Barometr Regionalny
Tom 14 nr 1

Verifying the Performance of Multiple Linear Regression in Predicting the Indicator of Mass Accumulation of Waste. The Case of Lubelskie Voivodship

Tomasz Szul, Krzysztof Nęcka
University of Agriculture in Krakow, Poland

Abstract

In this study, the effectiveness of classical regression models to forecast the indicator of mass accumulation of waste was investigated. The economic and infrastructural variables were used as explanatory variables. The conducted studies show that applying regression models can produce forecasting models generating errors at an acceptable level although only for the municipalities of urban and urban-rural administrative type. For the models where the following were selected as explanatory variables: income indicator, mean number of persons living in a residential building, proportion of arable land in the structure of land use, percentage of buildings in the municipality covered by the waste collection scheme, and the functional type of municipality, the error in the forecast obtained for the test set amounted to 12%–14%. Using the same set of explanatory variables for the rural municipalities caused the models to display forecasting errors for the test set ranging from 35% to 50%. Also, applying another combination of input variables gathered in the course of the studies did not result in developing models of better quality. Therefore, further studies are necessary in the search for more effective methods or other variables describing the mass waste accumulation indicator in rural municipalities.

Keywords: regression model, household, waste, forecasting

Introduction

The provisions of the Act on the Maintenance of Cleanliness and Order in Municipalities which came into law in January 2012 have revolutionised the waste management system in Poland.[1] Under the amendments, municipalities became the owners of waste, and — as a consequence — took over the full control of the management of waste within their respective territories. Waste management requires major financial outlays which in Poland amount to an estimated PLN 650–890 million per year, and constitute 8%–10% of all expenditures for environmental protection (Koneczna and Kulczycka 2011). When a system of waste management is created it has to consider not only economic criteria but also those of social acceptance and environmental effectiveness. The basis for the rational planning of waste management — e.g., taking into account the issues of transportation and storage is the so-called unit waste accumulation indicator whose proper selection is the most important task during the planning stage (Beigl et al. 2005; Kempa 1983). The groups of determinants affecting the quantity of waste generated include economic, social, and infrastructural factors. The distinction of these groups of elements affecting changes of the amount of waste generated is insufficient as the strength of their mutual interactions is not known (Beigl, Lebersorger, and Salhofer 2008; Bogner et al. 1993; Passarini et al. 2011; Sircar, Ewert, and Bohn 2003;

1. See: Obwieszczenie Marszałka Sejmu Rzeczypospolitej Polskiej z dnia 17 lutego 2012 r. w sprawie ogłoszenia jednolitego tekstu ustawy o utrzymaniu czystości i porządku w gminach. DzU z 2012 r. poz. 391.

Szul and Nęcka 2014; Tałałaj 2011). The choice of the method which permits the working out of the model to forecast the amount of waste generated in individual households, which provides the basis for planning waste management in a given area—e.g., the municipality, should consider a number of functions for which significant effects on the final outcome are expected (Malinowski et al. 2009a, 2009b). In practice, however, many of the variables which affect the waste accumulation indicator are very hard to obtain, or their value is burdened with a great level of uncertainty. Great attention has been recently devoted to the changes in quantity and quality of generated waste, depending on the functional type of a given municipality (Bański 2009). It seems that this information can significantly affect the amount of generated waste. In view of the current situation of local governments which are now obliged by law to manage waste in their territories at their own cost, an attempt was made to use classical regression models to forecast the mass waste accumulation indicator based on commonly available data.

1 Study methods

This paper presents a comparative analysis of the effectiveness with which the classical regressive methods can be used to determine the mass waste accumulation indicator. The studies were conducted in 208 municipalities of the Lubelskie Voivodship which were described using the following indicators where explanatory variables:

c_1—population density (persons per km^2)
c_2—mean number of persons living in a residential building (persons per building)
c_3—percentage of buildings in the municipality covered by waste collection scheme
c_4—income indicator (own revenues of municipalities – shares in the taxes constituting the revenues of the state budget, revenue tax from natural persons) (PLN per person per year)
c_5—area of arable land (hectares)
c_6—proportion of arable land in the structure of land use (%)
c_7—functional type of municipality (Bański 2009)

and dependent variables:

d_{a1}—overall mass waste accumulation indicator (kg per person per year)
d_{a2}—mass waste accumulation indicator from households (kg per person per year)

