OUTLIERS IN TIME SERIES

Luboš Marek

Abstract

In the study of economic time series we work often with large amount of data. Some observations take unusual values and are significantly different from other observations. These data are referred to as outliers. The causes of these outliers can be different. It may be a mistake in the data. Most of these errors are relatively easy to remove if you have a careful control of data. On the other hand, there may occur the data that are observed properly and still have the character of outliers. These observations make it difficult to build right models for time series and can misrepresent any predictions.

In the article we describe the different types of outliers and their integration into the stochastic time series. We work with transfer function models, which are mainly based on ARIMA models and on linear dynamic models. In our paper we show that each outlier must necessarily leave specific track in residuals. This fact is the base for detection outliers, to describe their type and methods of integration into the model. We describe in our article the theory of outliers including their detection. Described theory will be illustrated on practical data.

Key words: time series analysis, stochastic process, outliers, interventions, residuals

JEL Code: C40, F470

Introduction

Time series often contain observations caused by unexpected events, called interventions. Another type may be caused, e.g., by typographical errors when entering data or errors in data aggregation; and it may happen that some observation values are simply misaligned with most of the others. They are called outliers. If the time of and reasons for their occurrence are known, methods of intervention analysis can be applied to them. If the time of outlier occurrence is unknown, it is important to identify outliers and clean the time series from them. The entire theory of outliers in time series is clearly connected with methods of intervention analysis and, consequently, with the theory of dynamic regression models – cf., e.g., (Abraham, Ledolter 1983, Anderson 1976, Pankratz 1991, Granger,

Newbold 1986 and Wei 1990). The example is from monetary area (Stankovičová, Vlačuha, Ivančíková 2013, Bartošová, Forbelská 2012 and Buc, Klieštik 2013).

1 Types of outliers

Three basic types of models have been derived in the theory of outliers, describing their most frequently occurring types. Namely, they are:

- Additive Outlier (AO). It is, in essence, a pulse, and hence can be modeled as an intervention.
- Level Shift (LS). It is a shift; again, it can be described as an intervention.
- Innovational Outlier (IO). As we will see below, this is the most complex type of outlier; unlike in the previous two, an effect of distant observation after time T is present (i.e., in times $T, T+1, T+2, \ldots$).

1.1 Models for outliers

Let us consider time series z_t with zero mean value described with the aid of an AR(1) process; that is

$$(1-\varphi_1 B) z_t = \varepsilon_t, \tag{1}$$

equivalently

$$z_t = \frac{1}{(1-\varphi_1 B)} \varepsilon_t. \tag{2}$$

We will now illustrate outliers using a generated time series z_t governed by an AR(1) model with the parameter value $\varphi_1 = 0.7$. We have generated 100 values of the series to show visual and computational circumstances of the above-mentioned types of outliers. Our AR(1) model is, of course, a special type of the general ARIMA$(p,d,q)(P,D,Q)_L$ model; in this more general case the series z_t would be described by the formula

$$z_t = \frac{\theta(B^L)\theta(B)}{\varphi(B^L)\varphi(B)\Delta_L^D \Delta^d} \varepsilon_t. \tag{3}$$

Regarding general processes, they are supposed to be stationary (either immediately or after relevant transformations) and invertible; these assumptions of course imply certain requirements for their parameters (Coufal 2012). We are now going to illustrate the outliers using a series governed by the selected AR(1) model, it would be easy to generalize our considerations to a general ARIMA model characterized by formula (3).

Let us now suppose that, instead of series z_t, a "contaminated" series Y_t is observed, in which a contaminating component $f(t)$ is added to the original series z_t

$$Y_t = f(t) + z_t, \tag{4}$$

where function $f(t)$ corresponds to one of the above-mentioned outlier types. The contaminating component may be of various types in practice; for the sake of simplicity, we suppose that it can be expressed in the form of a rational transition function, as is usual in the theory of outliers; namely

$$f(t) = \frac{\omega(B)}{\delta(B)} X_t, \tag{5}$$

where X_t is a deterministic binary variable standing for an intervention in series z_t at time t; $\omega(B)$ is a polynomial of degree h in the numerator of the fraction in formula (5); and $\delta(B)$ is a polynomial of degree r in the denominator of the fraction in formula (5).

