WP. 32
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (v): Software & tools for data editing and imputation

Screening Methods and Tools for the UNIDO Industrial Statistics (INDSTAT) Databases

Prepared by
Matthias Templ (Statistics Austria & UNIDO Consultant & Vienna University of Technology & data analysis OG) and
Valentin Todorov (UNIDO)

I. INTRODUCTION

1. The United Nations Industrial Statistics Organization (UNIDO) maintains industrial statistics databases (INDSTAT) which contain information on key indicators broken down by industry sectors, countries and years. The databases are built around the International Standard Industries Classification (ISIC), which classifies economic data according to kind of economic activity in the fields of production, employment, gross domestic product and other statistical areas. While the ISIC classification covers all areas of economic activity, the UNIDO databases only relate to the it's manufacturing, mining and utilities sections. The first major source for data collection are industry data reported by the national statistical offices through a country questionnaires. Data for OECD member countries are collected in the form of a joint OECD/UNIDO questionnaire and transmitted to UNIDO. Next, National publications for industrial censuses, annual surveys and input-output tables as well as international sources, both published and unpublished, are used.

2. Although the data do not come directly from the field, a screening for erroneous and/or inconsistency values prior updating the databases is very important. The currently used screening methods are described and new methods are proposed. The methods discussed vary from logical relationships in the data (balanced edits, e.g. wages and salaries cannot be negative, number of female employees cannot be larger than number of employees, etc.), to highly sophisticated outlier detection procedures. The theoretical background of these new procedures is considered, their implementation in the programming environment R is presented and the possibilities for their introduction in the statistical production process are investigated. The multivariate aspect of the data collected in surveys makes the task of outlier identification particularly challenging.

3. UNIDO statisticians take considerable care in ensuring data consistency in the process of enlarging the INDSTAT and in improving the international comparability of its content. However, due to inconsistency inherent to many series reported by primary sources as well as to the wide variety of sources used, it is felt that a final screening of the data is needed. The purpose of this

final screening is to diagnose and display 'abnormal' entries in the database, to allow for possible corrections. The final screening takes place in two phases. First, possible abnormalities are identified through a computerized procedure. Second, UNIDO statisticians redress, to the extent possible, the identified abnormalities. Each of the following four variables in the database:

(1) Establishments (EST)
(2) Employment (E)
    (a) Number of persons engaged (PER)
    (b) Number of employees (EMP)
    (c) Number of female employees (FEM)
(3) Wages and salaries (WS) and
(4) Output (GO)
(5) Value added (VA)

per country, ISIC category, combinations of ISIC categories, and total manufacturing and year are considered simultaneously. An observation (one variable) or a combination of two observations in the form of a ratio of two variables pertaining to the same country, ISIC category (or combination of ISIC categories) and year is considered to be abnormal if it appears to be implausible on logical, statistical or economic grounds.

4. The criteria consist of acceptable ranges for (1) to (5) above. Other than purely logical ones, these ranges were set by screening a sample of countries having dissimilar economies, data collecting and reporting procedures. Some of the acceptable ranges are allowed to take different values depending on the degree of specialization and volatility of the manufacturing sector in a country.

The abnormality may stem from one or more of the following:

(1) An outright mistake, e.g., a typo. In the current statistical development environment this type of errors is not possible since the data are validated at the phase of initial loading or modification.
(2) A problem related to definitions and/or methods used in collecting and processing data, or changes in those definitions or methods over time;
(3) Actual extraordinary economic circumstances.

Only a fraction of the tests unambiguously point to a mistake. In all other cases the diagnosed abnormality may stem from any combination of the three problems stated above.

5. In the following sections, different screening methods are introduced. In Section II/A traditional editing rules are briefly discussed. Section II/B continuous by giving an outline how deviations of parameters of a time series can be formalized. Here, deviations from initial values, target values and deviations from the linear and non-linear trend of time series are expressed. A slightly different approach is presented in Section II/C where the time series is firstly filtered before the outlier detection is done. Another approach is presented in Section II/D where modeling of the time series is in focus. Here, the time series are transformed to white noise before outliers are detected. After the time series are white noise, methods from the previous chapters may be applied. The last section on methods, Section II/E outlines three approaches based on regression analysis, whereas two of them are widely used in Official Statistics - the additive and the innovational (outlier) model, implemented in the X12 and X13 free software of the US Bureau of Statistics [Time Series Staff, 2009], see also [see also Kowarik et al., 2012].