The values pertinent to particular municipalities were obtained from the local data bank and they pertained to the year 2013. In order to verify the admissibility and accuracy of the models developed, the gathered pool was divided randomly into a training set containing 70% of observations, whereas the remaining part constituted a test set. The regression models were developed using Statistica 10.0 software to estimate coefficients using the method of least squares. The selection of the optimum set of exogenous variables was performed on the basis of correlation analysis as well as using the function of forward and backward step-wise regression available in this software. The quality of the developed models was assessed based on the value of MAPE determined for particular sets

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|d_a - d_a^p|}{d_a}\cdot 100,$$

where:
da—real indicator of mass accumulation of waste,
d_a^p—forecast indicator of mass accumulation of waste,
n—number of monitoring commune.

Prior to estimations of the regression model parameters, the outliers were eliminated by using the three-sigma rule.

2 Results

The analyses presented in this paper were done based on statistical data for the Lubelskie Voivodship. This data was obtained from the Local Data Bank and pertained to the year 2013. In the

year under study, a total of 303 thousand tons of waste was generated which constituted 3,7% of the stream of waste in the whole of Poland. In the Lubelskie Voivodship, the indicator expressing the quantity of generated communal waste per single inhabitant was 140 kg per person per year thus it was 34% lower than the national average amounting to 212,9 kg per person per year.[2] The average household in the Lubelskie Voivodship produces 104,4 kg per person per year, whereas in the rural areas this value is lower by approximately 56%.

The characteristic features of the variability among the quantities characterising particular objects for which the explanatory variables were denoted by subsequent symbols c_1 – c_6, whereas the dependent variables were denoted as d_{a1} and d_{a2}.

Tab. 1. The characteristic features of explanatory and dependent variables by administrative type of municipality

Municipality	Measure	c_1	c_2	c_3	c_4	c_5	c_6	d_{a1}	d_{a2}
All	mean	127,7	3,9	61,3	277,8	6,0	85,2	72,0	57,0
	coefficient of variation	218,0	44,5	35,0	47,9	48,2	8,8	71,8	67,1
Urban	mean	883,8	7,6	79,8	549,7	3,0	82,4	182,2	135,4
	coefficient of variation	58,7	47,1	19,4	23,5	56,1	11,6	32,5	33,5
Urban-rural	mean	94,2	4,4	65,9	341,0	4,3	82,2	105,5	79,1
	coefficient of variation	66,2	44,5	24,6	36,8	32,0	11,7	46,5	44,0
Rural	mean	51,4	3,4	58,8	241,0	6,5	85,9	56,1	45,9
	coefficient of variation	47,0	12,8	36,7	38,1	43,9	7,9	55,2	53,7

The analysis performed indicates that the values of both particular conditional and dependent variables in the studied municipalities are characterised by great variability in the order of several dozen per cent. The population density and the proportion of arable land in the land-use structure constituted exceptions. The former displayed extreme variability exceeding 200% whereas the latter did not reach 10%. In order to reduce the variability, an attempt was made to divide the pool of municipalities into particular administrative types (i.e., urban municipalities, urban-rural municipalities, and rural municipalities). This division allowed the determination of average values of the analysed indicators for particular administrative types of municipalities which statistically significantly differed from one another. The greatest differences were noticeable in the population densities which in urban municipalities were 884 persons per square kilometre whereas in the rural municipalities it amounted to as little as 51 persons per km^2. Very large differences also occurred in the average value of the overall mass waste accumulation indicator and mass waste accumulation indicator from households. The quality of waste generated, both overall and from households in urban areas, is approximately three times lower than that in rural municipalities. It was observed during the studies that the variability of parameters for the municipalities of a given administrative type decreased, but still remained high in the order of several dozen per cent.

