TEKA. COMMISSION OF MOTORIZATION AND ENERGETICS IN AGRICULTURE – 2014, Vol. 14, No. 4, 175–180

Comparison of the Usefulness of Cluster Analysis and Rough Set Theory in Estimating the Rate of Mass Accumulation of Waste in Rural Areas

Tomasz Szul, Krzysztof Nęcka

Department of Power Engineering and Agricultural Processes Automation, Agricultural University of Cracow
Balicka Str. 116B, 30-149 Kraków, Poland, e-mail: tomasz.szul@ur.krakow.pl

Received December 01.2014; accepted December 17.2014

Summary. The study shows the comparison between *k-means* and *EM* methods of clustering and the rough set theory as far as determining the rate of mass accumulation of waste in rural areas is concerned. Performed comparative analyses reveal that the average mean absolute percentage error – MAPE for *k-means* and *EM* algorithm ranged between 33 and 41% for the training set and between 20% and 40% for the test set. The rough set theory was characterised by a much better quality of prognosis, for which MAPE value established for the test set was 14%.
Key words: cluster analysis, households, waste, rough set theory.

INTRODUCTION

Amendments to the Act on Maintaining Cleanliness and Order in the Municipalities of 13 September 1996 entered into force in January 2012 and were again amended in January 2013 [consolidated text: Journal of Laws of 2012, item 391]. These changes revolutionised waste management system. According to the amendments, the municipalities became owners of waste and as such took control over waste management in their areas. Waste management requires considerable financial resources, which are estimated at PLN 650–890 million per year in Poland, which constitutes 8–10% of all the environmental expenditures [Konieczna, Kulczycka 2011]. Apart from economic criteria, the creation of waste management system has to encompass also the criteria of social acceptability and ecological effectiveness. The basis of rational waste management planning is the rate of waste accumulation, whose proper selection is the most important task of the planning stage [Kempa 1983]. The amount of generated waste is influenced by economic, social and infrastructural factors. Determining the groups of elements that affect the change in the amount of generated waste is not enough, as it is not known how strong their interactions are [Malinowski et al. 2009, Tałałaj 2011, Szul et al. 2014]. In the choice of a method allowing to develop a model that predicts the amount of generated waste and is the basis of management planning in a given area, a municipality, should take into account a number of features, which will expectedly have an essential effect on the final outcome. Due to defining a number of features, i.e. mutually correlated quantitative and qualitative variables, it is an interesting alternative to apply methods using cluster analysis and the rough set theory.

Methods of cluster analysis are often used in objects and features clustering. This concept was introduced by Tryon in 1939. Currently, this term includes many different algorithms of classification [Hartigan 1975; Hartigan, Wong 1978; Nęcka 2013; Sneat & Sokal 1973; Ward 1963; Witten & Frank 2000], among which the popular ones are *k-means* and *EM* algorithms. Generally, it can be said that cluster analysis is an exploratory analysis of data, aiming at extracting objects from a large set in such a way that elements belonging to one group are as homogenous as possible within particular groups and as different as possible from objects belonging to other groups. Cluster analysis methods allow to identify structures present in a set, but they do not explain the mechanism of their creation.

A classic *k-means* clustering algorithm was popularised by Hartigan [Hartigan 1975; Hartigan, Wong 1978]. During its implementation, as the first step, observations are randomly allocated to an established *k* number of clusters. Next, observations are transferred between the clusters in such a way that the means in the clusters (for all variables) are as different from one another as possible. As the optimal number of clusters is not known, it has to be determined by an expert or on the basis of the developed algorithms. Usually the number of clusters is selected with the use of *v-fold cross validation* algorithm. The idea of this method is to divide the whole sample into *v* subsets, and then, the same analysis is in turn performed on the observations of *v-1* subsets, i.e. on the so-called training sample. Next, the results of the analysis are applied to the data of the training

sample, which had not been so far used in the analysis, and the measure of predictive power is determined on its basis. The results of *v* repetitions are aggregated and give one assessment of the model's stability, i.e. its ability to predict new observations.

Another popular procedure is *EM* method cluster analysis, whose detailed description was presented by Witten and Frank [2000]. This algorithm calculates the probability of cluster membership, with the assumption of one or many probability distributions. The aim of the algorithm is to maximise the general probability for a given division into clusters. The advantage of *EM* algorithm over *k-means* algorithm is the fact that it can be used both for quantitative and qualitative variables.