II. SCREENING METHODS

A. Balanced and Ratio Edits

6. Screening rules, at National Statistical Institutes often called *edit rules* or *edits* for short, are often used to determine whether a record is consistent or not. An example of an edit is PER = FEM + MALE , where PER is the number of employees, FEM and MALE the number of female and male employees. This edit expresses that the male and female employees should sum up to the total number of employees. Such an edit is referred to as a *balanced edit*. A *non-negative edit* is just defined that a value has to be non-negative if it passes the edit. For example, WS (Wages and Salaries) should not be negative. If a ratio of two variables should be less (or greater) than a certain threshold, then this edit is referred as *ratio edit*. For example, VA/GO ≤ 100. In order to construct a set of edits one usually starts with the *hard* (or logical) edits, which hold true for all correctly observed values. Balance edits are usually referred as *hard edits*. After the hard edits have been specified, one generally uses subject-matter knowledge and statistical analysis to add a number of *soft* edits, which hold true for a high fraction of correctly observed records but not necessarily for all of them [de Waal, 2008].

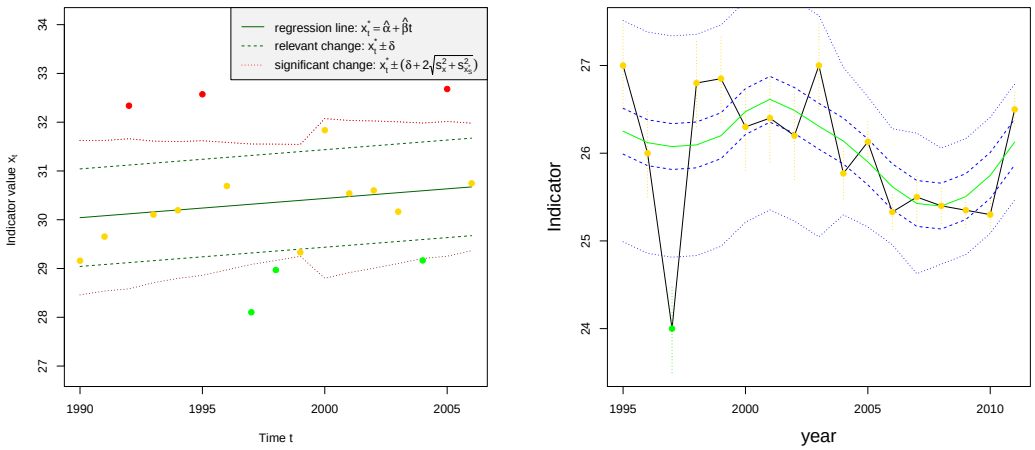
7. Ratio edits can be either hard or soft edits. The threshold related to ratio edits have to be determined carefully, so that on the one hand only few values may violate the edit and that on the other hand erroneous values are detected by these edits. This threshold is either fixed by a subject matter specialist (hard edit) or may vary depending on the input data (soft edit). To avoid over-editing one should in particular be careful not to specify too many soft edits. In general, users tend to apply more soft edits than necessary to the data.

B. Deviation from Pre-defined Parameters

8. Data analysts, policy makers or subject matter specialists may have knowledge about the trend and behaviour of their time series. In Templ et al. [2011] the detailed description is available that we applied within the screening of data with prior knowledge of the behaviour of the time series. Here, within so called **evaluation plots** [Hulliger and Lussmann, 2008, Templ et al., 2011], the deviation of pre-defined parameters, deviation of target values, or any other defined time-depended observations and the deviation of pre-defined trends is formally described.
For the evaluation of the indicator a reference value δ (the two dashed lines in Figure 1(a)) is needed, which defines whether a change is relevant or not. In addition, the uncertainty of x_S^* and the uncertainty of each data point x is expressed by their variances $s^2_{\delta(S)} = s^2_x + s^2_{x^*_S}$. The relevant deviation δ can also be seen as the real variability of the phenomena that is observed. The standard deviation s_δ could be seen as a measurement of variability, which will be added to the real variability and the measurement variability may change over time. Please see the exact mathematical description of the problem in Templ et al. [2011].
Not for every indicator or every region a course or target value is given. In order to get at least a better picture of the indicator and its trend a linear regression is applied. One easy possibility is to estimate the trend $\hat{\beta}$ and the intercept $\hat{\alpha}$ of an indicator by least squares estimation. Let $\hat{\beta}$ be the trend estimated by minimizing the sum of squared distances between the observed indicator x_t and the regression line x_t^*, this distances are called residuals.