Prior to developing the classical regression models allowing the determination of mass waste accumulation indicators, the coefficient of correlation and statistical significance of correlation between explanatory and dependent variables were measured. In table 2, the values of Pearson's linear correlation coefficients for dependent and explanatory variables were compiled. The analysis performed shows that for urban municipalities the strongest correlation exists between the mass waste accumulation indicator and the conditional attributes c_2 (mean number of persons living in a residential house) and c_4 (the income indicator). When jointly analysing the municipalities of urban and urban-rural administrative types, the increase in the correlation strength was observed in the majority of conditional attributes, with the highest value showed by the income indicator. On the one hand, the correlation with the area of arable land also turned out to be statistically significant and its direction was negative. This fact corroborated the earlier observations showing higher quantities of waste produced in urban areas. On the other hand, in rural municipalities only

2. [In the journal European practice of number notation is followed—for example, 36 333,33 (European style) = 36 333.33 (Canadian style) = 36,333.33 (US and British style).—Ed.]

Tab. 2. The Pearson's linear correlation coefficients between explanatory and dependent variables by administrative type of municipality

Municipality		c_1	c_2	c_3	c_4	c_5	c_6	c_7
Urban	d_{a1}	0,66*	0,72*	0,69*	0,71*	−0,22	−0,15	−0,67*
	d_{a2}	0,55*	0,72*	0,55*	0,66*	−0,20	−0,07	−0,55*
Urban-rural	d_{a1}	0,74*	0,77*	0,54*	0,84*	−0,48*	−0,01	−0,68*
	d_{a2}	0,70*	0,80*	0,51*	0,83*	−0,44*	0,03	−0,72*
Rural	d_{a1}	0,26*	0,04	0,32*	0,41*	−0,13	0,03	−0,07
	d_{a2}	0,24*	0,02	0,33*	0,33*	−0,10	0,03	−0,02

* $p < 0{,}05$

three among the studied conditional attributes were correlated statistically significantly with the mass waste accumulation indicator. The strengths of these correlations were, however, much lower than that for municipalities of urban and rural-urban administrative types.

Statistica software was used for the estimation of regression models. In the first step, individual exogenous variables for which the regression coefficients were determined had been entered one by one. When the formal assumptions pertaining to regression had been met, the assessment of admissibility and accuracy of the developed model began. In a subsequent step, the optimum combination of conditional attributes was used to explain the changes of the decision-making attribute to the highest degree. Several variables were gathered during the studies which are statistically significantly correlated with the mass waste accumulation indicator, and therefore the function of forward and backward step-wise regression available in the software was used. The characteristics of the developed models of the overall mass waste accumulation indicator is presented in table 3, whereas that of the mass waste accumulation indicator for household only — in table 4.

Tab. 3. The results of estimation of the model of the overall mass waste accumulation indicator (d_{a1}) for urban and urban-rural municipalities

Model number	Explanatory variables	Parameter values	p	R^2	MAPE for: Training set	MAPE for: Testing set
(1)	c_4	0,009	< 0,001	0,72	18,9	18,1
(2)	c_2	0,242	< 0,001	0,79	17,4	13,0
	c_4	6,964	0,003			
(3)	c_2	0,259	< 0,001	0,80	15,6	13,9
	c_4	6,682	0,003			
	c_6	0,990	0,050			
(4)	c_2	0,237	< 0,001	0,82	16,3	12,7
	c_3	6,104	0,006			
	c_4	1,002	0,064			
	c_6	0,542	0,110			
(5)	c_2	0,284	< 0,001	0,83	17,3	11,8
	c_3	6,474	0,004			
	c_4	0,553	0,103			
	c_6	1,188	0,036			
	c_7	5,591	0,208			

The developed forecasting models for mass index waste accumulation, characterized by the smallest mistakes to urban and semi-urban areas represent the relationship 1–5 (rate of mass accumulation of waste in total) and 6–9 (rate of mass accumulation of household waste).

The model cannot be regarded as suitable for forecasting based solely on the information that it has satisfactory quality. The capability of forecasting among the developed models was tested by comparing the actual value of the waste accumulation indicator with the values calculated on the basis of developed models, both visually on graphs (fig. 2) and on the basis of the mean absolute percentage errors (MAPE) of ex-post forecasts which were determined separately for the training set and test set (tab. 3 and 4).

