ENVIRONMENTAL PROBLEMS

Vol. 2, No. 1, 2017

FORECASTING AUTOMOTIVE WASTE GENERATION USING SHORT DATA SETS: CASE STUDY OF LITHUANIA

Aistė Karpušenkaitė[1], Tomas Ruzgas[2], Gintaras Denafas[1]

[1]Department of Environmental Technology, Kaunas University of Technology, 19, Radvilėnų pl., Kaunas, Lithuania, medianaa@gmail.com
[2]the Faculty of Mathematics and Natural Science, Kaunas University of Technology, 50, Studentų g., Kaunas, Lithuania

Received: 19.10.2016

Abstract. There were 1.83 million cars and average passenger car age was 18 years in Lithuania in 2013. Increasing number of cars has an insignificant effect on car age change but it is contrary to automotive waste, both hazardous and non-hazardous, that accumulates during vehicle exploitation and after it ends. The aim of this study was to assess different mathematical modelling methods abilities to forecast non-hazardous and hazardous automotive waste generation. Artificial neural networks, multiple linear regression, partial least squares, support vector machines, nonparametric regression and time series methods were used in this research. Results revealed that nearly perfect theoretical results in both cases can be reached by smoothing splines and other nonparametric regression methods. It is very doubtful that results would be so precise using data outside of currently used data set range and due to this reason further testing using 2014–2015 data is needed.

Key words: automotive waste, hazardous, car, smoothing splines, nonparametric regression.

1. Introduction

In Lithuania average passenger car age was 18 years and average freight vehicle age was 16 years in 2014. Comparing these statistics side by side with other European Union countries Lithuania's situation looks grim. Furthermore, there are plans in near future to introduce pollution taxes based on vehicle age and it is believed that it will cause a great stir among society members and widen the gap between rich and poor. Car number for 1000 residents is growing every year, bigger cities are choking on exhaustion gasses, cities courtyards can't fit all the resident's cars and it directly influences human health. Automotive waste generation is also growing every year and along with it grows the danger that is caused by hazardous components in vehicles which are still not always treated responsibly and end up in the city landfill. The need to know how and which socio-economic factors influence automotive waste generation is growing bigger when Lithuania's government and waste managers want to effectively control and manage total and hazardous waste collection, management, and treatment system. Forecasting would make it easier to prepare necessary capacities to ensure that such public and private sectors plans would reach its maximum goal.

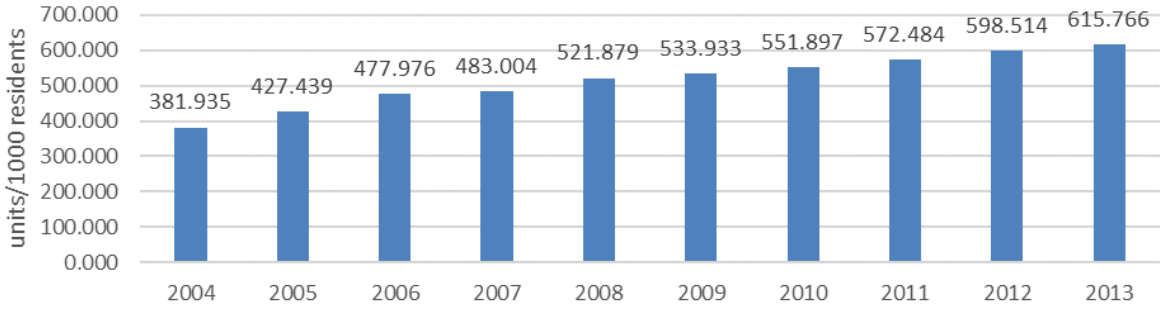

Fig. 1. Number of cars 2004–2013 in Lithuania, units/1000 residents [9]

The overview of previously conducted research on waste matter in general showed that various mathematical forecasting methods were used to predict solid waste generation. Only a few published research papers that could be found using a wide access to scientific journals and subscribed scientific databases uncovered successful application of mathematical prognostic methods applicable for medical waste. This paper will only overview those research papers that showed the most promising forecasting results.

A study conducted by Abdol et al presented the approach to unravel the interpolating problem of various structures of artificial neural networks (ANN) for the long-term prediction of solid waste generation (SWG). Results indicate that the multilayer perception approach has more advantages in comparison with traditional methods, like MLR, in predicting the municipal SWG [2].