$$x_t = \hat{\alpha} + t \cdot \hat{\beta} + \varepsilon, \qquad t_1 \leq t \leq t_N \tag{1}$$

where ε is a random variable (errors) which accounts for the discrepancy between the actually observed responses x_t and the predicted outcomes $\hat{\alpha} + t \cdot \hat{\beta}$. The evaluation is calculated as follows [for exact



(a) Evaluation of an indicator for a pre-defined trend.

(b) Evaluation of an indicator using Loess.

FIGURE 1. Evaluationplots. The points highlighted outside the outer bands in (a) or (b) are detected as outliers following not the trend.

mathematical description, see Templ et al., 2011]

$$Eval(x_t, \delta, s_\delta^2) = \begin{cases} -1 & x_t > \hat{\alpha} + t \cdot \hat{\beta} + \delta + 2 \cdot s_\delta^2 \\ 1 & x_t < \hat{\alpha} + t \cdot \hat{\beta} - \delta - 2 \cdot s_\delta^2 \\ 0 & \text{else} \end{cases} \tag{2}$$

Additional information in the default plot is represented as the dotted lines and the different colors of the values of the time series. The dotted lines determine the region where the indicator will be called 'neutral', it is the region around the line reaching from the first value to the last value in our time series. By default this marks a 1% margin of the median of the indicator around the course. Outside this corridor we call an indicator 'bad' if it has a higher value or 'good' if it has a lower value.

9. The mentioned approach in the last paragraph has one great disadvantage, its non-robustness. We therefore applied a regression method which is not sensitive to the presence of outliers. For this reason an MM-estimator is chosen. See Yohai [1987] for further details on robust regression.

B.1. *Considering a Non-linear Trend of an Indicator.*

10. For a nonlinear regression the LOESS model may be used. LOESS (locally weighted scatterplot smoothing) was proposed by Cleveland [1979] as a model which uses locally weighted polynomial regression. For every point in the data a polynomial of low degree is fit to a subset of the data. We are going not into details about this method, but show one result in Figure 1(b), whereas one outlier is detected.

C. Detection of Outliers After Filtering the Time Series

11. The motivation of the following section comes from online monitoring of time series with applications to low frequency heart rates [Fried, 2004, Gather and Fried, 2007, Fried, 2004] where new measurements are taken online of which is the aim to decide if such a new measurement is an

measurement error or not. This is similar to UNIDO INDSTAT2 database where new values come into the database. The aim is to robustly filter a time series by a given window size. The idea is to move a window through the time series where in the window the time series are filtered. The trend in the current time window is extrapolated to the next point in time and the residual of the incoming observation is standardized by the current scale estimate [see, e.g., Fried and Gather, 2007].
If more than half of the residuals in the right part of the window are larger than σ_t, a shift is detected. A more detailed description of the filter can be found in Fried [2004]. The adaption of the window width is described by Gather and Fried [2007] and Fried [2004], for shift detection we refer to Fried and Gather [2007]. Figure 2 shows an application of the robust filtering approach with

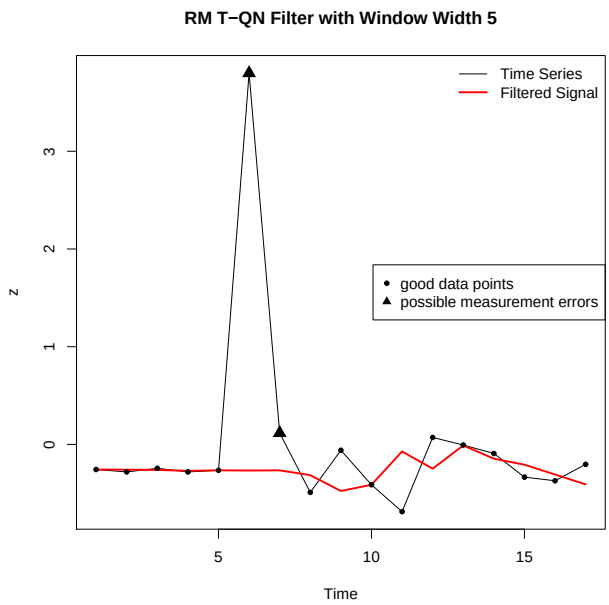


FIGURE 2. Establishments in ISIC 369. Repeated median filtered time series (window width equals 5) and outliers detected.

