DE GRUYTER OPEN

INTERNATIONAL
Agrophysics
www.international-agrophysics.org

Int. Agrophys., 2014, 28, 257-260
doi: 10.2478/intag-2014-0015

Note

Application of the Akaike criterion to detect outliers for the analysis of ash content in barley straw

Andrzej Kornacki

Department of Applied Mathematics and Computer Science, University of Life Sciences, Akademicka 13, 20-950 Lublin, Poland

Received March 17, 2013; accepted July 11, 2013

A b s t r a c t. This study presents the method of detection of outliers based on the Akaike information criterion. This method has been applied to experimental data on ash content resulting from the combustion of barley straw.

K e y w o r d s: biomass, barley straw, ash content, Akaike criterion, outliers

INTRODUCTION

In recent years there has been a growing tendency worldwide to use renewable energy sources (RES). One of them is biomass that is produced in agriculture. At present it is the third most important energy source after coal and petroleum. Biomass satisfies worldwide energy demand to a large extent (Roszkowski, 2008; Sotannde, 2010; Werther *et al*., 2000). In Poland biomass plays definitely the leading role among the renewable energy sources (Szyszlak-Bargłowicz *et al*., 2012). The energy obtained from biomass constitutes 85.5% of total energy from renewable energy sources (GUS, 2010).

One of the green energy sources obtained on a large scale in agriculture is straw. The annual production of straw in Poland amounts to *ca* 25 mln t (Dziewanowska and Dobek, 2006, 2009; Gradziuk, 2006; Niedziółka and Zuchniarz, 2006).

From the point of view of biomass energy properties the most important energy indices include the heat of combustion and the calorific value. For ecological reasons it is also important that environmental pollution during biomass combustion be as low as possible (Kowalczyk-Juśko *et al*., 2009; Maj and Piekarski, 2013; Majtkowska and Majtkowski, 2005; Nilsson *et al*., 2006, 2011; Wawrzosek and Piekarski, 2006). It is for that reason that the research, inter alia on the percentage of ash as a result of biomass combustion, is carried out.

It is often the case that among the results of such research we can identify the results that grossly differ from the remaining ones. They are called outliers. During the statistical analysis of test results it is important to identify whether outliers come from a different population than the rest of the results, whether they result from equipment failure *etc*. In such a case they should be rejected. It is also possible (although fairly unlikely) that some strange observations appear at the same distribution as for the remaining results. In such a case, such observations should be preserved for further statistical analysis increasing that way its effectiveness.

Hypothesis testing methods are most often used in order to detect unexpected observations. Research for one-dimensional (univariate) normal sample was conducted by Breuning *et al.* (2000), Grubbs (1969), Ferguson (1961), Ramaswamy *et al.* (2000). In the multi-dimensional normal model the rejecting of outliers was studied by Rousseeuw and Leroy (2003), Srivastava and Von Rosen (1998).

However, in the hypothesis testing method the conclusions are dependent on the assumed significance level and can be different for its various values. Moreover, there may appear the 'masking' effect of outliers. As regards the data concerning the strength of plastic elements (Grubbs, 1969) describes a situation when the tests do not discover 1 smallest observation, whereas 2 smallest observations are identified as outliers (some discrepancy).

The current study suggests the application of the Akaike information criterion (AIC) to detect outlier observations. This criterion, which is derived from the information theory, allows the selection, out of models describing experimental data, the one that maximizes entropy (Akaike, 1977). The value of this criterion equals (Sakamoto *et al*., 1986):

*Corresponding author e-mail: andrzej.kornacki@up.lublin.pl

$$\text{AIC}=-2\ln(\text{max likelihood})\ 2K \tag{1}$$

where the maximum likelihood denotes the likelihood calculated for parameter estimators obtained using the maximum likelihood method, whereas K means the number of the parameters.

The model for which the AIC value is the lowest is selected. Such a method does not depend on the significance level, the number of outliers and the fact whether the unexpected observations are the lowest or the highest.