Artificial neural networks (ANNs) and multiple linear regression (MLR) were applied to predict the total rate of medical waste generation and in different types of sharp, infectious and general waste. ANNs showed high performance measure values (R^2 = 0.99). Such results were attributed to the non-linear nature of ANNs in problem-solving, which provides the opportunity for relating independent variables to dependent ones non-linearly [13].

In the Rimaitytė et al study, the municipal solid waste (MSW) generation was forecasted using time series analysis. The combination of autoregressive integrated moving average (ARIMA) and seasonal exponential smoothing (SES) techniques was found to be the most accurate. This method proved to be very valuable for forecasting the weekly variation of waste generation data (R^2>0.87) [16].

A support vector machine (SVM) as an intelligence tool, combined with partial least squares (PLS) as a feature selection tool were used to produce a weekly prediction of MSW generated in Tehran, Iran. Research showed that PLS-SVM is superior to the SVM model in predictive ability and is also calculation time saving. In addition, results demonstrate that PLS could successfully identify the complex nonlinearity and correlations among input variables and minimize these [1].

In a study conducted by Noori and others, the hybrid of wavelet transform (WT) – adaptive neuro-fuzzy inference system (WT-ANFIS) and wavelet transform-artificial neural network (WT-ANN) was used to predict the weekly waste generation in Tehran. The achieved results indicate the positive effect of input variables pre-processing by WT, which led to the noticeable increase in the accuracy of two model calculations. However, the WT-ANFIS model had better results than the WT-ANN model, because of the smaller uncertainty than the WT-ANN model [15].

The study of Denafas provides results from municipal waste composition research campaigns conducted during the period 2009–2011 in four cities of Eastern European countries. The quantitative estimation of seasonal variation was performed by fitting the collected data into time series forecasting models, such as non-parametric seasonal exponential smoothing, winters additive and winters multiplicative methods [6].

A different approach to predict medical waste generation was conceived by Eleyan et al and Chaerul et al, who present a new technique using system dynamics modelling to predict generated medical solid waste. Eleyan et al findings indicate that this forecasting approach may cover a variety of possible causative models and track inevitable uncertainties when traditional statistical least-squared regression methods are unable to handle such issues (Eleyan et al, 2013). A hospital waste management model, by Chaerul et al, determines the interaction among factors in the system using a software package. A simulation was made to find out when the existing final medical waste disposal sites will reach their capacity [4].

The range of applicable mathematical modelling methods was limited by the scarce official annual data that was obtainable from the Department of Statistics in Lithuania and the Agency of Environmental Protection, due to the institution's data collection system specifics.

One of the main aims of the research purpose was not to initially create a new total and hazardous automotive waste generation forecasting models, but to investigate how traditional mathematical modelling methods will respond to the data collected for this research and then to use the models with the best performance characteristics in the creation of a hybrid hazardous waste forecasting model. Therefore, no ownership of the applied model development can be credited to the authors of this paper.

2. Materials and methods

2.1. Data sets

Two data sets were developed in the progress of this research. A freely accessible data from Department of Statistics in Lithuania, the Agency of Environmental Protection, government institution "Regitra" and European Commission Directorate – General for Energy and Transport of the 2004–2013 period was used to develop these data sets. Freely accessible official data was used to make developed forecasting methods simple and easy to use in the future for anyone interested from private and public sectors.

The generation of total annual automotive waste is a dependent variable and 5 socio-economic indicators acted as dependent variables in a first data set named TOTAL.

These indicators were picked from a larger primary data set after the calculation of Spearman's rank correlation showed which coefficients could have the most influence on the automotive waste generation. Independent variables in TOTAL data set are a total number of new vehicles, a total number of new cars (M1 type), distance, that passengers travelled by cars and distance, that passengers travelled by buses. Due to the intention to compare the results of this study with foreign countries in the future, this data set was normalized and adjusted to represent 1000 residents. To conclude, TOTAL data set consists of 6 variables and 10 observations.