1.2 Additive outliers

Let us now consider the simplest polynomials in formula (5), i.e., set all coefficients in $\omega(B)$ and $\delta(B)$ equal to zero except for ω_0. Since we are now considering an additive outlier, let us denote ω_0 by ω_A for our purposes. This means

$$f(t) = \omega_0 X_t \tag{6}$$

and consequently

$$Y_t = \omega_A X_t + z_t, \tag{7}$$

where

$$X_t = \begin{cases} 1, & t = T \\ 0, & t \neq T \end{cases}. \tag{8}$$

The series Y_t is thus the same as series z_t except for a sole observation at time $t = T$. At time $t = T$ the value of time series Y_t is either larger (for $\omega_A > 0$) or smaller (for $\omega_A < 0$) than z_t. In this case the *outlier is additive*. Figure 1 shows the effect of an additive outlier in the generated time series z_t: the additive outlier has been artificially introduced into the generated time series by adding the value $\omega_A = 8$ to the observation at time $t = T = 57$. Apart from time $t = 57$, both series are identical and their charts are indistinguishable. The additive outlier is depicted by a dotted line at time $t = 57$.

If we know the time at which the outlier occurs, we can estimate the value of ω_A within the framework of the transition-function model. Using a preliminary approximation, we substitute for z_t in formula (7) from formula (2), obtaining formula

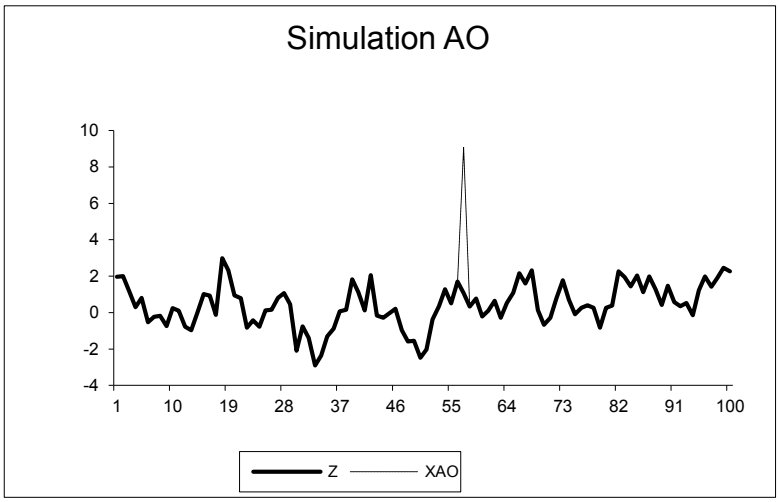


Fig. 1: Original and AO-contaminated time series

Source: Own calculations

$$Y_t = \omega_A X_t + \frac{1}{1-\varphi_1 B}\varepsilon_t . \tag{9}$$

The first term in formula (9) is the pulse intervention variable (known from intervention models), while the second term is the error component governed by an AR(1) model, by which series z_t is also governed. We can see it is the classical transition-function model. In most instances, however, we do not know the time of the outlier occurrence and have to determine that time first.

Let us generalize the assumptions about time series z_t. If we suppose that the series is governed by a general ARIMA $(p,d,q)(P,D,Q)_L$ model, formula (9) becomes

$$Y_t = \omega_A X_t + \frac{\theta(B^L)\theta(B)}{\varphi(B^L)\varphi(B)\Delta_L^D\Delta^d}\varepsilon_t . \tag{10}$$

1.3 Permanent level shift

Let us again consider function $f(t)$ in formula (5) and set all coefficients in $\omega(B)$ and $\delta(B)$ equal to zero except for ω_0 and δ_1. This time, we denote ω_0 by symbol ω_S, and we set $\delta_1 = 1$. Then it is true that

$$1 - \delta_1 B = 1 - B = \Delta, \tag{11}$$

hence

$$f(t) = \frac{\omega_S}{\Delta} X_t, \tag{12}$$

where again

$$X_t = \begin{cases} 1, & t = T \\ 0, & t \neq T \end{cases}. \tag{13}$$

The expression $(\omega_s / \Delta) X_t$ is one of the known ways to describe a jump intervention, which is, in our case, identical with a *permanent level shift* of the time series. formulas (12) and (13) can be rewritten as

$$X_t = \begin{cases} 0, & t < T \\ 1, & t \geq T \end{cases}, \tag{14}$$

where $\delta(B) = 1$, and consequently $f(t) = \omega_S X_t$. Going back to formulas (12) and (13), we can write

$$Y_t = \frac{\omega_S}{\Delta} X_t + z_t. \tag{15}$$

Formula (15) shows that, for a permanent level shift, the contaminated series Y_t is identical with z_t for $t < T$; at time T the permanent shift by a value of ω_S occurs. The values of Y_t will be increased (for $\omega_S > 0$) or decreased (for $\omega_S < 0$) for all $t \geq T$.