Tab. 4. The results of estimation of the model of the mass waste accumulation indicator in households (d_{a2}) for urban and urban-rural municipalities

Model number	Explanatory variables	Parameter values	p	R^2	MAPE for: Training set	MAPE for: Testing set
(6)	c_4	0,244	$< 0{,}001$	0,70	17,2	21,9
(7)	c_2	0,155	$< 0{,}001$	0,80	16,0	13,9
	c_4	6,479	$< 0{,}001$			
(8)	c_2	0,169	$< 0{,}001$	0,83	12,8	15,3
	c_4	6,230	$< 0{,}001$			
	c_6	0,871	0,024			
(9)	c_2	0,158	$< 0{,}001$	0,84	12,7	13,3
	c_3	5,921	$< 0{,}001$			
	c_4	0,878	0,022			
	c_6	0,290	0,208			

(1) $$d_{a1}^{p} = -4{,}798 + 0{,}009c_4$$

(2) $$d_{a1}^{p} = -3{,}865 + 0{,}242c_2 + 6{,}964c_4$$

(3) $$d_{a1}^{p} = -90{,}845 + 0{,}259c_2 + 6{,}682c_4 + 0{,}990c_6$$

(4) $$d_{a1}^{p} = -118{,}211 + 0{,}237c_2 + 6{,}104c_3 + 1{,}002c_4 + 0{,}543c_6$$

(5) $$d_{a1}^{p} = -173{,}710 + 0{,}284c_2 + 6{,}474c_3 + 0{,}553c_4 + 1{,}188c_6 + 5{,}591c_7$$

(6) $$d_{a2}^{p} = 0{,}249 + 0{,}244c_4$$

(7) $$d_{a2}^{p} = 1{,}116 + 0{,}155c_2 + 6{,}479c_4$$

(8) $$d_{a2}^{p} = -75{,}452 + 0{,}169c_2 + 6{,}230c_4 + 0{,}871c_6$$

(9) $$d_{a2}^{p} = -90{,}083 + 0{,}158c_2 + 5{,}922c_3 + 0{,}878c_4 + 0{,}290c_6$$