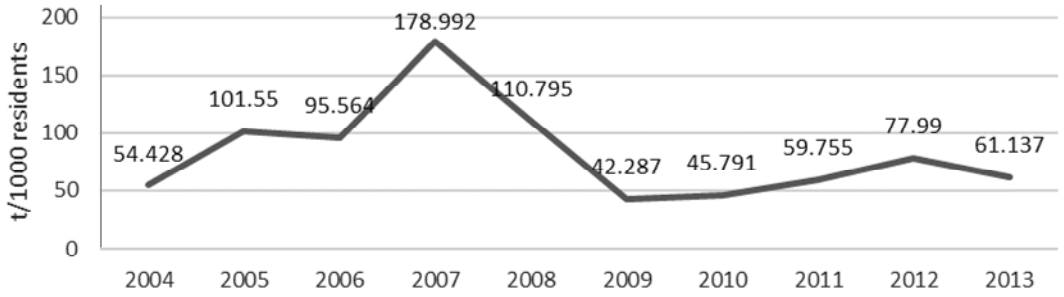

Fig. 1. Data on collected total annual automotive waste generation 2004–2013 in Lithuania, t/1000 residents [3]

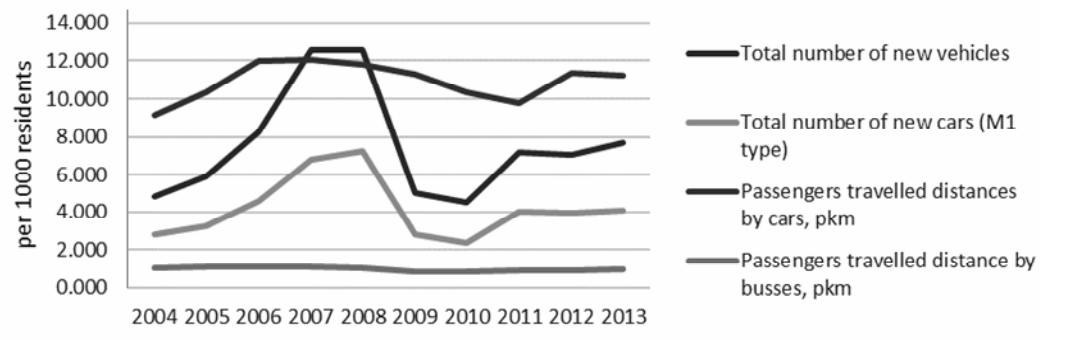

Fig. 2. Data on collected socio-economic indicators 2004–2013 in Lithuania, adjusted to correspond 1000 residents [8, 9]

The second data set, named HAZ, consists of total annual hazardous automotive waste generation as dependent variable and 13 socio-economic indicators. These indicators were picked from a larger primary data set after the calculation of Spearman's rank correlation showed which coefficients could have the most influence for the hazardous automotive waste generation. Independent variables in HAZ data set are: number of registered used cars (M1 type), number of checked out retired vehicles, total number of registered cars (M1 type), number of registered passenger cars (M1-M3 types), number of registered freight vehicles (N1-N3 types), number of registered mopeds (L1-L2 types), number of registered trailers (O1-O4 types), total number of registered vehicles, number of traffic accidents, GDP, international and total haulage by vehicles registered in Lithuania. Due to the intention to compare the results of this study with foreign countries in the future, this data set was normalized and adjusted to correspond 1000 residents. To conclude, HAZ data set consists of 14 variables and 10 observations.

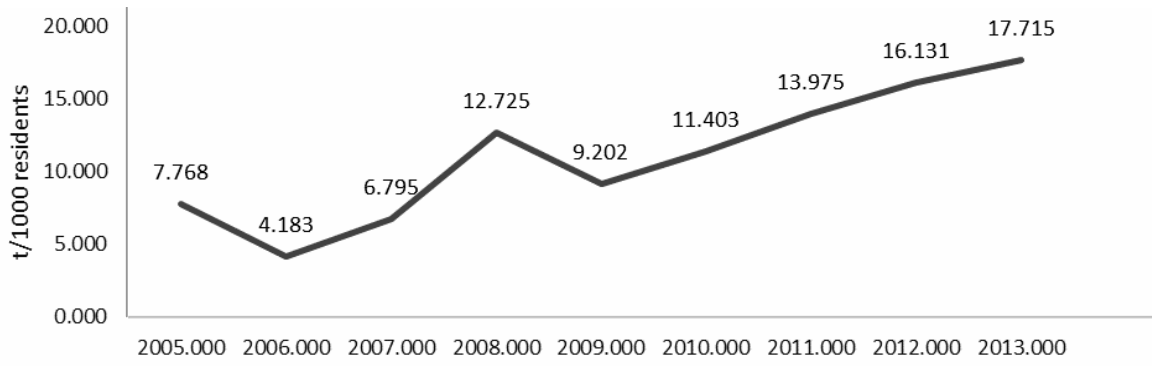

Fig. 3. Data on collected annual hazardous automotive waste generation 2004–2013 in Lithuania, t/1000 residents [3]