Another method is the rough set theory, which was introduced in the 1980s by professor Zdzisław Pawlak [1982]. It is a relatively new mathematical method of data analysis. It is used as a tool for the synthesis of advanced and effective analyses methods and for the reduction of data sets [Muruszkiewicz 2004]. Rough sets serve as a methodology in the process of discovering knowledge in databases. This process is usually both iterative and interactive (a lot of decisions are made by the user) [Sodłacki 2001]. The rough set theory is very significant in the process of data extraction due to the fact that it is one of the fastest developing areas of artificial intelligence. It is used to describe imprecise, uncertain knowledge, to model decision-making systems and approximate reasoning [Semeniuk-Polskowska 2001]. Methodology of deduction that uses rough set theory refers only to the qualitative nature of the objects' features. This causes limitations and difficulties when we deal with quantitative and not qualitative features. In such a case, the integration of valued tolerance relation proves useful [Stefanowski and Tsoukias 2000]. It allows to introduce more flexibility to the rough set theory when examining data and to analyse observations expressed in a quantitative form. This course of action is aimed at selecting the most important conditional attributes which are necessary to make the right decision in individual decision-making subgroups [Renigier 2008]. Standard assumption of the rough set theory is based on the indiscernibility relation concept as a precise equivalence relation, which means that the objects will be indiscernible only when they have similar attributes (0 – 1 system). Application of valued tolerance relation to the rough set theory allows determination of the upper and lower approximation of a set with different levels of indiscernibility relation [d'Amato 2006]. Owing to this solution, one can compare two sets of data and achieve a result in the 0...1 range, which constitutes the level of indiscernibility relation. This range is a membership function derived from the assumptions of the fuzzy set theory. The closer the result to 1, the more similar are the objects (indiscernible) with regard to the analysed attribute, and the closer the result to 0, the more discernible they are [Renigier-Biłozor 2008, 2008a, Renigier-Biłozor, Biłozor 2013, Stefanowski 2001].

The rough set theory is a certain theory of knowledge (theory of information systems) and serves as a tool for describing uncertain, imprecise knowledge, for modelling approximation reasonings and decision making systems as well as systems of feature and classification recognition. The results of theoretical study within RST involve logics, set theory, knowledge representation, data filtering, algorithmic problems connected with information systems [Nguyen 2013, Semeniuk-Polskowska 2001]. Although developed a short time ago, rough set theory is used in a number of new fields of study. Nowadays, it is used both in medicine, pharmacology, economics, banking, chemistry, sociology, acoustics, linguistics, general engineering, neuroengineering as well as in the diagnostics of machines, geography, land management and environmental engineering – the publications of results can be found, inter alia, in [Bondar-Nowakowska 2000, Deja 2000, Hachoł, Bondar-Nowakowska and Reinhard 2008, Komorowski et al. 1999, Mróżek and Płonka 1999, Pawlak 1997, Polkowski and Skowron 2001, Renigier 2006, Renigier-Biłozor 2011, Renigier-Biłozor and Biłozor 2007, 2013, Renigier-Biłozor and Wiśniewski 2012, Słowiński 1999, Szul et al. 2014].

METHODOLOGY

The research was limited to two commonly used methods of cluster analysis, i.e. *k-means* and *EM* methods. In both the methods, calculations began with dividing objects into the training and test set, then input variables were standardised and the number of clusters established. *V-fold cross validation* was performed in order to establish an optimal number of clusters. Next, "Generalised cluster analysis" module available in *Statistica 10.0* program was used. It transferred objects between these clusters in such a way as to minimise variability within the clusters and maximise variability between the clusters. The following distances were set while performing analyses with *k-means* method:

Euclidean distance – geometric distance within a multidimensional space:

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n}\left(X_i - Y_i\right)^2}\ , \qquad (1)$$

Manhattan distance – sum of differences measured along dimensions:

$$\text{Manhattan distance} = \sum_{i=1}^{n} |X_i - Y_i|\ . \qquad (2)$$

Chebyshev distance – it is used when we want to define two objects as „different", when they differ in one dimension:

$$\text{Chebyshev distance}\ (X, Y) = \max|X_i - Y_i|\ , \qquad (3)$$

As the next step, these methods were compared with the calculations using the rough set theory, whose detailed methodology was presented inter alia in the following studies: [Pawlak 1997, Renigier-Biłozor 2008, Szul et al. 2014].