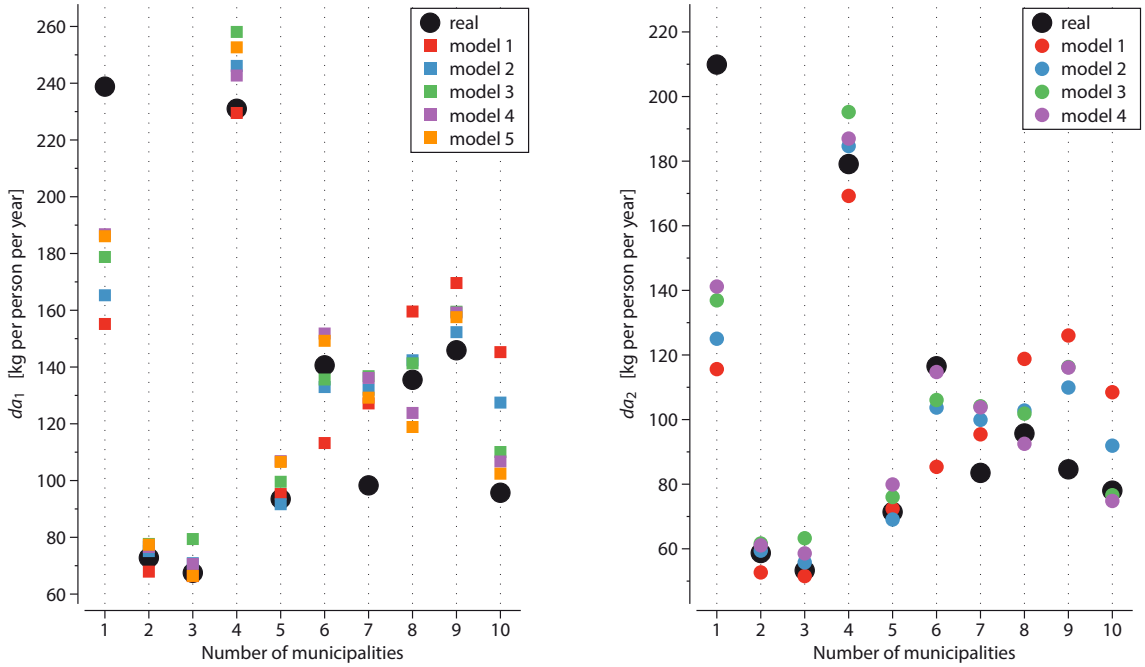

Fig. 1. The actual and forecast values of the mass waste accumulation indicators for urban and urban-rural municipalities: total overall (on the left), overall for households and its ex-post forecasts for the test set (on the right)

The presented characteristic shows better suitability for forecasting the mass waste accumulation indicator in the urban and urban-rural municipalities than the models with greater numbers of input variables, despite the fact that the estimated statistical parameters differed from zero only at $p < 0{,}2$ significance level.

The lowest ex-post forecast error determined in the test set was characteristic for the model of the overall mass waste accumulation indicator, described by: income indicator, mean number of persons living in a residential building, proportion of arable land in the structure of land use, percentage of buildings in the municipality covered by the waste collection scheme, and the functional type of municipality. Although the model developed on the basis of the first three of the aforementioned variables was characterised by the lowest forecasting error for the training set and the significance of the estimated parameters at $p < 0{,}05$ level, the introduction of two more subsequent variables resulted in the increased fit of the model up to 83%, and reducing MAPE for the test set to 11,8%.

The most effective model of the mass waste accumulation indicator for households was determined on the basis of: income indicator, mean number of persons living in a residential building, proportion of arable land in the structure of land use, and the percentage of buildings in the municipality covered by waste collection scheme. The model was characterised by a forecasting error for the training set and test set at respective values of 12,7% and 13,3%. It was observed that for these models, the quality of forecasts produced increased with the increases in numbers of input variables.

The attempt to develop effective forecasting models for rural municipalities was made in an analogous manner as for urban and urban-rural municipalities. The characteristic features of the developed models are presented in tables 5 and 6, and in figure 4.

Tab. 5. The results of estimation of the model of the overall mass waste accumulation indicator (d_{a1}) for rural municipalities

Model number	Explanatory variables	Parameter values	p	R^2	MAPE for: Training set	MAPE for: Testing set
(10)	c_4	0,211	< 0,001	0,23	44,6	35,8
(11)	c_3	0,485	< 0,001	0,32	40,4	38,7
	c_4	0,180	< 0,001			
(12)	c_2	−11,941	0,054	0,33	38,4	40,8
	c_3	0,539	< 0,001			
	c_4	0,184	< 0,001			

Tab. 6. The results of estimation of the model of the mass waste accumulation indicator in households (d_{a2}) for rural municipalities

Model number	Explanatory variables	Parameter values	p	R^2	MAPE for: Training set	MAPE for: Testing set
(13)	c_4	0,122	< 0,001	0,13	52,6	40,5
(14)	c_2	−111,619	0,050	0,29	42,5	47,8
	c_3	0,503	< 0,001			
	c_4	0,097	0,001			

The most effective predictive models of mass accumulation rate of waste for rural communities represent the relationship 10–12 (rate of mass accumulation of waste in total) and 13–14 (rate of mass accumulation of household waste).

