumed that the absence of the value was because the collection of the given type of waste is not taking place in the municipality and the missing value was replaced by zero. Legislative changes can also cause missing values as well as other problems.

3.1. Outlier analysis

A process using the above-mentioned robust measure (see equation (1)) was used. In addition, compared to the independent variables, the boundaries were checked based on expertise for waste generation data. Also in this case, the boundary for Q was set to 3.

Table 2 shows the corresponding upper bound values according to which the respective values of different types of waste were assessed. Table 2 also shows the percentages of identified measurements. At municipal level, in case of identification of outliers of some of the fractions, the MW values were also updated. If only one observation was identified in the time series of the respective municipality and fraction, it was (for the purposes of MW updating) replaced by linear inter/extrapolation. At the regional level, only a few observations with MMW and MW were identified as outliers.

At the level of municipalities, approximately 3% of the values for independent variables were identified and removed by this procedure. A total of 2% of the micro-region values were removed. For regions, it was 5% of values. Original dataset contained absurd values of annual MW generation (e.g. over 300 and 20 tonnes per capita for MMW and PLA respectively). However clear boundaries between outliers (i.e.

Table 3
Overview of tests p-values and CV performance measures for LR models – regional level.

	Test p-values			Performance measures			
	Normality	Mean value	Correlation	*RMSE*	*MAE*	*MAPE*	R^2
lr_MW	0.4625	1	0.002	0.01	0.01	3.1	0.74
lr_MMW	0.06101	1	0.072	0.01	0.01	2.59	0.73
lr_PAP	0.2473	1	0.832	0	0	9.83	0.32
lr_PLA	0.01577	1	0.034	0	0	10.64	0.73
lr_GLA	0.4797	1	0.018	0	0	5.33	0.77
lr_MET	0.2868	1	0.168	0	0	34.95	0.68
lr_BIO	0.0006377	1	0.078	0.01	0.01	35.69	0.66
lr_BW	0.3061	1	0.33	0	0	12.05	0.38

errorneous inputs) and extremes (i.e. values distant from majority of data) were not found. Thus, our approach does not make a distinction between extreme values and outliers. This distinction can be quite complicated even if the author possesses expertise in a given field. In case of analysis of data from multiple fields, it is getting even more complicated. One of the possible solutions is to focus just on 'bulk' of data by excluding both extremes and outliers. Such a solution is far from perfect and conclusions should not be generalized to excluded cases. For more proper handling of this problem, distribution properties of analysed variables should be examined in greater detail. Considering waste generation data for each territorial unit separately with respect to specific time development should be also helpful for distinction between outliers and extremes.

3.2. Principal component analysis

In the case of many potential predictors which, in addition, are strongly correlated with one another (i.e. a high risk of multicollinearity in regression models), it is advisable to use the principal component analysis (Hastie et al., 2009).

Prior to PCA application, variables containing many missing values (over 20%) and at the same time not strongly correlated with dependent variables (for the regional level about two thirds of all available variables) were removed. The remaining variables were centralized and standardized before PCA was performed (individual variables are often in significantly different orders, which would strongly affect the results).

At the regional level, 40 PCs (10 in use) were created, 19 PCs (6 in use) for micro-regions and 21 (9 in use) for municipalities. The choice about the number of PC in use for further analysis was made based on the so-called Kaiser rule (Kaiser, 1960).

3.3. Linear regression–based models

This section describes the results of selected models, including their diagnosis. All models covered by this section were created using PCs as independent variables. Moreover, at the municipal level, so-called dummy variables were added for BIO and MET models to indicate whether they were collected in the given municipality.

3.3.1. Multiple linear regression (LR)

For regional data, most models *MAPE* range from 3% to 12%. The results for MET and BIO are significantly worse, but this is not surprising. In some municipalities, the collection of these types of waste is not in place, which may also affect the aggregated data. The R^2 values (at the level of the regions) are in a relatively wide range (from 0.32 for PAP to 0.77 for GLA). It should be noted that a lower value of R^2 does not necessarily mean a bad model; it may be the case that the average value already provides a sufficiently accurate estimate and the deviations from it are not very large. Table 3 shows p-values of performed tests. All models at regional level pass *t*-test, two models (lr_PLA, lr_BIO) violate assumption of normality and three models (lr_MW, lr_PLA, lr_GLA) have not uncorrelated residuals. However, it should be noted that some of the values are close to selected significance level ($\alpha = 0.05$). In such a case, no strict conclusions should be made about the test results.

At the micro-regional level, the accuracy deteriorated dramatically compared to the regional level (most of the *MAPE*s ranged from 10 to 30%); for MET and BW, it was more markedly deteriorating (*MAPE* value over 100%), and for BIO, the absurd value in the order of tens of thousands. This phenomenon is due to the presence of very small productions, where the model makes a small absolute error, but the relative error is very high. Pure zero values were omitted for calculating *MAPE* for BW and BIO; otherwise this calculation would not make sense. Compared to the level of regions, there are already significant problems with fulfilling the assumptions. Strictly speaking, none of the models meet the assumptions; however, virtually, the normality violence or independence of residuals (judged by correlation) in PAP or PLA models were rejected only tightly, and such models could be considered applicable.

At the level of municipalities, another expected deterioration occurs in terms of accuracy for all criteria (with exception of R^2). Out of all modelled fractions, just MW and MMW reached $MAPE < 100\%$. The reason is the same as with BIO at the micro-regional level. Extremely small observations could, of course, be eliminated (as with MMW), but in this case, it would be very difficult to deduce when the value is erroneous and when the low value is because the collection of the given fraction in the area is just beginning. The trend of deteriorating accuracy is not maintained with R^2 for BIO and MET where the improvement at the municipal level is due to the addition of the so-called dummy variable indicating whether waste is collected in a given municipality in a given year. At this level, there is already a severe violence of LR assumptions in all models. Results for micro-regions and municipalities are reported in Appendix A.