In this method, municipalities selected for analysis were divided in a way analogous to the analysis of clusters into two subsets: the training set and the test set. Objects within the training set were presented in the form of a decision table where the features characterising the municipalities were

described with condition attributes. The decision attribute of the rate of mass accumulation of waste in households, kg·(person·year)$^{-1}$ was also established. Next, the matrix of „valued tolerance relation" was calculated for condition attributes:

$$R_j(x,p) = \max\left(\sum_{j=1}^{n} R_j(x,p)\right), \quad (4)$$

where:
R_j – valued tolerance relation,
x – attribute of the considered object,
p – attribute belonging to the conditional part of the considered decision rule,
where

$$R_j(x,y) = \frac{\max(0,\min(c_j(x)\ c\ (y) + k - \max((c_j(x)\ c\ (y)))}{k}, \quad (5)$$

where:
$R_j(x,y)$ – is the relation between two sets with a membership function [0,1],
$c_j(x), c_j(y)$ – variable of the analysed object,
k – coefficient taken as a standard deviation in the set of a given attribute of the analysed object.

After having determined indiscernibility classes for condition and decision attributes, quality and accuracy indicators of approximations within individual decisions sub-clusters were calculated:

$$\gamma_c(X) = \frac{card(POS_c(U))}{card(\underline{O}X_i)}, \quad (6)$$

where:
$\underline{O}X$ – the number of lower approximation objects (cardinality of the lower approximation of X set),
$\bar{O}X$ – the number of upper approximation objects (cardinality of the upper approximation of X set),
POS_c – the number of objects in the indiscernibility class of a decision attribute.

$$\beta_c(X) = \frac{card(\underline{O}X)}{card(\bar{O}X)}, \quad (7)$$

where:
$\underline{O}X$ – the number of lower approximation objects (cardinality of the lower approximation of X set),
$\bar{O}X$ – the number of upper approximation objects (cardinality of the upper approximation of X set).

Having distinguished representative decision rules the author determined the rate of mass accumulation of waste. For this purpose the municipalities from the test set were used. Applying valued tolerance relation (VTR), the author checked to which of the decision rules selected above the analysed municipality has the highest level of membership.

The quality of the match between the predicted rate of mass accumulation of waste in households and its real value was estimated by determining the value of error:

$$ME = \frac{1}{n} \cdot \sum_{i=1}^{n} d_i - d_i^p, \quad (8)$$

$$APE = \frac{\left|d_i - d_i^p\right|}{d_i} \cdot 100, \quad (9)$$

$$MAPE = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\left|d_i - d_i^p\right|}{d_i} \cdot 100, \quad (10)$$

where:
d_i – the rate of mass accumulation of waste in the households, kg·(person·year)$^{-1}$'
d_i^p – predicted rate of mass accumulation of waste in the households, kg·(person·year)$^{-1}$.

RESULTS

Analyses presented in the study were performed on the basis of statistical data from Małopolska Voivodeship of 2012 [GUS 2013]. During this time in the analysed area 1001 thousand Mg of waste was generated, which constitutes 7,8% of the waste stream on a national scale. The indicator expressing the amount of waste produced per one inhabitant in 2012 was 300 kg·(person·year)$^{-1}$ for Małopolska Voyvodeship and it was only slightly lower than the national average, i.e. 314 kg·(person·year)$^{-1}$. An average household in Małopolska produces 118 kg·(person·year)$^{-1}$, while in the rural areas this value is lower by about 45%.

The comparative analysis of individual methods' effectiveness while determining the rate of waste accumulation for rural areas was done on the example of the set of 60 randomly chosen rural municipalities and rural areas of urban and rural municipalities of Małopolska Voivodeship. The number of objects within the set was chosen in a way to enable the level of confidence of 95%. Then, the municipalities chosen for the analysis were divided into two subsets: the training set containing 40 objects and the test set comprising 20 objects.

Objects within the training set were presented in the form of a decision table (Table 1) where the features characterising the municipalities were marked with symbols $c_1 \div c_7$, and the rate of mass accumulation of waste in households, which is a decision attribute was marked with *d* symbol.