Figure 2 shows the effects of a permanent level shift. We have artificially introduced into series z_t an outlier at time $t = 57$ with a value of $\omega_S = 8$. Until time $t = 56$, both series are identical; and after time $t = 57$, series Y_t is permanently shifted by a value of $\omega_S = 8$ (larger by eight units), which is depicted by a dotted line. The original series z_t is depicted by a full line, as the explanations make clear.

Both the additive outlier considered above and the present permanent level shift are such that the outlier effects are clearly visible in charts, despite the AR(1) autocorrelation structure of the underlying series z_t and the presence of random component ε_t. Another

factor is the size of dispersion σ_ε^2 of the random component; if its value is large, visual detection of the outlier may be more difficult. The structure of the ARIMA model for series z_t may also play a role by partially suppressing the outlier effect.

If we know the time at which the outlier occurs, we can, similarly to the instance considered above, estimate the value of ω_S within the framework of the transition-function model.

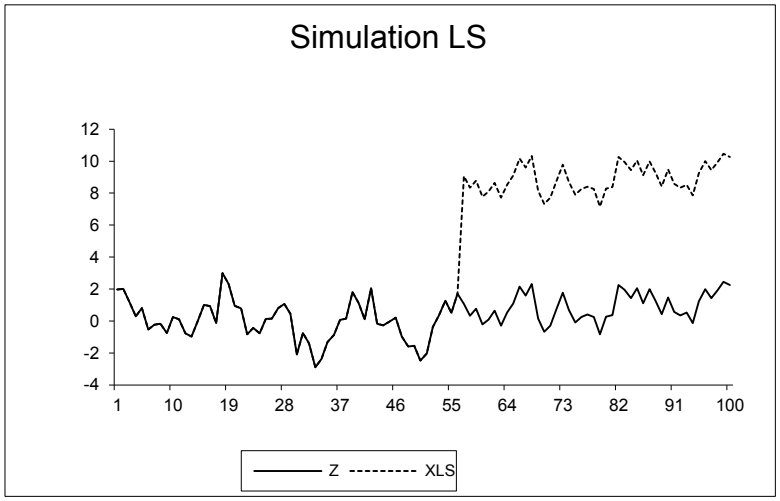


Fig. 2: Original and LS-contaminated time series

Source: Own calculations

Using a preliminary approximation, we substitute for z_t in formula (15) from formula (2), obtaining

$$Y_t = \frac{\omega_S}{\Delta} X_t + \frac{1}{1-\varphi_1 B} \varepsilon_t . \tag{16}$$

The first term in formula (16) is the intervention variable, and the second term is the error component governed by an AR(1) model, by which series z_t is also governed. We can again see that it is the classical transition-function model.

Similarly to the instance of additive outlier, we can consider a general ARIMA$(p,d,q)(P,D,Q)_L$ model for z_t, obtaining

$$Y_t = \frac{\omega_S}{\Delta} X_t + \frac{\theta(B^L)\theta(B)}{\varphi(B^L)\varphi(B)\Delta_L^D \Delta^d} \varepsilon_t . \tag{17}$$

1.4 Innovational outlier

Let us again consider function $f(t)$ in formula (5) and set all coefficients in $\omega(B)$ and $\delta(B)$ equal to zero except for ω_0 and δ_1. This time, we denote ω_0 by symbol ω_I, and suppose $\delta_1 = \phi_1$. Now the form of function $f(t)$ becomes