2.2. Mathematical modelling methods

2.2.1. Artificial neural networks

The ANN-based models are meant to interact with objects in the real world in the same way that the biological nervous system does. Each neuron in the network is connected to several of its neighbours, with varying coefficients or weights representing the relative influence of the different neuron inputs to other neurons. The weighted sum of the inputs is transferred to the hidden neurons, where it is transformed using an activation function, such as a tangent sigmoid activation function. In turn, the outputs of the hidden neurons act as inputs to the output neuron where they undergo another transformation [19].

In order to apply the neural network method Neural Tools 6. software was used.

2.2.2. Multiple linear regression

Multiple linear regression (MLR) belongs to a family of linear models that are used for mapping a set of independent input parameters, also called regressors, to one dependent output parameters using a set of coefficients. Generally, the regression model parameters are obtained using the least squares method (LSM). The MLR can be presented as follow [11]:

$$Y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_i x_i$$

where Y is the dependent variable (daylight lamps waste and waste that has mercury in its composition generation data), x_i is the independent parameter (variable) and γ_i are coefficients resulting from multiple linear regression.

Multiple linear regression analysis was conducted using SPSS software.

2.2.3. Partial least squares

Partial least squares (PLS) is an effective technique to identify a latent space for two variable spaces by projecting them on a low-dimensional and common space. Since PLS is good at modelling the covariance relations between two sets of variables, it is widely used in machine learning and particularly suitable to the high-dimensional data, where the conventional learning techniques often fail. PLS can also play the role of discriminant analysis, e.g., prediction and classification, like in the case of this research [12].

Minitab software was used for the appliance of PLS method.

2.2.4. Support vector machine

SVMs were developed by Cortes & Vapnik (1995) for binary classification. This method looks for the optimal separating hyperplane between the two classes by maximizing the margin between the classes' closest points—the points lying on the boundaries are called support vectors, and the middle of the margin is the optimal separating hyperplane. Data points on the "wrong" side of the discriminant margin are weighted down to reduce their influence ("soft margin"). When a linear separator cannot be found, data points are projected into a (usually) higher-dimensional space where the data points effectively become linearly separable (this projection is realized via kernel techniques). The whole task can be formulated as a quadratic optimization problem, which can be solved by known techniques.

A program able to perform all these tasks is called a Support Vector Machine and R studio software's package called e1071 was used to make these calculations [20].

2.2.5. Nonparametric regressions

Nonparametric regression relaxes the usual assumption of linearity and enables a more flexible exploration of data, uncovering structure in the data that might otherwise be missed. SAS software was used to apply nonparametric regression methods to the research datasets.

Hastie and Tibshirani [10] proposed a generalized additive models which enable the mean of the dependent variable to depend on an additive predictor through a nonlinear link function. The models permit the response probability distribution to be any member of the exponential family of distributions [17].

The local regression procedure in SAS allows greater flexibility because no assumptions about the parametric form of the regression surface are needed. Furthermore, it is suitable when there are outliers in the data and a robust fitting method is necessary.

Smoothing splines procedure uses the penalized least squares method to fit the data with a flexible model in which the number of effective parameters can be as large as the number of unique design points. This allows greater flexibility in the possible form of the regression surface and this method also makes no assumptions of a parametric form for the model [17].

Kernel regression evaluates the possibilities of the relative random variable. The main task is to find nonlinear relations between random values X and Y. The main purpose of kernel regression is to calculate and use fitted weights. The calculated regression curve is finished with the development of forecasts. In the whole data sample, the kernel function is not defined [17].

2.2.6. Time series

The primary objective of time series analysis is to develop mathematical models that provide plausible descriptions for sample data. To provide a statistical

setting for describing the character of data that seemingly fluctuate in a random fashion over time, it is assumed that a time series can be defined as a collection of random variables indexed according to the order they are obtained in time. For example, a time series maybe be considered as a sequence of random variables, x_1, x_2, x_3, . . . , where the random variable x_1 denotes the value taken by the series at the first time point, the variable x_2 denotes the value for the second time period, x_3 denotes the value for the third time period, and so on. In general, a collection of random variables, {xt}, indexed by t is referred to as a stochastic process. The observed values of a stochastic process are referred to as a realization of the stochastic process [18].