3.3.2. Generalized linear model (GLM)

GLM models have comparable results to LR models at regional level. At the level of micro-regions and municipalities, it was necessary to specify the starting value of the calculation for some waste fractions; LR estimates served this purpose. With this setting the resulting GLM models were comparable to LR apart from the municipal level (GLMs are better). This is probably due to the omission of zero values in gamma regression models (selected GLM approach). With the lowest units – municipalities, it was not possible to create a model for BIO (therefore performance measures and test results are not reported) and the quality of models was deteriorated due to the presence of zero production. See Appendix B for performance measures and test results.

3.4. Tree-based models

Unless otherwise stated, all models in this section are based on the same data as the models of LR and GLM. *RMSE* is minimized to find optimal values of model parameters. The same number remains in terms of considered inputs (PC) and observations.

3.4.1. Decision trees (DT)

Commonly optimized parameters of DT are *complexity parameter* (cp – denoted by α in some literature) and *maxdepth* (maximum tree depth).

Table 4
Overview of parameter setup and CV performance measures for RF models – regional level (both PCA and OV versions).

	Mtry		Performance measures - PCA				Performance measures - OV			
	PCA	OV	*RMSE*	*MAE*	*MAPE*	R^2_{RF}	*RMSE*	*MAE*	*MAPE*	R^2_{RF}
rf_MW	7	14	0.01	0.01	2.99	0.75	0.01	0.01	2.79	0.75
rf_MMW	5	15	0.01	0.01	2.51	0.8	0.01	0.01	2.62	0.67
rf_PAP	4	26	0	0	8.74	0.41	0	0	8.27	0.41
rf_PLA	6	8	0	0	9.14	0.86	0	0	8.22	0.82
rf_GLA	6	26	0	0	5.55	0.74	0	0	4.81	0.79
rf_MET	7	8	0	0	46.9	0.54	0	0	36.18	0.67
rf_BIO	7	22	0.01	0.01	37.46	0.75	0.01	0.01	29.7	0.73
rf_BW	7	17	0	0	9.09	0.64	0	0	8.05	0.64

Table 5
Overview of best models (based on MAPE) performance measure values for modelled waste fractions – micro-region and municipality level.

	Model	Micro-region				Model	Municipality			
		RMSE	*MAE*	*MAPE*	R^2		*RMSE*	*MAE*	*MAPE*	R^2
MW	RF-OV	0.03	0.02	7.54	0.50	RF-OV	0.09	0.07	53.92	0.37
MMW	RF-OV	0.02	0.02	8.52	0.56	RF-OV	0.06	0.04	56.84	0.44
PAP	RF-OV	0	0	24.29	0.33	RF-OV	0.01	0.01	307	0.26
PLA	RF-OV	0	0	19.69	0.53	RF-PCA	0.01	0.01	435	0.08
GLA	RF-OV	0	0	13.23	0.44	RF-OV	0.01	0	90.01	0.24
MET	GBRT	0	0.00	126.21	0.09	RF-OV	0.01	0	1369	0.41
BIO	RF-OV	0.01	0.01	25,673	0.44	RF-OV	0.03	0.02	13,400	0.25
BW	GBRT	0.01	0.01	85.66	0.18	RF-OV	0.01	0.01	640	0.28

Their goal is to find a balance between the complexity and the depth of the tree. Both values are first optimized separately due to computational demands. The values found in this way serve as a basis for optimizing both values simultaneously. In both steps of parameters optimization (individual and combined), repeated cross-validation is used with division into 10 subsets and 3 repetitions. The models were subsequently pruned, resulting in the final DT models. In terms of selected performance criteria, DT models appear to be worse (compared to LR and GLM, see Appendix C for regional level results and parameter settings), which was the expected result. However, the results confirm the conclusions of previous models (higher accuracy with increasing levels of territorial breakdowns or the same "problematic" types of waste).

3.4.2. Random forest (RF)

The main advantages of RF include high accuracy and little to no need of parameter tuning (Kalmar and Nilsson, 2016). For this reason, only the *mtry* parameter, which specifies the number of randomly selected variables to be selected at each division, is optimized. The default setting for this parameter is one-third of the number of independent variables available, with values ranging from one to two-thirds of the predictors tested during the search. This parameter was optimized at the level of municipalities due to computational demands, so the values corresponding to the basic algorithm settings were used.

Since the splitting nodes are only selected from a subset and the model can be easily created in parallel, so the RF for original variables (OV) were also created. The performance results and parameter settings for regional level are given in Table 4.

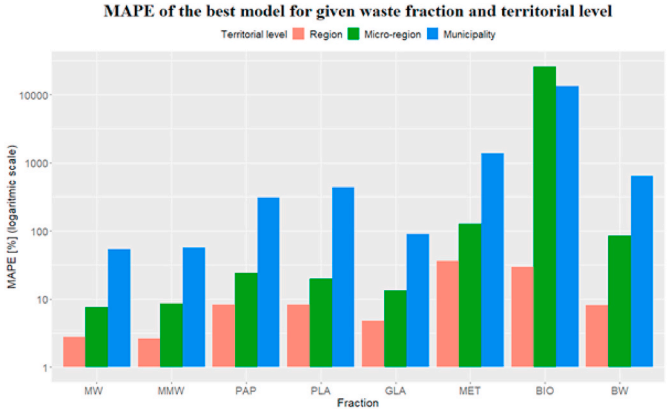

Fig. 2. Comparison of *MAPE* for all modelled fractions and territorial levels (best models).

3.4.3. Gradient boosted regression trees (GBRT)

For GBRT, two parameters were optimized, in particular: *learning rate* (eta - lower values reduce the risk of overfitting but leads to a higher computational time) and *maxdepth* (maximum tree depth). To find optimal parameter values, eta values of 0.001, 0.005, 0.01, 0.05, 0.1 were used, *maxdepth* of values from one to the number of variables (PC) were examined. At the same time, the numbers of iterations (*nrounds*) were recorded, after which the values of the purpose function in the last 10 steps no longer improved. Values of selected performance measures and parameter settings for regional level are presented in Appendix C.