$$f(t) = \frac{\omega_I}{1-\varphi_1 B} X_t \,, \tag{18}$$

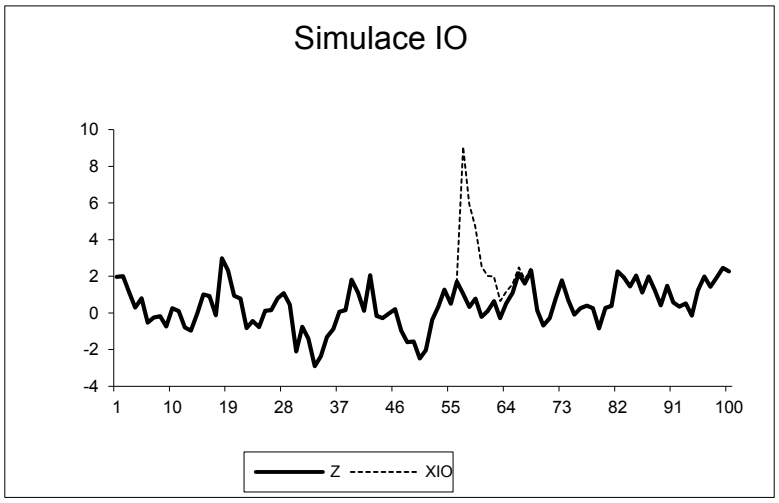
where again

$$X_t = \begin{cases} 1, & t = T \\ 0, & t \neq T \end{cases} . \tag{19}$$

Formula (18) for function $f(t)$ looks like a temporary pulse intervention and represents an example of the Koyck-type model with a rate of decrease equal to $\delta_1 = \phi_1$. This type of outlier is called *innovational*. A time series with an innovational outlier is expressed as follows:

$$Y_t = \frac{\omega_I}{1-\varphi_1 B} X_t + z_t \,. \tag{20}$$

Fig. 3: Original and LS-contaminated time series



Source: Own calculations

What is clear here is that series Y_t is identical with series z_t until time T; then the values of Y_t are shifted by ω_I units at time T; namely, shifted upwards (if $\omega_t > 0$) or downwards (if $\omega_t < 0$); after time T the outlier effect exponentially fades out, and the rate of this decay depends on the value of $\delta_1 = \phi_1$. Figure 3 shows the effects of an innovational outlier.

If we know the time at which the outlier occurs, we can, similarly to the instances considered above, estimate the value of ω_I within the framework of the transition-function model. Using a preliminary approximation, we substitute for z_t in formula (20) from formula (2), obtaining

$$Y_t = \frac{\omega_I}{1-\varphi_1 B} X_t + \frac{1}{1-\varphi_1 B}\varepsilon_t . \tag{21}$$

The first term in formula (21) is the intervention variable, and the second term is the error component governed by an AR(1) model, by which series z_t is also governed. We can again see that it is the classical transition-function model of Koyck type.

Similarly to the instance of additive outlier, we can consider a general ARIMA$(p,d,q)(P,D,Q)_L$ model for z_t, obtaining

$$Y_t = \frac{\theta(B^L)\theta(B)\omega_I}{\varphi(B^L)\varphi(B)\Delta_L^D\Delta^d} X_t + \frac{\theta(B^L)\theta(B)}{\varphi(B^L)\varphi(B)\Delta_L^D\Delta^d}\varepsilon_t . \tag{22}$$

2 Example

The following figure shows the daily time series of the CZK/USD exchange rate. The source of the date is the Czech National Bank (CNB – www.cnb.cz) and the series covers the time period from August 2013 until April 12, 2014. The chart clearly shows the CNB intervention at the beginning of November 2013, when the CNB made a decision to artificially weaken the CZK exchange rate with respect to EUR (and USD too), which was manifested by a large jump divided into several consecutive days.

At first sight it is clear that it is modified outlier type Level Shift – cf. Figure 2. We can see that till to the time point $T = 69$ (November, 6) is the character of time series stable and a point $T = 69$ occurs permanent level shift range of ω_S. We now have to estimate a suitable model for this time series and incorporate the level shift outlier into the model. We use formulas (12) and (13), or (14) a (15). After a number of analyses (study of stationarity, ACF, PACF, EACF, and other identification methods) the following model was identified as the most suitable one:

$$(1-\theta_1)Y_t = (\omega_0 + \omega_1 B) X_t + \varepsilon_t , \tag{23}$$

where Y_t is the CZK/EUR exchange rate time series, X_t is an intervention (binary) variable, and ε_t is normally distributed white noise.

Fig. 4: Rate CZK/USD

Source: Own graph

We obtained final model as

$$Y_t = 0.9801 Y_{t-1} + 0.9626 X_t - 0.9595 X_{t-1} + \varepsilon_t . \quad (24)$$

This model was successfully confirmed at several stages of verification – tests of stationarity and of unit roots, as well as analyses of residua of the input and output series. The quality of this model is very good, with index of determination value 0.967. The series Y_t itself is governed by an AR(1) model and the value of parameter θ_1 is close to one, as could be expected (this model is close to a random walk). The quality of this model is also clear from the autocorrelation and partial autocorrelation functions, whose values do not significantly depart from zero. We can use this model for forecasts building (Soukup 2012).