the standard parameters (the window width was set to 5) of function `robust.filter` of package `robfilter`. Possible measurement errors are highlighted - here with a triangular symbol.

D. Outlier Identification Based on De-Trended and White Noise Time series

To determine if a time series include outliers it is often necessary to remove the trend and to make the time series stationary.

12. In order to start with the outlier detection, **systematic or deterministic effects over time have to be removed**, i.e. first the time series has to be transformed to *white noise*. White noise is a random process, which has mean zero, a constant variance over time, and covariance equals zero between different time points. To transform the time series to white noise, different methods can be used. The most common used models for short time series are the ARIMA(1,0,0) and the ARIMA(0,d,0) model [for details about ARIMA models have a look, e.g., in Box et al., 2008]. When non-stationary behaviour is expected for the time series, differencing the time series by degree d may induce stationarity, i.e. to apply an ARIMA(0,d,0) model. This is also known as *de-trending*. In the following Δ denotes the differencing operator, i.e. $\Delta(x_t) = x_{t+1} - x_t$ for $t = 1, \ldots, T-1$. Hence Δ^d is the differencing operator applied d times on the time series x_t, i.e. $x_{t_d} = \Delta^d X_t$. In practice, d is

usually 0, 1 or 2. In general, this approach is common to remove deterministic components of time series. An ARIMA(1,0,0) model is also often used to transform the time series to white noise. An ARIMA(1,0,0) equals an autoregression of order 1 (AR(1)). This means that each point of the time series is dependent on the previous. The third option is to apply an ARIMA(1,d,0) model. This equals to apply an AR(1) model to the de-trended data.
To transform the series to white noise the following steps are necessary:

(1) First an *ARIMA* (autoregressive integrated moving average) model is fit to the time series.
(2) Transformation of the correlated input series x_t to the uncorrelated white noise series, which consists in fact of the residuals of the fitted time series.
(3) Apply the outlier detection on the corresponding residuals.

This is clearly a univariate approach which do not consider the multivariate cross-correlation between time series that may give additional power to detect outliers. Here the time series have to be pre-whiten - we refer to Box et al. [2008], Templ [2012] for the therory behind and applications.

13. There exists a never ending discussion of some researchers about the kind of model used for transformation of the time series to white noise. In the following, we therefore show the effect of the two most applicable models to short time series, the ARIMA(0,1,0) and the ARIMA(1,0,0) model.

14. The time series visualised in Figure 3 (see the graphic in the lower right for the original values) shows a clear trend. Depending on the method used for detecting the outliers, the trend may to be removed. The standardized residuals of the modeled time series of Wages and Salaries in Austria, ISIC code 369 (Manufacturing n.e.c.) are shown in the first three graphics in Figure 3. The outlier in 1996 is clearly visible in all graphics, but the value in year 1997 is best modeled by ARIMA(1,0,0). However, it is clearly visible that the residuals of the ARIMA(1,0,0) model are not stationary and white noise. Accounting for this, ARIMA(1,1,0) and it's robust counterpart seems to be the best choice and the behaviour of the values in the original time series is best reflected one of this two models. However, some problems remain. For example, level shifts can hardly be detected