Table 1. Information system (decision table)

Municipality/ object number	Condition attributes							Decision attribute
	c1	c2	c3	c4	c5	c6	c7	d

Source: own study on the basis of General Statistical Office's data

For the aforementioned attributes, domains were determined according to the following assumptions:
c_1 – population density, [people·km^{-2}],
c_2 – average area of agricultural land, [ha],
c_3 – building's age rate (established as a weighted arithmetic mean of the number of buildings from different periods of time i.e. before 1944, 1945-1970, 1971-1988, 1989-2002, 2003-2012),
c_4 – participation of buildings heated with natural gas,
c_5 – municipality type, (1 – suburban, 2 – tourist, 3 – agricultural),
c_6 – participation of households deriving income from agricultural activity,

c_7 – income rate (municipalities' own income – participation in taxes comprising national budget income personal income tax), [PLN·(person·year)$^{-1}$],

d – the rate of mass accumulation of waste in the households, [kg·(person·year)$^{-1}$].

The values of particular attributes were established on the basis of statistical data included in the Regional Data Bank of General Statistical Office for 2012 and 2010 General Agricultural Census available on the General Statistical Office's website [GUS 2013].

With the use of condition attributes ($c_1 – c_7$), the training set was divided by *Statistica 10.0* program into an optimal number of clusters on the basis of *v-fold cross validation*. Observations from the test set were assigned to different groups. Then, on the basis of the training set, mean values of the decision attribute ($d_{śr}$ [kg·(person·year)$^{-1}$]), its variability and the coefficient (V[%]) were determined for individual clusters. The results of achieved analyses are presented in Table 2.

Table 2. Decision attribute's variability for determined clusters

cluster:	Clustering algorithm:							
	k-means – distance between the objects:						EM	
	Euclidean		Manhattan		Chebyshev			
	$d_{śr}$	V	$d_{śr}$	V	$d_{śr}$	V	$d_{śr}$	V
1	50	47	52	36	53	33	60	40
2	55	26	56	35	93	31	84	34
3	86	32	85	33			97	38
4	107	26	107	26				

As the final step, individual clusters of the test set were attributed with decision algorithm values i.e. the rate of mass accumulation of waste from households, determined on the basis of the training set. The achieved values were compared with real data and the error level was established. The achieved results are presented in Table 3.

The performed analyses show that the mean value of the rest for all methods of cluster analysis determined for the training set was 0 [kg·(person·year)$^{-1}$]. For the training set it oscillated between – 3 [kg·(person·year)$^{-1}$] for the rough set theory and 22 [kg·(person·year)$^{-1}$] for *k-means* method, for which the Euclidean distance between objects was calculated. In case of the training set observations, *k-means* method of clustering generated underestimated prognoses independently of the type of distance calculated between objects, contrary to *EM* method and the rough set theory. The average MAPE for *k-means* and *EM* algorithm ranged between 33 and 41% for the training set and between 20% and 40% for the test set. The rough set theory was characterised by a much better quality of prognosis, for which MAPE value was 14%.

With the purpose of a better presentation of the changes in differences generated for individual methods, empirical distribution function for APE has been shown in Figure 1.

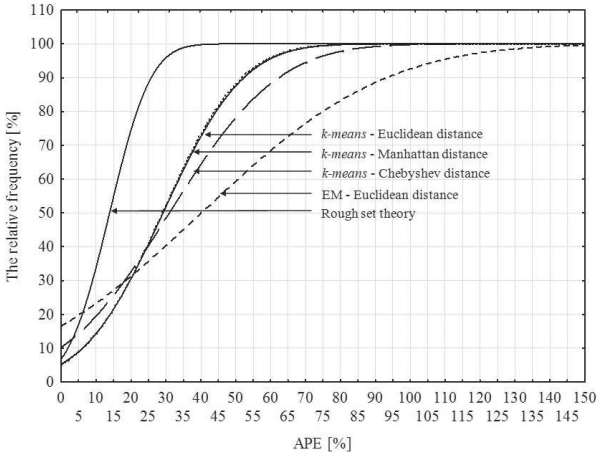

Fig. 1. Comparison of empirical distribution function for APE

On the empirical distribution function for APE diagram (Fig. 1) we can see that the rough set theory was characterised by the best quality of prognosis of the rate of mass accumulation of waste in households for the test set. Participation of error with the lowest value was similar to *k-means* clustering method and it was less than 10% of the observations. The advantage of rough set theory for such a small test was very much visible for bigger errors of the prognosis. Maximum APE value determined with this method did not go beyond 40% whereas for *k-means* method it was around 80-90%. The lowest quality of prognosis despite the highest percentage of low value errors was characteristic of *EM* method of clustering. Low value errors accounted for almost 20% of observations, but at the same time the maximum values of errors were as high as150%.