StatTools software was used to calculate moving average (MA), single exponential smoothing (SES) and Holt's methods which belong to time series analysis.

2.3. Performance evaluation

To evaluate the performance of the applied model, four statistical indices were used: the root mean square error (RMSE) (see Eq. 2), coefficient of determination (R^2) (see Eq. 3), the mean absolute error (MAE) (see Eq. 4) and the mean absolute percentage error (MAPE) (see Eq. 5). They are calculated using the outputs given by used models or calculated along with other calculations while running models.

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_0+Y_p\right)^2} \quad (2)$$

$$R^2=1-\frac{\sum_{i=1}^{n}\left(Y_0+Y_p\right)^2}{\sum_{i=1}^{n}\left(Y_0-Y_{ave}\right)^2} \quad (3)$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|Y_0+Y_p\right| \quad (4)$$

$$\mathrm{MAPE}=100\,\%\times\frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_0+Y_p}{Y_0}\right| \quad (5)$$

Where Y_0 is the observed value of medical waste generation, Y_{ave} is the average value of observed medical waste generation and Y_p is the predicted value of medical waste generation. The R^2 represents the proportion of the overall variance explained by the model. Additionally, the MAE represents the most absolute and relative meaningful measure of the model's error, and the RMSE is a measure for the remaining measurement variance not explained by the model. MAPE measures prediction accuracy, usually expressed as a percentage, of a forecasting method in statistics.

3. Results and discussions

Tests on TOTAL data set were made using models described previously and gained results are shown in Table 1. Due to model characteristics and already too short data sets only ANN and SVM models represent true forecasting potential outside of given data range. This is very important notion because regression models (PLS, MLR, nonparametric regression) makes most accurate forecast in the limits of their training data – it is unknown how accurate the predictions would be using data outside of used data sets range. For time series models, the whole chronological order of data is very important and it is also unknown how models would work if there would be missing annual data. Having this in mind, further testing will be needed as soon as more data will be available.

Table 1

Applied models performance results by using TOTAL data set

Methods		RMSE	R^2	MAE
ANN	PN/GRN auto-testing	–	–0.134	–
	MLF auto-testing	1730.572	0.107	57.928
MLR	Enter method	105.879	0.862	11.544
PLS	(after 4 selections)	105.879	0.862	11.544
SVM	Classification	–	–0.75921	–
	Nonlinear Q regression	–	–0.78950	–
Nonparametric regression	Generalized additive	19.554	0.974	4.851
	Local regression	–	–4.064	–
	Smoothing splines	0.0001	1.000	0.009
	Kernel regression	54.170	0.929	6.209
Time series	MA	1554.874	0.999	42.296
	SES	1056.699	0.999	35.748
	Holt's	1056.879	0.999	35.657

Despite these limitations, results gained by the authors in this research represents if and how models would work in such conditions. ANN and SVM models mostly failed in making accurate forecasts and only ANN MLF method showed positive result. Local regression also failed. All other models except time series showed very good or nearly perfect combination of mathematical indices which represents how accurate the predictions would be. Even though all three time series models showed nearly perfect score of coefficient of determination, other indices are much higher and makes authors question the forecast possibilities of time series in this case. If a model does not show good enough results in theory it is likely to fail in practice also.

In conclusion, nearly perfect results are shown by smoothing splines (RMSE=0.0001, R^2=1, MAE=0.009, MAPE=0.008) method. Second best results belong to generalized additives (RMSE=19.554, R^2=0974, MAE=4.851, MAPE=6.52) and kernel regression (RMSE=54.170, R^2=0.929, MAE=6.209, MAPE=0.609) models. These and other models performance results are represented in Fig. 2.