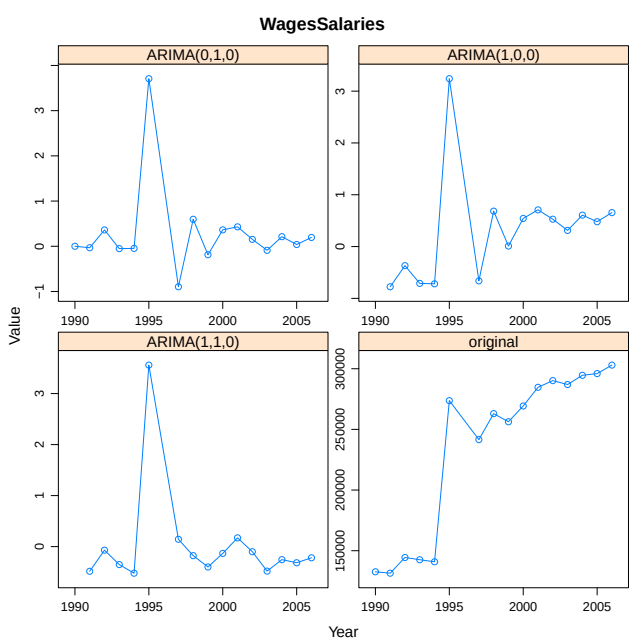


FIGURE 3. Wages and Salaries. Original and modelled time series for Austria at 3-digits level 369 (Manufacturing n.e.c.).

with that approach. For this problems we refer to the detection of outliers with innovation models, see Section F.

15. An observation declared as outlier when the residuals fitted from the autoregressive model, ARIMA(1,1,0) and robust ARIMA(1,1,0), are larger than pre-defined threshold. Two different kinds of thresholds are proposed:

fixed threshold:: the standard model assumes that the residuals are normally distributed and therefore each observation with standardized residual greater than 2.5 could be determined as an outlier.

variable threshold:: each observation where the absolute value of the standardized residual is greater than $\tilde{x} + c \cdot 1.4826 \cdot MAD(x)$ with $\tilde{x}$ the median of the standardized residuals of the model, $MAD(x) = med|x_i - med(x)|$ the median absolute deviation, 1.4826 a scale factor for consistency and c a constant usually set to 2.5.

E. Outlier Detection By Dummy Regression Models

16. The standard procedure for outlier detection in time series implemented in almost all statistical software tools is based on regression the time series to candidate regression variables. These regression variables differs is case of *additive outliers* (AO) and *level shifts* which are often named *innovational* outliers (IO) in this context. An additive outlier concern to a measurement error whereas an innovential outlier represents an extraordinary shock at time point t influencing the values at time points $t+1, \ldots, T$.

17. The differential model: Despite of the given models in literature, we define a simplified model which are not based on complex model estimation, which is not very applicable when dealing with short time series. We define the model as follows:

$$y_{t_d} = \beta\zeta_t + \epsilon_t \quad , \tag{3}$$

where y_{t_d} is the vector of differentiated values $y_{t_d} = \Delta^d y_t = y_{t+1} - y_t$ for $t = 1, \ldots T-1$. β are the regression coefficients and

$$\zeta_t = \begin{cases} 1, & \text{if } t = T \\ 0, & \text{otherwise} \end{cases} \quad .$$

F. The Additive and Innovational Outlier Model:

18. The most popular model in literature which is also implemented in various statistical software tools are the additive outlier model and the innovational outlier model. Let x_t be a stochastic process following an autoregressive integrated moving average (ARIMA) model [see also Box et al., 2008] that is [see also Chang et al., 1988],

$$\phi(B)\alpha(B)x_t = \theta(B)a_t \quad , \tag{4}$$

with B the backshift operator such that $Bx_t = x_{t-1}$; $\phi B = (1 - \phi_1 B - \ldots - \phi_p B^p)$ and $\theta(B) = (1 - \theta_1 B - \ldots - \theta_q B^q)$ are two polynomials in B; $\alpha(B) = (1-B)_1^d(1-B^s)_2^d; d = d_1 + sd_2$; and $\{a_t\}$ are $\sim N(0, \sigma_a^2)$. Then the model for an IO is

$$z_t = x_t + \frac{\theta(B)}{\phi(B)\alpha(B)}\omega\zeta_t \quad , \tag{5}$$

and for the AO the model is given by

$$z_t = x_t + \omega\zeta_t \quad , \tag{6}$$

where ω describes the impact on the least squares fit and has to be estimated [we refer to Chang et al., 1988].