(10) $$d_{a1}^{p} = 5{,}998 + 0{,}211c_4$$

(11) $$d_{a1}^{p} = -14{,}972 + 0{,}485c_3 + 0{,}180c_4$$

(12) $$d_{a1}^{p} = 21{,}275 - 11{,}941c_2 + 0{,}539c_3 + 0{,}184c_4$$

$$d_{a2}^{p} = 16{,}384 + 0{,}122c_4 \tag{13}$$

$$d_{a2}^{p} = 32{,}208 - 11{,}619c_2 + 0{,}503c_3 + 0{,}097c_4 \tag{14}$$

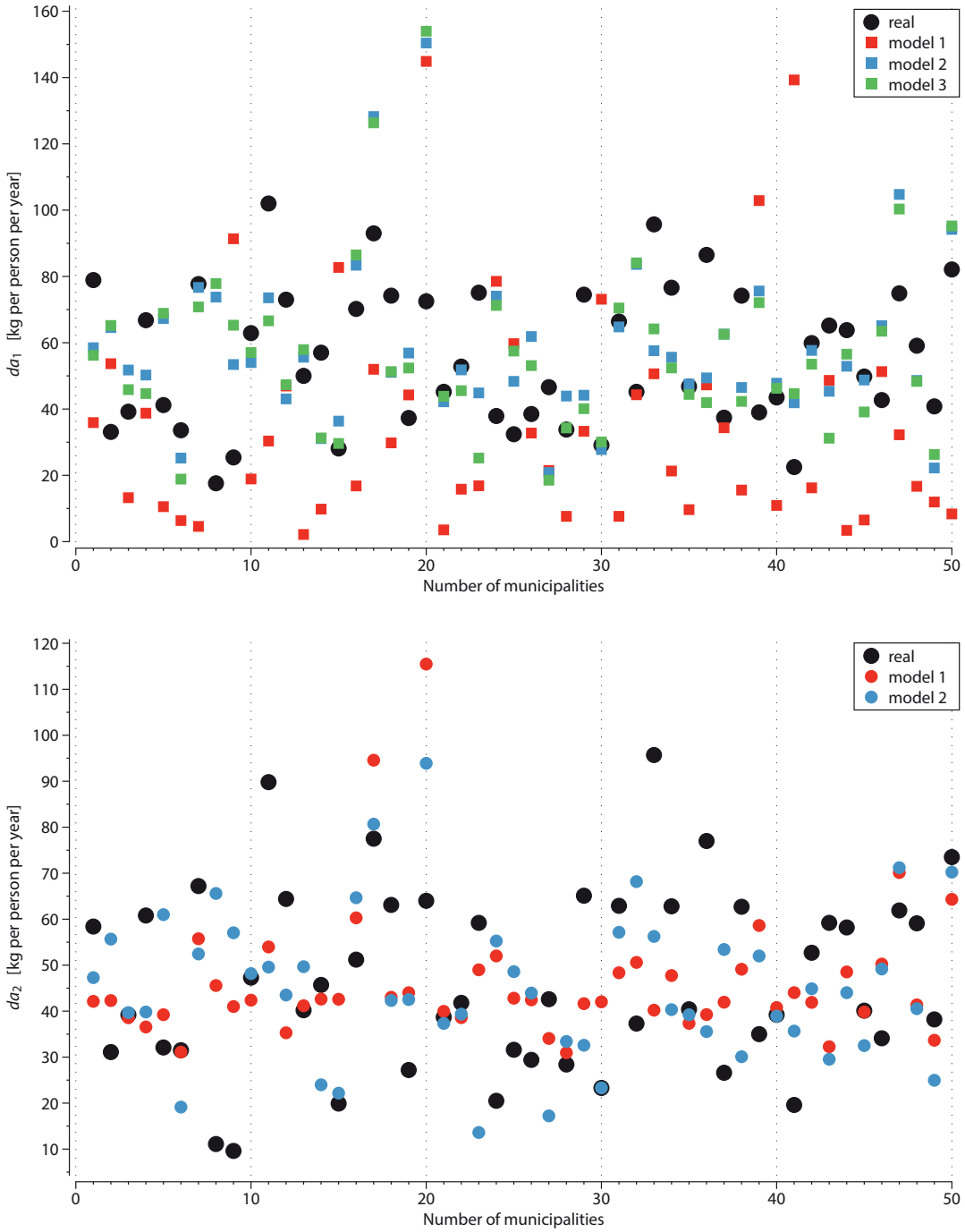

Fig. 2. The actual and forecast values of the mass waste accumulation indicators for rural municipalities: overall (above), overall for households and its ex-post forecasts for the test set (below)

The performed studies show that the effective forecasting model of the mass waste accumulation indicator for rural municipalities cannot be developed on the basis of gathered data describing rural municipalities by the use of classical regression analysis. The model developed on the basis of gathered information was characterised by the low level of fit to actual data, reaching approximately 30% and MAPEs in the order of 40%–50% for both the training set as well as the test set.