The quality of models varies considerably for different territorial levels. Variations and inaccuracies at municipal level are often smoothed for higher territorial units due to aggregation. At the regional level R^2 values vary considerably for different waste fractions, the same is true for *MAPE*. At micro-regional level, RF shows superiority in terms of *MAPE* (see Tables 3 and 5). The use of original variables for RF models brings a slight improvement of *MAPE* in most cases (about 10% in comparison with LR or GLM). However, micro-regional level models bring unsatisfactory results for some of the fractions (MET, BIO, BW). This pattern continues (and strengthens) at municipality level (see Table 5). The results also confirm that RF are strong in 'decorrelation' of predictors (without using PCA).

The results of LR models on municipality level confirm problems with accuracy of modelling of MW (fractions) in the Czech Republic via socio-economic and demographic variables, see Rybová and Slavík (2016), Rybová et al. (2018). In comparison with Prades et al. (2015) or Lebersorger and Beigl (2011), results from the Czech Republic are much worse. One of the reasons could be data quality. Another reason is that Prades et al. (2015) analysed only cities with population over 5000 and Lebersorger and Beigl (2011) results are for transformed variable $MW_t = \log(MW)$. Moreover, Lebersorger and Beigl (2011) analysed only data

from one year (with census data available) and only one region of the country. So, none of these results are fully comparable with our results. Our analysis is also more comprehensive. Moreover, findings of this paper may be beneficial even for forecasting models like Ghinea et al. (2016), in two ways. First, better predictors can be found and second, accuracy of the models for other fractions or territorial levels can be estimated.

Up to our knowledge, there is no paper dealing with micro-regional and regional level for the Czech Republic data. For MMW production modelling comparison on regional level, Keser et al. (2012) can be used as Turkish 'provinces' are roughly comparable to Czech 'regions' in terms of population. R^2 of Keser et al. (2012) for MMW modelling ranges from 0.4 to 0.62 depending on the model. Our models for MMW on regional level perform better (see Table 3, Table 4 and Appendix) with DT as the only exception. Moreover, our results are based on CV and thus should be more reliable. Fig. 2 shows comparison of *MAPE* for all modelled waste fractions.

3.5. Influential factors

Influencing factors are analysed based on *t-statistics* from regression models. At the municipal level, our results indicate three groups of socio-economic variables with the significant impact on MW generation - age, number of applicants for a job, and amount of expenditures. According to our results elderly people of age 65+ are more prone to separate PLA, GLA, MET, BIO, and BW. Our results are consistent with previous findings (based on individual, household, or municipal level data) that found elderly people more willing to participate at recycling activities (Sidique et al., 2010). Sidique et al. (2010) concluded that older people use drop-off recycling sites more often, especially when the travel distance from home to a site is shorter. Therefore, the localization of drop-off recycling sites matters when encouraging people to separate waste, mainly when elderly people stay in the centre of attention.

Nevertheless, the role of age in explaining the waste (or recycling) behavior seems to be tricky. Our results also found elderly people of age 65+ to generate more MW. Struk and Soukopova (2016) came to the same result for the people aged 50–79. They explained this result by the specific habits of elderly people when they prepare households for being retired - reconstructions, replacement of household goods, or sorting and disposal of goods accumulated during the active life. On the other hand, these results contradict with Sterner and Bartelings (1999) who concluded that based on their frugal lifestyle older people generate less waste. Obviously more detailed work with age groups is needed when explaining behavioral settings of elderly people.

Some authors replace missing data about income at the municipal level by the unemployment rate. Rybová and Slavík (2016) concluded that increasing unemployment leads to a higher MW generation. Our results indicate lower separation of PAP, PLA, and BIO when the number of applicants for a job increases (incl. of those applicants who are in an evidence for more than 12 months). However, the number of applicants for a job seems to be of lower significance when explaining the MW, or MMW generation.

Based on the data on the municipal level our research indicates a significant relationship between MW generation and expenditures of a given municipality. Expenditures are commonly highly correlated with income and thus can be used as a proxy for the economic level of a municipality. Moreover, high expenditures can indicate general development of the municipality. Lebersorger and Beigl (2011) found that per capita tax income of municipality (another possible proxy for economic level) increases MW. According to our results, the same is true also for individual fractions of MW.

Our results for micro-regions and regions are compared mainly to the results of papers at the lower territorial levels, due to lack of relevant papers for higher levels. Socio-economics factors influencing the MW generation and separate collection change at the micro-regional level. The most significant factors are the population density, number of economic subjects with no more than 10, and 50 employees, share of flats in family houses, and age. Our results indicate a positive relationship between the population density and the MW generation, and the separate collection of PAP, PLA, MET, BIO, or BW. These results for MW are in accordance with Mazzanti et al. (2008). However, the negative coefficients for MMW and GLA imply some specific management settings based on density economies - e.g. shared collection containers, or lack of space increases a pressure on waste prevention, minimization, and separate collection.

Our results for the relationship between the number of economic subjects and MW generation are ambiguous. While results for economic subjects with no more than 10, and 50 employees indicate some relationship, results for more than 50 employees are insignificant. Probably, the reason lies in the difference in management of MWM between those economic subjects. While economic subjects with more than 50 employees generate types of industrial waste with lower amounts of waste that is similar to MW and have their own contract to waste company to collect and dispose their waste, economic subjects with 0, or no more than 10, or 50 employees generate waste similar to MW and they are connected on the MWM system (incl. the contract with the municipality). Therefore, the results for economic subjects with no employees indicate higher generation of MMW with an increasing number of these subjects (especially small traders, services etc.) participating in the municipal system and generating waste similar to MW. Results for separate collection fractions are ambiguous. While fractions with their functioning market for secondary raw materials (PAP, MET, or some types of BW) evidence lower generation with increasing number of economic subjects with no employees, the amount of other fractions (PLA, and GLA) increases in the municipal system. The involvement of economics subjects in the MWM system is crucial when planning the financing, operation, and investments in the system. Therefore, the active cooperation between economic subjects and municipalities is considerable.