Let us consider the n sample of the observation which, after ordering according to rising values, forms the following set:

$: x_{(1)} \le x_{(2)} \le x_{(n)}$. Therefore $x_{(k)}$ means the *k*-th value of the order statistics $X_{k,n}$. In the subsequent part of this paper the following notation will be used: $\Theta(x;\mu,\sigma^2)$ is the probability density function of the normal distribution with the mean μ and variance σ^2, $\Phi(x;\mu,\sigma^2)$ cumulative distribution function of this distribution, whereas $g_{r,n}(x;\mu,\sigma^2)$ is the density of the *r*-th of the order statistics from the normal population:

$$\Theta(x,\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \tag{2}$$

$$\Phi(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{x}\exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\}dt, \tag{3}$$

and (David, 1979):

$$g_{r,n}(x;\mu,\sigma^2)=B(r,n-r+1)^{-1}\Phi(x;\mu,\sigma^2)$$
$$\left\{1-\Phi(x;\mu,\sigma^2)^{n-r}\right\}\psi(x;\mu,\sigma^2). \tag{4}$$

In the Eq. (4) $B(p,q)$ denotes function:

$$B(p,q)=\int_0^1 t^{p-1}(1-t)^{q-1}dt,\quad p>0,\ q>0. \tag{5}$$

It is known that:

$$B(p,q)=\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}=\frac{(p-1)!(q-1)!}{(p+q-1)}, \tag{6}$$

where: p and q are natural numbers and $p!$ denotes factorial.

We will examine the following model describing data with possible outlier observations set by the density function:

$$S_r(x)=\begin{cases}\Theta(x;\mu,\tau^2) & r=1,\dots,n_1\\ f_{r-n_1,n-n_1-n_2} & r=n_1+1,\dots n-n_2\\ \Theta(x;\mu,\tau^2) & r=n-n_2+1,\dots,n\end{cases}. \tag{7}$$

The model described by Eq. (7) denotes that: median observations: $x_{(n_1+1)},\dots,x_{(n-n_2)}$ are the realization of normal variables with the mean μ and variance: σ^2. The initial results $x_{(1)},\dots,x_{(n_1)}$ and final results $x_{(n-n_2+1)},\dots,x_{(n)}$ are derived from the normal population with the same mean μ but a different variance τ^2. In this model initial and final observations are considered as the 'candidates' for outlier observations.

The likelihood function of Eq. (7) is presented by the formula:

$$L(x;n_1,n_2,\mu,\sigma^2,\tau^2)=\prod_{i=1}^{n_1}\Theta(x_{(1)},\mu,\tau^2)$$
$$\prod_{i=n_1+1}^{n-n_2} g_{i-n_1,n-n_1-n_23}(x_{(i)};\mu,\sigma^2)\prod_{i=n-n_2+1}^{n}\Theta(x_{(i)},\mu,\tau^2) \tag{8}$$

Therefore, the likelihood logarithm is obtained in the form of:

$$l_1=-\frac{1}{2}\left\{n\ln 2\pi+\sum_{i=1}^{n}\ln\sigma^2(i)+\sum_{i=1}^{n}\frac{(x_{(i)}-\mu)^2}{\sigma^2(i)}\right\}-$$
$$\sum_{i=n_1+1}^{n-n_2}\left[\ln B(j,k-j+1)-(j-1)\ln\left\{\Phi(x_{(i)})\right\}-\right.$$
$$\left.(k-j)\ln\left\{1-\Phi(x_i)\right\}\right], \tag{9}$$

where :

$$j=i-n_1,\quad k=n-n_1-n_2, \tag{10}$$

and

$$\sigma^2(i)=\begin{cases}\sigma^2, & i=n_1+1,\dots,n-n_2\\ \tau^2, & i=1,\dots n_1 \text{ or } i=n-n_2+1,\dots,n.\end{cases} \tag{11}$$

Thus the value of Akaike criterion *ie* the minimum value (1) equals:

$$AIC(i,j)=\begin{cases}-2l_1(x;i,j,\hat\mu,\hat\sigma^2)+2\times 2\ (i=j=0)\\ -2l_1(x;i,j,\hat\mu,\hat\sigma^2,\tau^2)+2\times 3\ (i\neq 0 \text{ or } j=0),\end{cases} \tag{12}$$

where: $\hat\mu,\hat\sigma^2,\hat\tau^2$ denote parameter estimators achieved using the minimum likelihood method.

MATERIAL AND METHODS

In the Department of Power Engineering and Vehicles of the University of Life Sciences in Lublin tests were conducted to determine ash content in barley straw (Maj, 2011). The tests were conducted according to the standard

PN-80/G-04512. The measurements were carried out for two different combustion temperatures: 600 and 815°C and two levels of moisture content (Table 1). All measurements were conducted in six replications. The combustion of biomass was carried out using the muffle furnace Nabertherm L3/11/B180. It is equipped with ceramic heating plates with integrated heating wire which is protected against splashes and exhaust emissions. The muffles are made of refractory clay with chamotte addition. The maximum combustion temperature in this furnace is 1 200°C.