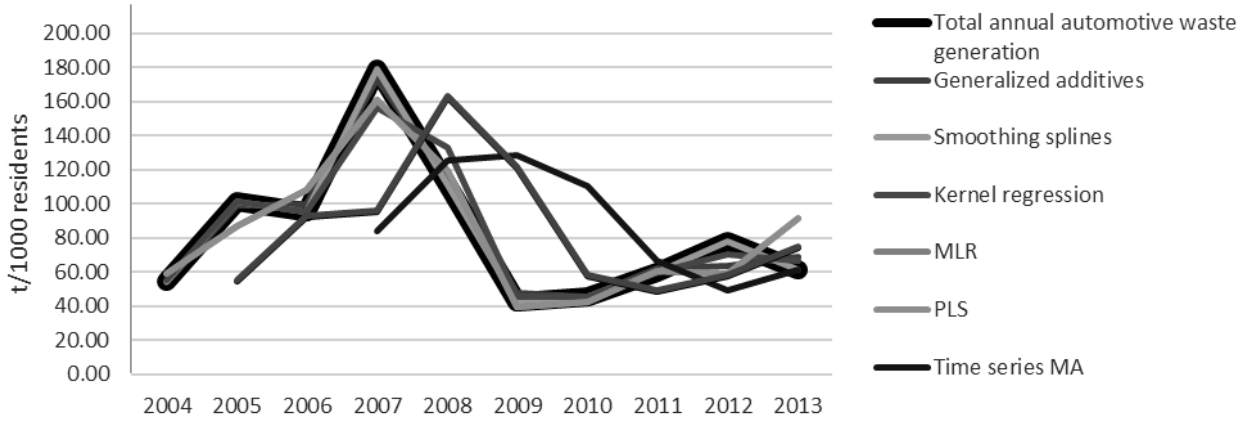

Fig. 2. Observed and predicted values using different mathematical modelling methods on TOTAL dataset.

Table 2

Applied models performance results by using first HAZ data set

Methods		RMSE	R^2	MAE
ANN	PN/GRN auto-testing	–	-5.416	–
	MLF auto-testing	0.977	0.419	1.010
MLR	Enter method	0.283	0.974	0.593
PLS	(after 4 selections)	0	1	0
SVM	Classification	–	-0.749	–
	Nonlinear ϱ regression	–	-0.995	–
Nonparametric regression	Generalized additive	0	1	0
	Local regression	0	1	0
	Smoothing splines	0	1	0
	Kernel regression	0.003	0.999	0.050
Time series	MA	6.755	0.998	3.246
	SES	5.663	0.998	3.055
	Holt‘s	1.834	0.992	1.577

Similarly like in the case of TOTAL data set, all regression models showed very good and higher results (Table 2). PLS, generalized regression, smoothing splines and local regression showed perfect results, but it is very unlikely to happen when real forecasting possibilities would be tested and authors believe, that such perfect results were reached by chance and “convenient” numbers which occurred in the calculation process and were influenced by the length of used data set. MLR high results may be influenced by the same factors. Time series also showed good results but quite high MAPE (18–32 %) causes authors some doubts. Further testing outside of current data set limits is necessary. Currents forecasting results are represented in Fig. 3.

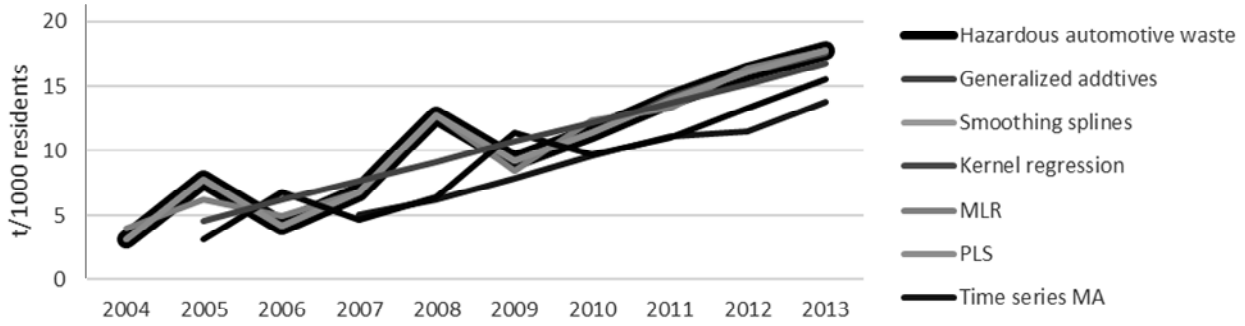

Fig. 3. Observed and predicted values using different mathematical modelling methods on HAZ dataset

Table 3

Mean absolute percentage errors (MAPE) of models tested with the TOTAL and HAZ data sets and which showed positive results

MAPE, %	ANN MLF	MLR	PLS	Generalized additives	Local regression	Smoothing splines	Kernel regression	Time series MA	Time series SES	Time series Holt's
TOTAL data set	62.83	14.43	14.43	6.52	–	0.08	6.09	65.66	46.93	47.02
HAZ data set	15.26	8.99	0	0	0	0	0.37	25.43	32.51	18.55

Results presented in Table 3 supplement those represented in Table 1 and Table 2 and are discussed in this section.