The share of flats in family houses largely expresses the possibilities of households in the MWM. Barr et al. (2003) confirmed that households living in the family house separate more waste than households in a block of flats. The reason lies in better conditions for the storage of waste and its fractions (e.g. garage). Furthermore, according to Alexander et al. (2008) these households perceive less barriers in the waste separation. Nevertheless, our results seem to be in the contradiction with these conclusions - increasing share of flats in family houses leads to lower amount of PAP (maybe because of the combustion of the PAP in the heating chamber - see Lebersorger and Beigl, 2011), MET (sell for market price), or BW (e.g. combustion of wood components). Higher MMW generation could result from the higher production of ash (Dennison et al., 1996). Another possible reason is that results from lower levels (all references in this paragraph) are simply not preserved for micro-regional level.

Dependence of the MW generation on the age structure at the micro-regional level has analogous results as at the municipal level.

At the regional level especially, following variables are significant: age structure, number of applicants for a job, education, or number of economic subjects. The impact of age structure seems to be the same as for micro-regions and municipalities. Higher share of people aged 15–64 increases MMW generation and decreases separate collection. Age group 65+ and mean age behave in the opposite way. However, the effect on PAP seems to be weak for age structure related variables. In the case of applicants for a job results from the municipal level were confirmed. The increasing number of applicants for a job leads to lower separate collection of PAP, MET, or BIO. Probably, the reason lies in the sale of PAP and MET at the secondary raw materials market. The results for other fractions seem to be insignificant, however, the number of applicants for a job negatively influences the generation of MW (because of lower amounts of PAP, MET, and BIO).

At the regional level the level of education seems to be significant when explaining MW generation. According to our results the increasing

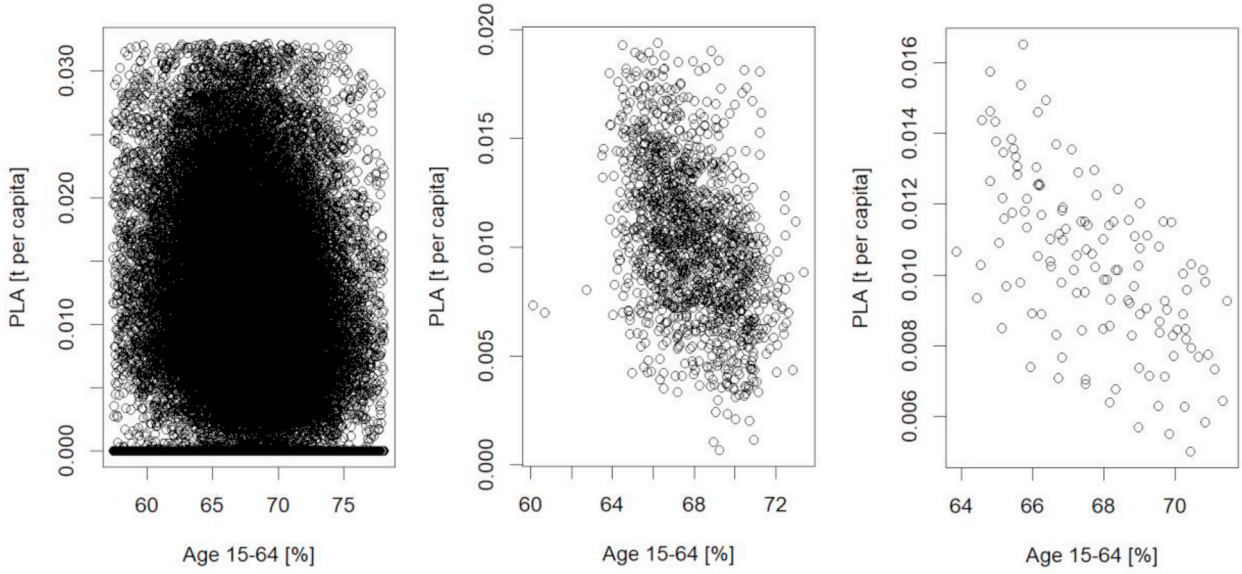

Fig. 3. Example of deterioration of relationship when moving from higher to lower territorial levels. Points graphs for Age 15–64 and PLA for level of municipality (left), micro-region (middle) and region (right).

number of people with secondary education (without graduation) leads to decreasing separate collection and increasing generation of MMW. Because the impact on separate collection is stronger than on MMW generation, MW generation decreases. These results are in accordance with Rybová and Slavík (2016) who found the level of education as a crucial variable. The higher share of higher educated people in the population the higher the separate collection. Our results partially contradict with Keser et al. (2012) who found higher generation of MMW in regions with higher share of higher educated people. However, this effect is significant only for 60% of Turkish regions (Keser et al., 2012). On the other hand, our results in the case of education with graduation show a higher separate collection of PLA, GLA, and BIO.

Considering the number of economic subjects at the regional level our results confirm the results from micro-regional level. The higher number of economic subjects with no employees, the higher generation of MMW and lower separate collection of PAP. Results for other fractions are not significant. Based on the current research of socio-economic variables influencing MW generation especially the average household size seems to be of key importance. Increasing average household size leads to the decrease of average generation of the waste (Lebersorger and Beigl, 2011). Our results indicate the decreasing separate collection of PAP when the average number of household members increases. As Johnstone and Labonne (2004) stated the reason lies in the opportunity to share some types of products (e.g. newspapers, but also books). Therefore, facing the trend of increasing number of single households nowadays, the average amount of (M)MW per person is going to increase. Not only P&E campaigns enhancing environmentally sound behavior (waste prevention and minimization), but also sharing of the waste infrastructure (e.g. containers for MMW) could limit the negative environmental consequences of the changing lifestyle.