The determination of ash content involves weighing the analytical sample containing 1-2 g of fuel and placing it in the heated furnace and then heating it to the temperature of 500°C for 30 min and after another 30-60 min by ±15°C to the temperature determined during the test. The sample is left at that temperature for 90 min. After it is cooled down and weighed, the sample is heated again for 15 min at 500°C. This procedure is repeated until the constant mass is achieved. The ash content in the analytical sample of solid fuel (A^a, %) is calculated according to the formula:

$$A^a = \frac{m_3 - m_1}{m_2 - m_1} 100, \tag{13}$$

where: container mass: m_1 – that was heated up and cooled down, m_2 – with the weighed amount of solid fuel, m_3 – with ash.

The sample and the crucibles were weighed using an analytical balance with 0.0001 g accuracy. The combustion took place in high temperature resistant ceramic crucibles. The heating programme of controller B180, which is installed in the furnace, was linear and had an interruption time.

RESULTS AND DISCUSSION

Table 1 presents the results of the percentage of ash in biomass pellets for barley straw obtained during the summer and winter season (Maj, 2011).

The results suggest that observations 19.13 and 18.98 are outliers. That will be checked using the Akaike information criterion. Using the Eq. (12) the AIC values for various configurations of outliers were calculated. The results obtained are presented in Table 2.

The data from Table 2 indicate that observation 19.13 and 18.98 are an outlier because for this configuration the Akaike criterion value (marked in the table as *) is the lowest. This conclusion is supported by classical statistical tests used to detect outliers (Grubbs, 1969). For one outlier the tests are as follows:

$$T_1 = \frac{\bar{x} - x_{(1)}}{s}, \quad T_n = \frac{x_{(n)} - \bar{x}}{s}, \tag{14}$$

where: S denotes sample standard deviation, and:

$$L_1 = \frac{nS_1^2}{nS^2}, \quad L_n = \frac{nS_n^2}{nS^2}, \tag{15}$$

T a b l e 1. Percentage of ash in barley straw

Moisture content (%)	Ash content (%)	Moisture content (%)	Ash content (%)
Summer		Winter	
10.48	6.03	11.60	6.74
	6.40		6.64
	6.72		6.77
	19.13		**18.98**
	6.32		6.78
	6.38		6.76

T a b l e 2. Values of the Akaike information criterion for the percentage of ash in barley straw

Lowest outliners	Highest outliners			
	Summer			
		None	**19.13**	**19.13** **6.72**
	None	35.98	4.6*	11.44
	6.03	37.50	11.79	10.78
	6.03 **6.32**	37.75	20.44	17.34
	Winter			
		None	**18.98**	**18.98** **6.78**
	None	36.23	-8.75*	2.68
	6.64	37.18	-8.62	4.33
	6.64 **6.64**	38.88	1.63	14.23

In Eq. (15) the following notations are accepted:

$$\begin{cases} nS^2 = \sum_{i=1}^{n} (x_{(i)} - \bar{x})^2, \quad nS_1^2 = \sum_{i=2}^{n} (x_{(i)} - \bar{x}_1)^2, \\ \bar{x}_1 = \frac{1}{n-1} \sum_{i=2}^{n} x_{(i)}, \quad nS_n^2 = \sum_{i=1}^{n-1} (x_{(i)} - \bar{x}_n)^2, \\ \bar{x}_n = \frac{1}{n-1} \sum_{i=1}^{n-1} x_{(i)}. \end{cases} \tag{16}$$

In our case it is said that we have:

$$\begin{cases} T_6 = \frac{19.13 - 8.50}{4.76} = 2.234 \\ L_6 = \frac{0.2416}{135.92} = 0.0018 \ . \end{cases} \tag{17}$$

Because T_6 is greater than 5 of the critical value 1.996 and L_6 is smaller than 0.2002 respectively, thus all tests confirm that observation 19.13 is an outlier. Similarly, observation 18.98 is an outlier in relation to the values of the statistics:

$$\begin{cases} T_1 = \dfrac{18.98 - 8.78}{4.56} = 2.236 \\ L_1 = \dfrac{0.0129}{124.90} = 0.0001 \ . \end{cases} \tag{18}$$

CONCLUSIONS

1. A method of detecting outliers was suggested based on the Akaike information criterion as an alternative to classical statistical tests. The suggested method is an objective procedure independent of the assumed significance level, quantity of outliers and of whether the 'suspicious' observations are the lowest or the highest.

2. The explicit indication of the method based on the Akaike criterion allows us to avoid the 'masking' effect of outlier observations.

REFERENCES

Akaike H., 1977. On entropy maximization principle. Proc. Symp. Applications of Statistics (Ed. P.R. Krishnaiah), North-Holland, Amsterdam, The Netherlands.

