Outlier Detection of the Agricultural Time Series

I Vorotnikov, A Rozanov, M Sidelnikova, S Tkachev and L Volochuk
Saratov State Vavilov Agrarian University, Saratov, Russia
E-mail: nir@sgau.ru

Abstract. To increase the reliability of the results of processing the data of automated monitoring of the dynamics of time series in the agricultural field of activity, it is necessary to assess the probability distribution of the data flow. It is shown that, if the distribution deviates from the normal one, adaptive and robust procedures for identifying and correcting random outliers should be applied.

1. Introduction
In modern agro-industrial production, the values of many environmental, industrial or economic processes are recorded at certain moments of time and, therefore, can be represented as a time series [1].

Real processes, as a rule, are distorted by various types of random and/or systematic errors caused by the uncontrolled influence of external or internal factors of various natures. Statistical data processing can significantly reduce the influence of random factors. An important component of the general procedure for data processing and analysis is the identification and correction of abnormally large or abnormally small quantities, called outliers of random processes.

2. Research methodology
The problem of identifying and correcting outliers consists in the uncertainty fulfillment of standard assumptions about the normality distribution of probabilities and resistance to the action of extremely small, abnormally large or missing values in advance for the series under consideration. In this regard, at the initial stage, when analyzing "raw" data, it is necessary to carry out a procedure for checking the normality distribution of probabilities.

Using an automated procedure recommended to carry out this assessment [2]. Under condition (1), the distribution considered normal:

\[
\left| \frac{MAD}{\sigma} - 0.7979 \right| < \frac{0.4}{\sqrt{n}}
\]  

Here \( n \) – the number of members of the series, \( \sigma \) – the standard deviation of the sample, \( MAD \) – the mean absolute deviation:

\[
\hat{\sigma} = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad i = 1, 2, ..., n, MAD = \frac{1}{n} \sum_{i=1}^{n} |x_i - x_{cp}|,
\]
Here \( x_i \) and \( x_{cp} \) – sample element values (current and average). In certain cases, it is necessary to carry out an additional procedure for checking for normality using more complex criteria for the sample under study. Provided that the normal distribution is confirmed, it is possible to establish “questionable” values of the series. This check is required for each member of the series:

\[
x_i - x_{cp} > t_q \cdot \hat{\sigma}, \quad i = 1, 2, \ldots, n,
\]

Here \( t_q \) is the critical value of the quantile of the Student's distribution, which is determined in accordance with the tables indicators of \( n \) and \( q \). Index \( q \) value is selected from the range of 0.05; 0.01 or 0.001, based on a given level of confidence (95%; 99%, etc.). If condition (3) performed, then the corresponding sample value recognized as a random outlier and not taken into account. Usually, the dataset does not have a perfectly normal distribution. As a result, the use of parametric estimates has low efficiency.

3. Robust procedure
The most effective way to increase the efficiency of data processing is to use robust estimates. The common robust methods based on m-estimates. The likelihood principle is used as a basis for determining the parameters of the time series [3]. For a sequence of random variables \( x_1, x_2, \ldots, x_n \), having a continuous probability density \( f(x) \), one should consider the logarithm of the function \( L(m) \) with respect to the sample position parameter \( m \):

\[
\ln[L(m)] = \sum_{i=1}^{n} \ln[f(x_i - m)] = -\sum_{i=1}^{n} r(x_i - m) = K(m),
\]

\[
r(x) = -\ln[f(x)], \quad K(m) = -\sum_{i=1}^{n} r(x_i - m)
\]

The maximum likelihood method consists in finding the maximum \( \ln(L(m)) \) or the minimum of the functions \( K(m) \). The definition of \( K(m) \) carried out by differentiating the function \( K(m) \) and solving the equations:

\[
\frac{dK(m)}{dm} = 0, \quad \sum_{i=1}^{n} \psi(x_i - m) = 0,
\]

\[
\psi(x) = -\frac{1}{f(x)} \frac{df(x)}{dx}
\]

Equation solution

\[
\sum_{i=1}^{n} \psi(x_i - m) = 0,
\]

The minimization of the function \( K(m) \) for parameter \( m \) is the maximum likelihood estimate or M-estimate for the parameter \( m \). The function \( \psi(x) \) choses so that the estimate of the desired parameter weakly depends on the outliers in the measurement results.