Transferability across the levels can be summarised as follows:

- Age structure seems to be important for all analysed levels for most of the fractions. Moreover, the direction of relationships preserves. Higher share of population aged 15–64 increases generation and decreases other fractions as well as MW. Average age and share of population aged 65+ have opposite impact.
- The number of economic subjects for micro-regional and regional level do not behave in exactly the same way, but do not contradict each other.
- Education seems to be important at the regional level since the share of the people with secondary education (without graduation) ranks among the top 5 predictors for all but two (MET, BW) analysed wastes. It could be helpful to use this information also for municipalities and micro-regions, if possible.

Regarding the public policy a significance of mentioned factors at different territorial levels is of key importance for practical decision-making processes. As Dahlén and Lagerkvist (2010) stated influential factors can be divided in factors controlled by local/regional WM authorities, national authorities and factors beyond the possibility of control. Share of the people with secondary education (without graduation) can be considered as 'super predictor' for regions, age structure related variables (especially mean age) and share of flats in family houses for micro-regions and expenditures for municipalities. This conclusion is in accordance with the subsidiarity principle that calls for making decision at the level of public administration most appropriate for finding solutions (more in Buclet, 2002, or Deszczka-Tarnowska and Wąsowicz, 2016). While municipal representatives focus on public expenditures determining the socio-economic aspects and convenience of MWM system for households, and other subject using this system (Soukopová et al., 2016), micro-regional authorities are able to influence the efficiency of the system through economies of scale (e.g. Dijkgraaf and Gradus, 2007), and economies of density (e.g. Abrate et al., 2012), and regional authorities could shape the education of students aimed at increasing the general and instrumental knowledge (Slavík et al., 2018).

Revelation of relationships between influencing factors and waste generation enhances quality of scenario approaches for forecasting. This allows decision-makers to make the system more flexible both from short-term (waste collection, material recovery, operation of WtE plants) and long-term (capacity and WtE plants) point of view. It should be noted that conclusions about influencing factors for municipality and micro-regional level are questionable since the quality of these models is quite poor in most cases (especially for municipality level). Moreover, the results should not be generalized to the whole population, since extreme cases were excluded (not distinguished from outliers). Although (spearman) correlations are safely significant (i.e. p-values are far below significance level of 0.05) for multiple predictors, they are very weak (especially for municipality level) as can be seen in Fig. 3. No strong statements and actions should be made based on such results.

4. Conclusions

Multiple predictive models for various MW fractions using social, economic and demographic data were created for different territorial levels of the Czech Republic. Moreover, assumptions of LR and GLM were checked. The results for particular waste fraction are usually quite similar across used models. Therefore, most of the information provided by the input data was probably used. In such cases, "white-box" models like LR should be preferred. The models at the regional level can be considered useful (with exception of BIO and MET). In contrast, the models for lower levels (micro-regions and municipalities) have almost a negligible predictive value and are thus basically useless for description of data dependencies.

Our research confirmed the significance of socio-economic and demographic variables that highly influence MW generation. Especially age, education level, share of flats in family houses, number of applicants for a job, amount of expenditures, or population density matter in MWM. However, the significance of these variables differs considering municipal, micro-economic, and regional perspective of the analysis. Furthermore, the significance of the number of economic subjects confirms the close relationship between MWM and the market for secondary raw materials that absorbs fractions from the separate collection.

In accordance with the subsidiarity principle our results indicated that there is not one single effective public policy dealing with circular economy strategy that fits all territorial levels. Public representatives should focus on policies effective at specific territorial level. To enable public bodies at different territorial levels to undertake an effective action resulting in circular economy the crucial attention should be paid to various factors influencing the generation of MW fractions at regional, micro-regional and municipal level in the Czech Republic.

Furthermore, problems with missing values, outliers (and extremes) and distinction between 'statistical' and 'real-world' significance were addressed. Future research should involve a more sensible approach to missing values and outliers/extremes in MWM. Another possibility to improve model selection for linear models is to test other ways for model selection (e.g. lasso regression) or use other performance measures (e.g. Bayesian information criterion). At the municipality level creation of more specific models (e.g. according to municipality size) or spatial models (e.g. geographically weighted regression) may be also helpful. Furthermore, the future research should evaluate the environmental effectiveness and efficiency of chosen policy measures on different territorial levels.

Declaration of competing interest

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

Acknowledgements

The article was written as part of the project "Prognosis of waste production and determination of the composition of municipal waste". The authors gratefully acknowledge the financial support provided by TACR (Technology Agency of the Czech Republic) [grant number TIRSMZP719], ERDF within the research project "Strategic Partnership for Environmental Technologies and Energy Production" [grant number CZ.02.1.01/0.0/0.0/16_026/0008413], and SMART City, SMART Region, SMART Community project [grant number CZ.02.1.01/0.0/0.0/17_048/0007435].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2021.112584.