Tab. 1: ACF and PACF output

```
AUTOCORRELATIONS
  1- 12    -.07 -.05 -.05 -.01 -.11  .00 -.15  .14 -.05  .03  .10  .08
  ST.E.     .07  .08  .08  .08  .08  .08  .08  .08  .08  .08  .08  .08
   Q        1.0  1.4  1.8  1.8  4.1  4.1  8.4 12.0 12.4 12.6 14.4 15.7

 13- 24    -.06 -.15 -.00  .06 -.09 -.00  .09  .06 -.09 -.05 -.01 -.07
  ST.E.     .08  .08  .08  .08  .08  .08  .08  .08  .08  .08  .09  .09
   Q       16.3 20.8 20.8 21.6 23.2 23.2 24.9 25.6 27.1 27.6 27.6 28.6
```

```
PARTIAL AUTOCORRELATIONS
  1- 12    -.07 -.05 -.06 -.02 -.12 -.02 -.13  .10 -.06  .01  .10  .07
  ST.E.     .07  .07  .07  .07  .07  .07  .07  .07  .07  .07  .07  .07

 13- 24    -.01 -.13  .03  .04 -.07  .00  .06  .06 -.12 -.04 -.04 -.11
  ST.E.     .07  .07  .07  .07  .07  .07  .07  .07  .07  .07  .07  .07
```

Source: SCA output

Conclusion

The aim of the article was to describe the basic types of outliers in time series and their integration in the model. Our example describes outlier type of Level Shift. This type implies the permanent level shift time series. As a suitable time series we selected a series of rates CZK/USD. This time series and its noise component is operated by stochastic process (process ARIMA or SARIMA). It is therefore impossible to apply classic regression analysis model. It is necessary to use more advanced techniques, which are dynamic regression models, namely models with transfer function. The model we have created can be used for subsequent analysis or for making predictions. The model quality is very good, its determination index is close to one, and the model successfully passed all stages of verifications. This model will be further used in analyses to follow this paper.

Acknowledgment

This paper was written with the support of the Czech Science Foundation project No. P402/12/G097 „DYME – Dynamic Models in Economics“.

References

Abraham, B., Ledolter, J. (1983). Statistical Methods for Forecasting, John Wiley & Sons, New York.

Anderson, O. D. (1976). Time series analysis and forecasting – Box-Jenkins approach. London, Butterworth.

Bartošová J., Forbelská M. (2012). Modelling of the Risk of Monetary Poverty in the Czech Regions. *In: The 6th International Days of Statistics and Economics. Conference Proceedings. September 13–15, 2012. Prague, Czech Republic*.

Buc D., Klieštik T. (2013). Aspects of Statistics in Terms of Financial Modelling and Risk. *In: The 6th International Days of Statistics and Economics. Conference Proceedings. September 13–15, 2012. Prague, Czech Republic*.

Coufal J. (2012). An Application of Discrete Mathematical Model in Economics. *In: The 7th International Days of Statistics and Economics. Conference Proceedings. September 19–21, 2013. Prague, Czech Republic*.

Granger, C. W. J., Newbold, P. (1986). Forecasting Economic Time Series. Academic Press. New York.

Pankratz, A. (1991). Forecasting with dynamic regression models. John Wiley & Sons inc., New York.

Soukup J. (2012). The Accuracy of Macroeconomic Forecasts. *In: The 6th International Days of Statistics and Economics. Conference Proceedings. September 13–15, 2012. Prague, Czech Republic*.

Stankovičová I., Vlačuha R., Ivančíková L. (2013). Trend Analysis of Monetary Poverty Measures in the Slovak and Czech Republic. *In: The 7th International Days of Statistics and Economics. Conference Proceedings. September 19–21, 2013. Prague, Czech Rep.*

Wei, W. W. S. (1990) Time series analysis – Univariate and multivariate methods. Redwood City, California, Addison-Wesley Publishing Company.

Contact

Luboš Marek

University of Economics, Prague

W. Churchill sq. 4, 130 67 Prague 3, Czech Republic

marek@vse.cz