Breuning M., Kriegel H.P., Ng T.R., and Sander J., 2000. LOF: Identifying density-based local outliers. Proc. Conf. ACM SIGMOD, On Management of Data, May 5-7, Dalles, TX, USA.

David G., 1979. Order Statistics (in Russian). Mockva Nauka Press, Moskva, Russia.

Dziewanowska M. and Dobek T., 2006. Heat values of leaves of some species of trees collected in urban areas. Acta Agrophysica, 141, 551-558.

Dziewanowska M. and Dobek T., 2009. Energy and environmental assessment of the process of obtaining heat from the combustion of leaves collected in urban areas. Inżynieria Rolnicza, 1(110), 115-122.

Ferguson T.S., 1961. On the rejection of outliers. Proc. 4th Berkeley Symp. Math. Statistic. Prob., 1, 253-287.

Gradziuk P., 2006. Technical and economic aspects of the use of straw and grain for energy purposes. Regional Renewable Energy Forum.

Grubbs F.E., 1969. Procedures for detecting outlying observations in samples. Technometrics, 11, 1-21.

GUS, **2010**. Energy from renewable sources in 2009. Warsaw, Poland.

Hadley P. and Fordham R., 2003. Vegetables of temperate climates. Miscellaneous Root Crops. Encyclopedia of Food Sciences and Nutrition, (Eds B. Caballero, L. Trugo, P. Finglas). Elsevier, New York, USA.

Kościk B., 2007. Energy materials of agricultural origin (in Polish). Highest Technical School, Jaroslaw, Poland.

Kowalczyk-Juśko A., Kościk B., and Kwapisz M., 2009. Possibilities and limitations of agricultural utilization for energy purpose. Scientific Papers of Polish Society of Ecological Engineering and Polish Soil Science Society South-Eastern Branch, 11, 155-160.

Nilsson D., Bernesson S., and Hansson P.A., 2011. Pellet production from agricultural raw materials – A system study. Biomass Bioenergy, 35, 679-689.

Nilsson L., Pisarek M., Buriak J., Oniszk-Popławska A., Bućko P., Ericsson K., and Jaworski Ł., 2006. Pellet production from agricultural raw materials - A system study. Energy Policy, 34, 2263-2278.

Maj G., 2011. Gaining energy from pellets made of biomass. Ph.D. Thesis, University of Life Sciences, Lublin, Poland.

Maj G. and Piekarski W., 2013. Cultivation conditions, pellet manufacturing parameters and physicochemical properties of prairie cordgrass (Spartina pectinata) as a dedicated energy crop. Acta Agrophysica, 20(1), 103-112.

Majtkowska G. and Majtkowski W., 2005. Grasses as a source of energy (in Polish). Agro Serwis, 9, 94-97.

Niedziółka I. and Zuchniarz A., 2006. Energy analysis of selected types of biomass plant. Motrol., 8A, 232-237.

PN-80/G-04512, **2002.** Designation of ash content by balance method (in Polish). Polish Committee for Standardization, Warsaw, Poland.

Ramaswamy S., Rastogi R., and Shim K., 2000. Efficient algorithms for mining outliers from large data sets. Proc. ACM SIGMOD Conf. Management of Data, Dallas, TX, USA.

Roszkowski A., 2008. Biomass versus agriculture. Inżynieria Rolnicza, 10(108), 201-208.

Rousseeuw P. and Leroy A., 2003. Robust Regression and Outlier Detection. Wiley Press, New York, USA.

Sakamoto Y., Ishigura M., and Kitagawa G., 1986. Akaike Information Criterion Statistics, Reidel Comp. Press, Tokyo, Japan.

Sotannde O.A., Oluege A.O., and Abah G.B., 2010. Physical and combustion properties of charcoal briquettes from neem wood residues. Int Agrophys., 24, 189-194.

Srivastava M.S., 1997. Slippage tests of mean for a single outlier in multivariate normal data. Amer. J. Manag. Sci., 17, 117-145.

Srivastava M.S. and Von Rosen D., 1998. Outliers in multivariate regression models. J. Mult. Anal., 65, 195-208.

Szyszlak-Bargłowicz J., Zając G., and Piekarski W., 2012. Energy biomass characteristics of chosen plants. Int Agrophys., 26, 175-179.

Wawrzosek J. and Piekarski W., 2006. Model of CO emission level of exhaust gases in tractor engines fed with biofuels. Int. Agrophysics, 20, 353-358.

Werther J., Saenger M., Hartge E.U., Ogada T., and Siagi Z., 2000. Combustion of agricultural residues. Progress Energy Combustion Sci., 26(1), 1-27.