Nomenclature

MW	Municipal waste
MMW	Mixed municipal waste
PAP	Paper waste
PLA	Plastic waste
GLA	Glass waste
MET	Metal waste
BIO	Biodegradable waste
BW	Bulky waste
LR	Linear regression
GLM	Generalized linear model
DT	Decision tree
RF	Random forest
GBRT	Gradient boosted regression trees
GIS	Geographic information system
WM	Waste management
MWM	Municipal waste management
lr_MW	LR model for MW
lr_MMW	LR model for MMW
lr_PAP	LR model for PAP
lr_PLA	LR model for PLA
lr_GLA	LR model for GLA
lr_MET	LR model for MET
lr_BIO	LR model for BIO
lr_BW	LR model for BW
RMSE	Root mean square error
MSE	Mean square error
MAE	Mean absolute error

MAPE	Mean absolute percentage error
R^2	Coefficient of determination
R^2_{GLM}	Coefficient of determination for GLM
R^2_{DT}	Coefficient of determination for DT
R^2_{RF}	Coefficient of determination for RF
Maxdepth	Maximal depth of tree
Minsplit	Lowest number of observations to attempt split
Minbucket	Lowest number of observations in leaf node
complexity parameter	Minimal improvement needed at each node
Mtry	Number of variables available at each node
Learning rate	Step size for parameter tuning
Nrounds	Number of iterations
OV	Original variables
PCA	Principal components analysis
PC	Principal components
CV	Cross-validation
glm_MW	GLM for MW
glm_MMW	GLM for MMW
glm_PAP	GLM for PAP
glm_PLA	GLM for PLA
glm_GLA	GLM for GLA
glm_MET	GLM for MET
glm_BIO	GLM for BIO
glm_BW	GLM for BW
rt_MW	DT model for MW
rt_MMW	DT model for MMW
rt_PLA	DT model for PLA
rt_PAP	DT model for PAP
rt_GLA	DT model for GLA
rt_MET	DT model for MET
rt_BIO	DT model for BIO
rt_BW	DT model for BW
rf_MW	RF model for MW
rf_MMW	RF model for MMW
rf_PLA	RF model for PLA
rf_PAP	RF model for PAP
rf_GLA	RF model for GLA
rf_MET	RF model for MET
rf_BIO	RF model for BIO
rf_BW	RF model for BW
gbrt_MW	GBRT model for MW
gbrt_MMW	GBRT model for MMW
gbrt_PLA	GBRT model for PLA
gbrt_PAP	GBRT model for PAP
gbrt_GLA	GBRT model for GLA
gbrt_MET	GBRT model for MET
gbrt_BIO	GBRT model for BIO
gbrt_BW	GBRT model for BW
GDP	Gross domestic product
WtE	Waste-to-Energy

Credit author statement

Martin Rosecký: Methodology, Formal analysis, Data curation, Writing – original draft, Visualisation. **Radovan Šomplák**: Conceptualization, Supervision, Funding acquisition, Writing – review & editing. **Jan Slavík**: Investigation, Writing – original draft, Writing – review & editing, Funding acquisition. **Jiří Kalina**: Methodology. **Gabriela Bulková**: Conceptualization. **Josef Bednář**: Supervision

References

Abrate, G., Ferraris, M., 2010. The environmental Kuznets curve in the municipal solid waste sector. HERMES Work. Pap. 1.

Abrate, G., Erbetta, F., Fraquelli, G., Vannoni, D., 2012. Size and density economies in refuse collection. Carlo Alberto Notebk. 274.

Alexander, C., Smaje, C., Timlett, R., Williams, I., 2008. Improving social technologies for recycling. Waste Resour. Manag. WR0. https://doi.org/10.1680/warm.2009.162.1.15.

Alhassan, H., Kwakwa, P.A., Owusu-Sekyere, E., 2020. Households' source separation behavior and solid waste disposal options in Ghana's Millennium City. J. Environ. Manag. 259 https://doi.org/10.1016/j.jenvman.2019.110055.

Ayvaz-Cavdaroglu, N., Coban, A., Firtina-Ertis, I., 2019. Municipal solid waste management via mathematic modeling: a case study in Istanbul, Turkey. J. Environ. Manag. 244, 362–369. https://doi.org/10.1016/j.jenvman.2019.05.065.

Barr, S., Ford, N.J., Gilg, A., 2003. Attitudes towards recycling household waste in Exeter, Devon: quantitative and qualitative approaches. Local Environ. 8, 407–421. https://doi.org/10.1080/13549830306667.