G. Detecting Level Shifts

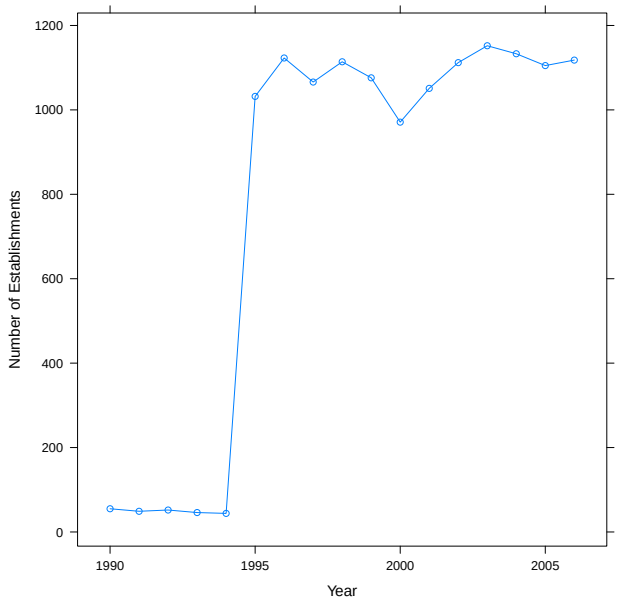


FIGURE 4. Original time series of Establishments in ISIC code 369 (Austrian data) including one clearly visible level shift.

19. Table 1 shows which outlier detection method has correctly classified an the level shift in Figure 4 in an example time series of Establishments in ISIC code 369. As expected, the existence of a level shift is well recognised by the innovational model applied on residuals for either the original time series, the de-trended time series and for the ARIMA(1,0,0) residuals. However, it is interesting to see that for ARIMA(1,1,0) the level shift is not detected. This tells us that level shifts might be only searched in original or de-trended time series, since a model fit might be heavily influenced from the level shift itself.

TABLE 1. Summary of detectingthe level shift in Figure 5.

Method	Original TS	de-trended TS	ARIMA(1,0,0)	ARIMA(1,1,0)
search for level shifts				
innovational model	✔	✔	✔	

H. Robustification of the approaches:

20. The standard estimation of the model in Equation 4, which is input to Equation 6 and 5 is already influenced by outliers. Because of the underlying short time series (17 observations at the maximum), an AR process is applicable while an full ARIMA is not (even there is no seasonal component in our time series). The parameter(s) or the AR process can be estimated robustly by applying an GM-estimator [Martin and Zeh, 1980] using bounded-influence autoregression via iteratively reweighted least squares.

III. Proposed Methodology

21. Section III/A starts with some hard ratio editing rules to flag inconsistencies. Section III/B.1, B.2 and B.3 corresponding to soft ratio editing methods which were implemented in the past for INDSTAT2 Rev. 3 data at two digit level. Section III/B.4 shows the previously described rules applied on 4-digit level data as well as an modification of that rule. All these rules were introduced in the first subsections of Section II. Section III/C highlights the application of detection of erroneous data points for residuals of fitted data points of the Austrian data, while Section III/D shows the application of robust filtering (introduced in Section II/C) to the same data set. Finally, Section III/E shows the results of dummy regression methods introduced in Section II/E.

A. Method 1: Hard Balanced Edits

22. The first three editing rules can be characterized as hard balanced edits described in Section II/A.

Edit 1 *[FWSWA] If* $WSVA = 100 \cdot WS/VA > 100$ *the row is marked by setting FWSWA to "<==" In fact, that are all units for which* $WS > VA$.

23. Edit 1 tests for inconsistencies in *Wages and Salaries* and *Value Added*, i.e. the value added should be higher than the wages and salaries. For the example data, 5 values fails for this edit.

Edit 2 *[FVAGO] If* $VAGO = 100 \cdot VA/GO > 100$ *the row is marked by setting FVAGO to "<==" In fact, that are all units for which* $VA > GO$.

24. Edit 2 is similar to Edit 1. It tests for inconsistencies in *Value Added* and *Gross Output*, i.e. the gross output should be higher than the value added. For the example data, all values pass edit 2.

Edit 3 *[FC] The row is marked by setting FC to "<==" if one of the following variables is equal to 0, but at least one of the others in the list is not missing and not 0: EST, PER, EMP, FEM, WS, GO, VA, GFCF. I.e. if one of them is 0, all must be 0 or missing.*

25. Edit 3 is also a hard balanced edit. The assumption is that when one value is reported then the other values should be as well available or vice versa, if one value is zero no other value should be reported. For the example data, all records pass this check.