Conclusions

The analysis performed in this study indicates that among the municipalities of urban and urban-rural administrative types, the strongest correlations of the mass waste accumulation indicator exist with the mean number of persons living in a residential building, and with the income indicator. The forecasting models of the highest quality, showing errors of 11,8 and 13,3% for the test set, were developed on the basis of the aforementioned variables supplemented by information on the proportion of arable land in the structure of land use, percentage of buildings in the municipality covered by the waste collection scheme, and the functional type of municipality.

The use of the same set for modelling the mass waste accumulation indicator within the territories of rural municipalities permitted the developing of models characterised by as little as 30% of the explanation of modelled changes, and the forecasting error for the test set in the order of 35%-50%. Thus, these are the models which cannot be used in practical applications. It is therefore necessary to continue research looking for more effective methods or the variables which better describe the mass waste accumulation indicators in the areas of rural municipalities.

References

Bański, J. 2009. „Typy obszarów funkcjonalnych w Polsce." In. www.igipz.pan.pl: IGiPZ PAN. http://www.igipz.pan.pl/en/zpz/zbtow/archiwum/1A.pdf (accessed 2015.04.13, not available at present).

Beigl, P., S. Lebersorger, and S. Salhofer. 2008. „Modelling Municipal Solid Waste Generation: A Review." *Waste Management* no. 28 (1):200–214. doi: 10.1016/j.wasman.2006.12.011.

Beigl, P., S. Salhofer, G. Wassermann, I. Maćków, M. Sebastian, and R. Szpadt. 2005. Prognozowanie zmian ilości i składu odpadów komunalnych. Paper read at VI Międzynarodowe Forum Gospodarki Odpadami „Efektywność gospodarowania odpadami", 29.05–01.06.2005, at Poznań–Licheń Stary, Polska.

Bogner, J.E., W.L. Rathje, M. Tani, and O. Minko. 1993. Discards as Measures of Urban Metabolism: the Value of Rubbish. Paper read at The International Symposium of Urban Metabolism, October 1993, at Kobe, Japan.

Kempa, E.S. 1983. *Gospodarka odpadami miejskimi*. Warszawa: "Arkady."

Koneczna, R., and J. Kulczycka. 2011. „Analiza ekonomiczna systemów gospodarki odpadami." In *Ocena systemu gospodarki odpadami. Cz. 2, Praktyczne zastosowania*, edited by A. Kraszewski and E. Pietrzyk-Sokulska, 120–173. Kraków: Wydawnictwo IGSMiE PAN.

Malinowski, M., A. Krakowiak-Bal, J. Sikora, and A. Woźniak. 2009a. „Ilości generowanych odpadów komunalnych w aspekcie typów gospodarczych gmin województwa małopolskiego." *Infrastruktura i Ekologia Terenów Wiejskich* (9):181–190.

————. 2009b. „Wykorzystanie analizy przestrzennej GIS do wyznaczenia wskaźników nagromadzenia odpadów w zależności od liczby mieszkańców i gęstości zaludnienia." *Infrastruktura i Ekologia Terenów Wiejskich* (9):193–205.

Passarini, F., I. Vassura, F. Monti, L. Morselli, and B. Villani. 2011. „Indicators of Waste Management Efficiency Related to Different Territorial Conditions." *Waste Management* no. 31 (4):785–792. doi: 10.1016/j.wasman.2010.11.021.

Sircar, R., F. Ewert, and U. Bohn. 2003. „Ganzheitliche Prognose von Siedlungsabfällen [Holistic Prognosis of Municipal Wastes]." *Müll und Abfall* (1):7–11.

Szul, T., and K. Nęcka. 2014. „Comparison of the Usefulness of Cluster Analysis and Rough Set Theory in Estimating the Rate of Mass Accumulation of Waste in Rural Areas." *TEKA. Commission of Motorization and Power Industry in Agriculture* no. 14 (4):175–180.

Tałałaj, I.A. 2011. „Wpływ wybranych czynników społeczno-ekonomicznych na zmiany ilości odpadów komunalnych w województwie podlaskim." *Inżynieria Ekologiczna* (25):146–156.