Beigl, P., Lebersorger, S., Salhofer, S., 2008. Modelling municipal solid waste generation: a review. Waste Manag. 28, 200–214. https://doi.org/10.1016/j.wasman.2006.12.011.
Benítez, S.O., Lozano-Olvera, G., Morelos, R.A., de Vega, C.A., 2008. Mathematical modeling to predict residential solid waste generation. Waste Manag. 28, 7–13. https://doi.org/10.1016/j.wasman.2008.03.020.
Berglund, C., 2006. The assessment of households' recycling costs: the role of personal motives. Ecol. Econ. 56, 560–569. https://doi.org/10.1016/j.ecolecon.2005.03.005.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.
Buclet, N. (Ed.), 2002. Municipal Waste Management in Europe: European Policy between Harmonization and Subsidiarity. Springer, Dordrecht. https://doi.org/10.1007/978-94-015-9910-8.
Cherian, J., Jacob, J., 2012. Management models of municipal solid waste: a review focusing on socio economic factors. Int. J. Finance Econ. 4, 131–139. https://doi.org/10.5539/ijef.v4n10p131.
Dahlén, L., Lagerkvist, A., 2010. Evaluation of recycling programmes in household waste collection systems. Waste Manag. Res. 28, 577–587. https://doi.org/10.1177/0734242X09341193.
Dai, C., Li, Y.P., Huang, G.H., 2011. A two-stage support-vector-regression optimization model for municipal solid waste management – a case study of Beijing, China. J. Environ. Manag. 92, 3023–3037. https://doi.org/10.1016/j.jenvman.2011.06.038.
Dangi, M.B., Pretz, C.R., Urynowicz, M.A., Gerow, K.G., Reddy, J.M., 2011. Municipal solid waste generation in Kathmandu, Nepal. J. Environ. Manag. 92, 240–249. https://doi.org/10.1016/j.jenvman.2010.09.005.
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. Resour. Conserv. Recycl. 24, 155–166. https://doi.org/10.1016/S0921-3449(98)00032-9.
De Feo, G., De Gisi, S., 2010. Public opinion and awareness towards MSW and separate collection pro-grammes: a sociological procedure for selecting areas and citizens with a low level of knowledge. Waste Manag. 30, 958–976. https://doi.org/10.1016/j.wasman.2010.02.019.
De Feo, G., Polito, A.R., 2015. Using economic benefits for recycling in a separate collection centre man-aged as a „reverse supermarket": a sociological survey. Waste Manag. 38, 12–21. https://doi.org/10.1016/j.wasman.2015.01.029.
Dennison, G.J., Dodd, V.A., Whelan, B., 1996. A socio-economic based survey of household waste characteristics in the city of Dublin, Ireland – II. Waste Quantities. Resour. Conserv. Recycl. 17, 245–257. https://doi.org/10.1016/0921-3449(96)01155-X.
Deszczka-Tarnowska, M., Wąsowicz, M., 2016. The principle of subsidiarity in rational waste management for example of cities with district rights. J. Mod. Sci. 30, 283–300.
Dijkgraaf, E., Gradus, R.H.J.M., 2007. Collusion in the Dutch waste collection market. Local Govern. Stud. 33, 573–588. https://doi.org/10.1080/03003930701417601.
Dobson, A.J., 1990. An Introduction to Generalized Linear Models, first ed. Chapman & Hall/CRC, Boca Raton.
Friedman, J., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189—1232. https://doi.org/10.1214/aos/1013203451.
Gellynck, X., Jacobsen, R., Verhelst, P., 2011. Identifying the key factors in increasing recycling and reducing residual household waste: a case study of the Flemish region of Belgium. J. Environ. Manag. 92, 2683–2690. https://doi.org/10.1016/j.jenvman.2011.06.006.
Ghinea, C., Drăgoi, E.N., Comăniţă, E.D., Gavrilescu, M., Câmpean, T., Curteanu, S., Gavrilescu, M., 2016. Forecasting municipal solid waste generation using prognostic tools and regression analysis. J. Environ. Manag. 182, 80–93. https://doi.org/10.1016/j.jenvman.2016.07.026.
González-Torre, P.L., Adenso-Díaz, B., 2005. Influence of distance on the motivation and frequency of household recycling. Waste Manag. 25, 15–23. https://doi.org/10.1016/j.wasman.2004.08.007.
Hage, O., Söderholm, P., 2008. An econometric analysis of regional differences in households waste collection: the case of plastic packaging waste in Sweden. Waste Manag. 28, 1720–1731. https://doi.org/10.1016/j.wasman.2007.08.022.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7.
Hatik, C., Gatina, J.-C., 2017. Waste production classification and analysis: a PCA-induced methodology. Energy Procedia 136, 488–494. https://doi.org/10.1016/j.egypro.2017.10.308.
Jaafar, I., Azmina Ibrahim, T., Awanis Ahmad, N., Abdul Kadir, A., Razali Md Tomari, M., 2018. Waste generation and characterisation: case study of Seberang Takir, Kuala Nerus, Terengganu, Malaysia. J. Phys. Conf. Ser. 1049 https://doi.org/10.1088/1742-6596/1049/1/012029.
Jalili, G.Z.M., Noori, R., 2008. Prediction of municipal solid waste generation by use of artificial neural network: a case study of Mashhad. Int. J. Environ. Res. 2, 13–22. https://doi.org/10.22059/IJER.2010.170.
Johnstone, N., Labonne, J., 2004. Generation of household solid waste in OECD countries: an empirical analysis using macroeconomic data. Land Econ. 80, 529–538. https://doi.org/10.2307/3655808.
Kaiser, H.F., 1960. The application of electronic computers to factor Analysis. Educ. Psychol. Meas. 20, 141–151. https://doi.org/10.1177/001316446002000116.
Kalmar, M., Nilsson, J., 2016. The Art of Forecasting – an Analysis of Predictive Precision of Machine Learning Models. Dissertation.
Keser, S., Duzgun, S., Aksoy, A., 2012. Application of spatial and non-spatial data analysis in determination of the factors that impact municipal solid waste generation rates in Turkey. Waste Manag. 32, 359–371. https://doi.org/10.1016/j.wasman.2011.10.017.
Kolekar, K.A., Hazra, T., Chakrabarty, S.N., 2008. A review on prediction of municipal solid waste generation models. Procedia Environ. Sci. 35, 238–244. https://doi.org/10.1016/j.proenv.2016.07.087.
Kontokosta, C.E., Hong, B., Johnson, N.E., Starobin, D., 2018. Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities. Comput. Environ. Urban Syst. 70, 151–162. https://doi.org/10.1016/j.compenvurbsys.2018.03.004.
Lebersorger, S., Beigl, P., 2011. Municipal solid waste generation in municipalities: quantifying impacts of household structure, commercial waste and domestic fuel. Waste Manag. 31, 1907–1915. https://doi.org/10.1016/j.wasman.2011.05.016.
Lee, C.K.M., Yeung, C.L., Xiong, Z.R., Chung, S.H., 2016. A mathematical model for municipal solid waste management–A case study in Hong Kong. Waste Manag. 58, 430–441. https://doi.org/10.1016/j.wasman.2016.06.017.
Ling, M., Xu, L., Xiang, L., 2021. Social-contextual influences on public participation in incentive programs of household waste separation. J. Environ. Manag. 281 https://doi.org/10.1016/j.jenvman.2020.111914.
Lu, H.W., Huang, G.H., He, L., Zeng, G.M., 2009. An inexact dynamic optimization model for municipal solid waste management in association with greenhouse gas emission control. J. Environ. Manag. 90, 396–409. https://doi.org/10.1016/j.jenvman.2007.10.011.
Mateu-Sbert, J., Ricci-Cabello, I., Villalonga-Olives, E., Cabeza-Irigoyen, E., 2013. The impact of tourism on municipal solid waste generation: the case of Menorca Island (Spain). Waste Manag. 33, 2589–2593. https://doi.org/10.1016/j.wasman.2013.08.007.
Mattson-Petersen, C.H., Berg, P.E.O., 2004. Use of recycling stations in Borlänge, Sweden - volume weights and attitudes. Waste Manag. 24, 911–918. https://doi.org/10.1016/j.wasman.2004.04.002.
Mazzanti, M., Montini, A., Zoboli, R., 2008. Municipal waste generation and socioeconomic drivers: evidence from comparing northern and southern Italy. J. Environ. Dev. 17, 51–69. https://doi.org/10.1177/1070496507312575.
Miafodzyeva, S., Brandt, N., 2013. Recycling behaviour among householders: Synthesizing determinants via a meta-analysis. Waste Biomass Valoriz. 4, 221–235. https://doi.org/10.1007/s12649-012-9144-4.
Miliute-Plepiene, J., Hage, O., Plepys, A., Reipas, A., 2016. What motivates households recycling behaviour in recycling schemes of different maturity? Lessons from Lithuania and Sweden. Resour. Conserv. Recycl. 113, 40–52. https://doi.org/10.1016/j.resconrec.2016.05.008.
Oliveira, V., Sousa, V., Vaz, J.M., Dias-Ferreira, C., 2018. Model for the separate collection on packaging waste in Portuguese low-performing recycling regions. J. Environ. Manag. 216, 13–24. https://doi.org/10.1016/j.jenvman.2017.04.065.
Ozkan, K., Isik, S., Ozkan, A., Banar, M., 2015. Prediction for the solid waste composition by use of different curve fitting models: a case study. In: International Symposium on Innovations in Intelligent SysTems and Applications, pp. 1–6. https://doi.org/10.1109/INISTA.2015.7276744.
Prades, M., Gallardo, A., Ibàñez, M.V., 2015. Factors determining waste generation in Spanish towns and cities. Environ. Monit. Assess. 187 https://doi.org/10.1007/s10661-014-4098-6.
Purcell, M., Magette, W.L., 2009. Prediction of household and commercial BMW generation according to socio-economic and other factors for the Dublin region. Waste Manag. 29, 1237–1250. https://doi.org/10.1016/j.wasman.2008.10.011.
R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rousseeuw, P.J., Hubert, M., 2011. Robust statistics for outlier detection. WIREs: Data Min. Knowl. Discov. 1, 73–79. https://doi.org/10.1002/widm.2.
Rybová, K., Slavík, J., 2016. Smart cities and ageing population - implications for waste management in the Czech Republic. In: SMART Cities Symposium Prague, pp. 1–6. https://doi.org/10.1109/SCSP.2016.7501025.
Rybová, K., Slavík, J., Burcin, B., Soukopová, J., Kučera, T., Černíková, A., 2018. Socio-demographic determinants of municipal waste generation: case study of the Czech Republic. J. Mater. Cycles Waste Manag. 20, 1884–1891. https://doi.org/10.1007/s10163-018-0734-5.
Saphores, J.D., Nixon, H., Ogundeitan, O.A., Shapiro, A.A., 2006. Household willingness to recycle electronic waste – an application to California. Environ. Behav. 38, 183–208. https://doi.org/10.1177/0013916505279045.
Saphores, J.D., Nixon, H., 2014. How effective are current household recycling policies? Results from a nation al survey of U.S. households. Resour. Conserv. Recycl. 92, 1–10.
Slavík, J., Remr, J., Vejchodská, E., 2018. Relevance of selected measures in transition to a circular economy: the case of the Czech Republic. Detritus 1, 144–154. https://doi.org/10.26403/detritus/2018.12.
Slavík, J., Pavel, J., Arltová, M., 2020. Variable charges and municipal budget balance: communicating vessels of the waste management. J. Environ. Manag. 257, 109976. https://doi.org/10.1016/j.jenvman.2019.109976.
Sidique, S.F., Lupi, F., Joshi, S.V., 2010. The effects of behavior and attitudes on drop-off recycling activities. Resour. Conserv. Recycl. 54, 163–170. https://doi.org/10.1016/j.resconrec.2009.07.012.
Soukopová, J., Ochrana, F., Klimovský, D., Meričková, B.M., 2016. Factors influencing the efficiency and effectiveness of municipal waste management expenditure. Lex Localis – J. Local Self-Gov. 14, 359–378. https://doi.org/10.4335/14.3.358-378.
Sterner, T., Bartelings, H., 1999. Household waste management in a Swedish municipality: determinants of waste disposal, recycling and composting. Environ. Resour. Econ. (Dordr) 13, 473–491. https://doi.org/10.1023/A:1008214417099.