CONCLUSIONS

The performed comparative analyses show that the average mean error – ME for all methods of cluster analyses determined for the training set was 0 [kg·(person·year)$^{-1}$]. For the training set it varied from – 3 [kg·(person·year)$^{-1}$] for rough set theory to 22 [kg·(person·year)$^{-1}$] for *k-means* method, for which the Euclidean distance between objects was calculated. In case of the training set observations, *k-means* method of clustering generated underestimated prognoses independently of the type of distance calculated between objects, contrary to *EM* method and the rough set theory.

Table 3. Characteristic of the estimation error of the rate of mass accumulation of waste in households

Sample:	Error clustering algorithm:									
	k-means – distance between the objects:						EM		Rough set theory	
	Euclidean		Manhattan		Chebyshev					
	ME	MAPE	ME	MAPE	ME	MAPE	ME	MAPE	ME	MAPE
training	0	33	0	35	0	35	0	41	-	-
test	22	29	10	31	20	29	-6	40	-3	14

The average MAPE of prognosis for *k-means* and *EM* algorithm ranged between 33 and 41% for the training set and between 20% and 40% for the test set. The rough set theory was characterised by the best quality of prognosis, for which MAPE value was 14%.

The performed analyses have proved that rough set theory should be used for estimating the rate of mass accumulation of waste from rural areas especially when the number of objects within the cluster is low.