M-estimates for the scale parameter \( S \) are found by solving the equation:

\[
\sum_{i=1}^{n} \psi \left( \frac{x_i - m}{S_{\text{rob}}} \right) = 0,
\]

Here \( S_{\text{rob}} \) is robust, outlier-resistant estimate of the scale parameter. The \( S_{\text{med}} \) parameter usually used as a sample estimate for \( S_{\text{rob}} \):
Here MED is the operator for calculating the median:

\[ MED = \begin{cases} x_{i+1}, & n = 2 \cdot i + 1 - \text{odd}, \\ \frac{(x_i + x_{i+1})}{2}, & n = 2 \cdot i - \text{even}. \end{cases} \] (11)

The normalization parameter 0.6745 is chosen from the condition that for \( n \to \infty \), the \( S_{med} \) estimate should tend to the usual standard deviation with a normal probability distribution. In this case, the criterion for classifying the “suspicious” values of the series as outliers is the condition:

\[ |x_i - x_{cp}| > k \cdot S_{med}, \quad i = 1, 2, ..., n, \] (12)

Here the value of \( k \) chosen equal to three, according to the semi empirical rule of "three sigma" [3].

4. Results and Discussion

Let us illustrate the effectiveness of the above methods for identifying emissions based on monitoring the dynamics of time series for the agro-industrial complex.

In figure 1 shows the change in the maximum prices for agricultural crops (wheat, grade 3) in the Saratov region [4].

Figure 1 shows the dynamics of prices for agricultural crops (wheat, grade 3, tons) in the Saratov region [1]. The circle defines a random outlier.

Figure 1 shows the change in the maximum prices for agricultural crops (wheat, grade 3) in the Saratov region [1]. An estimate of the normal distribution of probabilities by formula (1) shows that the condition:

\[ \left| \frac{MAD}{\sigma} - 0.7979 \right| = 0.02 < \frac{0.4}{\sqrt{n}} = 0.08 \] (13)

If (13) will be satisfied, then criterion (3) to determine the anomalous values of these series.

Figure 2 shows changes in the maximum prices for oilseeds - sunflower (for example, Altai Territory) [5].
Figure 2. Analysis of sunflower prices in the Altai Territory, rubles / ton [5]. The circle indicates an accidental release.

Based on the results of evaluating the fulfillment of the condition of the normal distribution of probabilities by formula (1), founded that the condition \( n = 25 \):

\[
\left| \frac{MAD}{\sigma} - 0.7979 \right| = 0.132 > 0.4 \sqrt{\frac{1}{n}} = 0.08
\]

As a result, the application of criterion (3) to determine the anomalous values of the analyzed series is not effective. The “suspicious” value highlighted in Figure 2 is significantly identified as an outlier only by the robust procedure (10) - (12) based on m-estimates of the position parameters \( m = MED \) and the scale \( S = S_{med} \).

5. Conclusion

Thus, if at first glance, outliers in the time series in Figures 1 and 2 easily identified visually, the use of the Gaussian type of procedures to identify and correct "suspicious" values can lead to serious errors, since their probabilistic characteristics might be fundamentally different [6-10]. In this regard, it becomes relevant to use alternative estimates in the automation of data processing in the field of agro-industrial complex.

References
[1] Frames P H, Dijk D van and Opschoor A 2014 Time series models for business and economic fore-casting. Cambridge University Press 36 312
[2] Montgomery D C 2006 Applied Statistics and Probability for Engineers. Wiley 4 784
[3] Huber P J and Ronchetti E M 2009 Robust statistics. Wiley 433 380
[4] Wheat price monitoring Retrieved from: https://agro-bursa.ru/prices/wheat/
[5] Sunflower price monitoring Retrieved from: https://agro-bursa.ru/prices/sunflower/
[6] Vorotnikov I L, Rozanov A V and Sidelnikova M V 2019 Econ. of agricul. and proc. Enterprises 10 79-81
[7] Nikiforov A, Kuchumov A, Terentev S, Karamulina I, Romanova I and Glushakov S 2020 Neural network method as means of processing experimental data on grain crop yields E3S Web of Conferences 161 01031
[8] Gnezdova J V, Semchenkova S V and Kuchumov A V 2020 State Backing of Modernization of Main Funds of Agricultural Purpose within a Matter of Increasing Food Security of Russia IOP Conference Series: Earth and Environmental Science 459(6) 062061
[9] Lakomiak A and Zhichkin K A 2019 Photovoltaics in horticulture as an opportunity to reduce operating costs. A case study in Poland. Journal of Physics: Conference Series 1399 044088
[10] Zhichkin K, Nosov V, Zhichkina L, Tkachev S and Voloshchuk L 2020 Prediction methodology for potential damage from misuse of agricultural lands. E3S Web of Conferences 161 01060