4. Conclusions

The main idea behind this research was that the variables used as independent were given by the Department of Statistics in Lithuania, government institution "Regitra" and European Commission Directorate – General for Energy and Transport, whose data is publicly available for everyone's use. Even if the mentioned institution in Lithuania collects more detailed data about more precise indicators that may influence the generation of automotive waste, this information is only available by special request, therefore indicators that are available to everyone and at any time were used for possible future user's practical reasons.

Four (ANN PN-GRN, ANN MLF, SVM classification, nonlinear regression) out of thirteen chosen for this research mathematical modelling methods currently can demonstrate the true potential of forecasting total and hazardous automotive waste. Unfortunately, only ANN MLF methods succeeded in showing at least positive results.

In the case of total annual automotive waste generation data set, smoothing splines method demonstrated best results. Second best results belong to generalized additives and kernel regression methods.

All regression methods showed perfect or nearly perfect results when used with hazardous automotive data set but authors fear that the length of the data set influenced such results and they may be distorted. Further testing will be needed as soon as 2014–2015 data will be available to test the real forecasting abilities of used regression and time series methods outside of currently used data sets limits.

References

[1] Abbasi, M. Abdul, M. A. Omidvar, B. Baghvand, A. International Journal of Environmental Research, 2013, 7(1), pp. 27–38.

[2] Abdol, M. A. Nezhad, M. F. Sede, R. S. Behboudian, S. Wiley Online Library, 2011, DOI 10.1002/ep.10591,.

[3] Agency of Environmental protection. http://gamta.lt/cms/index? lang=en(accessed June 29, 2016)

[4] Chaerul, M. Tanaka, M. Shekdar, AV. Waste Management, 2008, 28: 442–449.

[5] Cortes, C., Vapnik, V. Support-vector network. Machine Learning, 1995, 20, pp. 1–25.

[6] Denafas, G. Ruzgas, T. Martuzevičius, D. Shmarin, S. Hoffmann, M. and others. Resources, Conservation and Recycling, 2014, 89, pp. 22–33.

[7] Eleyan, D, Al-Khatib, I. A., Garfield, J. Waste Management & Research, 2013, 31(10):986–95.

[8] European Commission Directorate – General for Energy and Transport. Statistical pocket book 2004–2013 http://ec.europa.eu/transport/facts-fundings/statistics/ (accessed June 29, 2016)

[9] Government institution “Regitra”. http://www.regitra.lt/ (accessed June 29, 2016)
[10] Hastie, T. J. and Tibshirani, R. J. Generalized Additive Models. New York: Chapman & Hall, 1990.
[11] Heddam, S. Environmental processes, 2016, 3:525–536.
[12] Huawen, L. Zongjie, M. et al. International Journal of Machine Learning & Cybernetics, 2016.
[13] Jahandideh, S. Jahandideh, S. Asadabadi, E.B. Askarian, M. Movahedi, M.M. Hosseini, S. Jahandided, M. Waste Management, 2009, 29, pp. 2874–2879.
[14] Liutkevičiūtė, V. Neparametrinių regresinių metodų lyginamasis tyrimas. Kaunas, 2014.
[15] Noori, R. Abdoli, M.A. Farokhnia, A. Abbasi, M. Expert Systems with Application, 2009, 36, pp. 9991–9999.
[16] Rimaitytė, I. Ruzgas, T. Denafas, G. Račys, V. Martuzevičius, D. Waste Management & Research, 2012, 30(1), pp. 89–98.
[17] SAS user guide. http://support.sas.com/documentation/ (accessed June 29, 2016)
[18] Shumway, R., Stoffer, D. Time Series Analysis and Its Applications. With R examples. Springer texts in statistics, 2011.
[19] Yetilmezsoy, K. Ozkaya, B. Cakmakci, M. Artificial intelligence-based prediction models for environmental engineering. Neural Network World, 2011.
[20] Meyer, D. Support vector mashines. The interface to libsvm in package e1071. 2014.