Struk, M., Soukopová, J., 2016. Age structure and municipal waste generation and recycling – new challenge for the circular economy. In: 4th International Conference on Sustainable Solid Waste Management.

Struk, M., 2017. Distance and incentives matter: the separation of recyclable municipal waste. Resour. Conserv. Recycl. 122, 155–162. https://doi.org/10.1016/j.resconrec.2017.01.023.

Talalaj, I.A., Walery, M., 2015. The effect of gender and age structure on municipal waste generation in Poland. Waste Manag. 40, 3–8. https://doi.org/10.1016/j.wasman.2015.03.020.

Thanh, N.P., Matsui, Y., Fujiwara, T., 2010. Household solid waste generation and characteristic in a Mekong Delta city. Vietnam. J. Environ. Manag. 91, 2307–2321. https://doi.org/10.1016/j.jenvman.2010.06.016.

Tomić, T., Schneider, D.R., 2020. Circular economy in waste management - socio-economic effect of changes in waste management system structure. J. Environ. Manag. 267 https://doi.org/10.1016/j.jenvman.2020.110564.

Trang, P.T.T., Dong, H.Q., Toan, D.Q., Hanh, N.T.X., Thu, N.T., 2017. The effects of socio-economic factors on household solid waste generation and composition: a case study in thu dau Mot, Vietnam. Energy Procedia 107, 253–258. https://doi.org/10.1016/j.egypro.2016.12.144.