REFERENCES

1. **Bondar-Nowakowska E., 2000:** Oddziaływanie robót konserwacyjnych na florę i faunę koryt wybranych cieków nizinnych. Zesz. Nauk. AR Wrocław. Rozprawy CLXXIII, nr 391, 100.
2. **Deja R., 2000:** Zastosowanie teorii zbiorów przybliżonych w analizie konfliktów. Praca doktorska. Instytut Podstaw Informatyki Polskiej Akademii Nauk.
3. Główny Urząd Statystyczny. Bank Danych Lokalnych. http://stat.gov.pl/bdl/app/strona.html?p_name=indeks.
4. **Hachoł J., Bondar-Nowakowska E., Reinhard A., 2008:** Oddziaływanie wybranych elementów koryta cieku na zbiorowiska naczyniowych roślin wodnych. Infrastruktura i Ekologia Terenów Wiejskich. Nr 7/2008, Polska Akademia Nauk, Oddział w Krakowie, 255-266.
5. **Hartigan J. A., Wong M. A., 1978:** Algorithm 136. A k-means clustering algorithm. Applied Statistics. 28. 100.
6. **Hartigan, J. A., 1975:** Clustering algorithms. New York: Wiley.
7. **Kempa E.S., 1983:** Gospodarka odpadami miejskimi. Arkady. Warszawa.
8. **Komorowski, J. Pawlak, Z., Polkowski, L. & Skowron, A., 1999:** Rough sets: A tutorial." Rough fuzzy hybridization: A new trend in decision-making. 3-98.
9. **Konieczna R., Kulczycka J., 2011:** Analiza ekonomiczna systemów gospodarki odpadami. cz. II. IGSMiE PAN, Karków.
10. **Malinowski M., Krakowiak-Bal A., Sikora J., Woźniak A., 2009:** Ilości generowanych odpadów komunalnych w aspekcie typów gospodarczych gmin województwa małopolskiego.. Infrastruktura i Ekologia Terenów Wiejskich. Nr 9, Polska Akademia Nauk, Oddział w Krakowie, 181–191.
11. **Mróżek A., Płonka L. 1999.** Analiza danych metodą zbiorów przybliżonych. Akademicka Oficyna Wydawnicza PLJ, Warszawa.
12. **Nęcka K., 2013:** Selection of decisive variables for the construction of typical end user power demand profiles. TEKA Komisji Motoryzacji i Energetyki Rolnictwa Vol. 13 No 2. Lublin. 1 91-96.
13. Nutech Solution – Science for Businnes. 2005. http://www.nutechsolutions.com.pl/.
14. **Pawlak Z. 1997.** Rough set approach to knowledge-based decision support. European Journal of Operational Research , 99 (1), 48-57.
15. **Polkowski, L. Skowron, A., 2001:** Rough mereological calculi of granules: A rough set approach to computation. Computational Intelligence, 17 (3), 472-492.
16. **Renigier M., 2006:** Zastosowanie analizy danych metodą zbiorów przybliżonych do zarządzania zasobami nieruchomości. Studia i Materiały Towarzystwa Naukowego nieruchomości 14(1). 203-213.
17. **Renigier-Biłozor M, Biłozor A., 2013:** Opracowanie systemu wspomagania podejmowania decyzji z wykorzystaniem teorii zbiorów rozmytych oraz teorii zbiorów przybliżonych w procesie kształtowania bezpieczeństwa przestrzeni. Acta Scientiarum Polonorum. Administratio Locorum 12 (1). 38-47.
18. **Renigier-Biłozor M,. 2011:** Analiza rynków nieruchomości z wykorzystaniem teorii zbiorów przybliżonych. Studia i Materiały Towarzystwa Naukowego nieruchomości Vol 19 (1). 107-119.
19. **Renigier-Biłozor M. Wiśniewski R., 2012:** The Effectivencess of real estete market versus efficiency of its participants. European Spatial Research and Policy. Vo. 19 nr 1. 95-110.
20. **Renigier-Biłozor M., Biłozor A., 2007:** Application of the rough set theory and the fuzzy set theory in land management. Aplication 28 june 2007 r. Annual conference The European Real Estate Society – ERES. Londyn.
21. **Śleszyński P., 2009-2010:** a) Delimitacja obszarów wiejskich o najgorszych wskaźnikach rozwoju społeczno-gospodarczego; b) Delimitacja obszarów wiejskich o najgorszych wskaźnikach dostępności do usług publicznych; c) opracowania map i konsultacje kartograficzne, IGiPZ PAN, Warszawa, 7+5 + mapy, dla: Ministerstwo Rozwoju Regionalnego (Krajowa Strategia Rozwoju Regionalnego).
22. **Słowiński R., 1992:** Intelligent decision support. Applications and advances of the rough sets theory. Kluwer Academic Publshers., Dordrecht.
23. **Sneath, P. H. A., & Sokal, R. R., 1973:** Numerical taxonomy. San Francisco: W. H. Freeman & Co.
24. **Szul T., Knaga J., Nęcka K., 2014:** Application of rough set theory for establishing the rate of mass accumulation of waste in the households in rural areas. Ecological Chemistry and Engineering S. w druku.
25. **Tałałaj I.A., 2011:** Wpływ wybranych czynników społeczno-ekonomicznych na zmiany ilości odpadów komunalnych w województwie podlaskim. Inżynieria Ekologiczna. Nr 25.
26. **Tryon, R. C., 1939:** Cluster Analysis. Ann Arbor, MI: Edwards Brothers
27. **Ward, J. H. 1963.** Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58, 236.
28. **Witten, I., H., & Frank, E., 2000:** Data Mining: Practical Machine Learning Tools and Techniques. New York: Morgan Kaufmann.

PORÓWNANIE PRZYDATNOŚCI ANALIZY SKUPIEŃ ORAZ TEORII ZBIORÓW PRZYBLIŻONYCH DO SZACOWANIA WSKAŹNIKA MASOWEGO NAGROMADZENIA ODPADÓW NA OBSZARACH WIEJSKICH

Streszczenie. W pracy przedstawiono porównanie metod grupowania *k*-średnich i *EM* oraz Teorii Zbiorów Przybliżonych do wyznaczania wskaźnika masowego nagromadzenia odpadów odbiorców wiejskich. Z wykonanych analiz porównawczych wynika, że średnia wartość błędu MAPE dla algorytmu *k-średnich* i *EM* zawierała się w przedziale od 33 do 41% dla zbioru uczącego i w zakresie od 20% do 40% dla zbioru testowego. Dużo lepszą jakością prognozy wskaźnika charakteryzowała sie Teoria Zbiorów Przybliżonych, dla której wartość MAPE wyznaczona dla zbioru testowego kształtowała sie na poziomie 14%.

Słowa kluczowe: analiza skupień, gospodarstwa domowe, odpady, Teoria Zbiorów Przybliżonych.