B. Method 2: Soft Ratio Edits

B.1. *VA/GO Index.*

26. The following five editing rules corresponds to the class of soft ratio edits (see Section II/B), especially the deviation from initial values (see Section II/B). Ratios are calculated and checked if these ratios are below a certain threshold. In the following Δ denotes the differencing operator, i.e. $\Delta(x_t) = x_{t+1} - x_t$ for $t = 1, \ldots, T-1$. One editing rule which was implemented in ISIC revision 2 was defined as follows. For each country and ISIC code, the values of the time series are transformed in the following manner:

$$e_{(4)} = \Delta e^{VA/GO \cdot 10} \quad . \tag{7}$$

Edit 4 *[VA/GO index] For each value with necessary condition* $VA/GO > 0$ *a value is declared as possible measurement error when* $e_{(4)} > 30$.

This threshold has been fixed by subject matter specialists in order to fulfill both, to flag only few values and to flag the erroneous values.

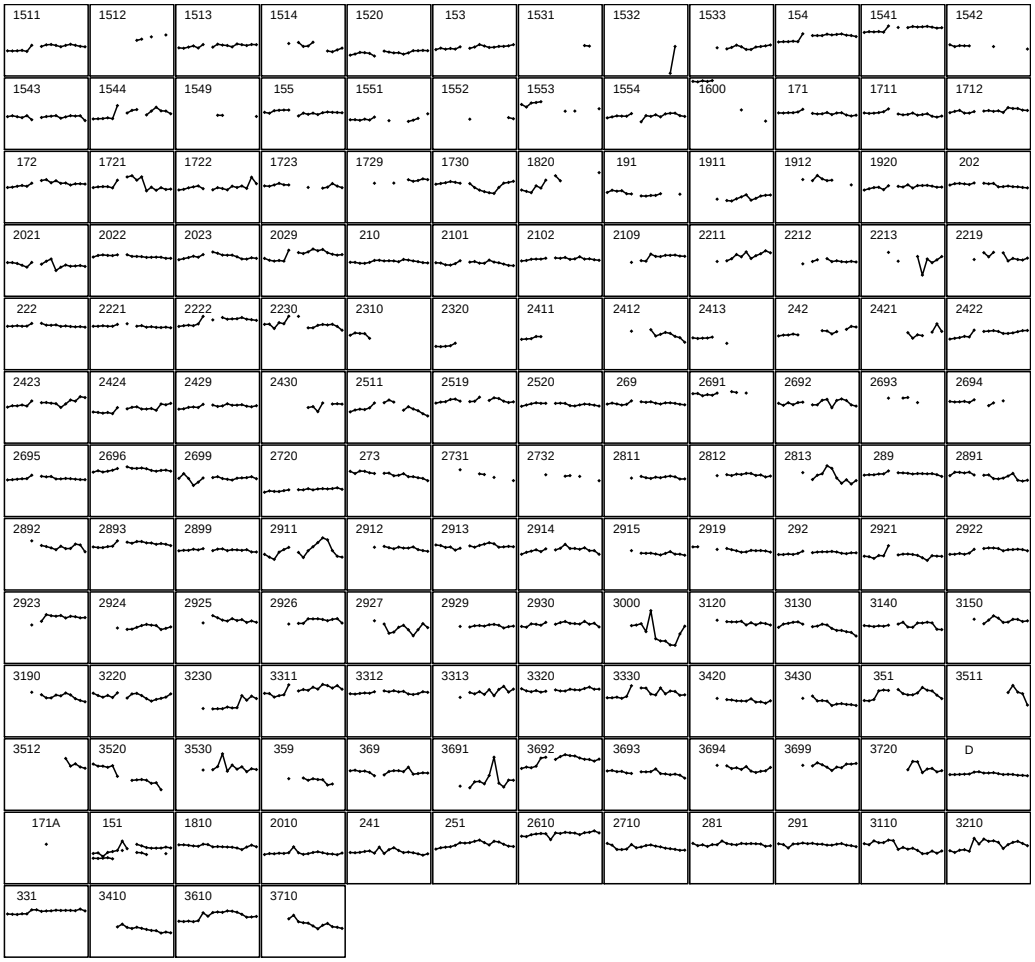


FIGURE 5. Time Series of VA/GO of Austria.

27. Figure 5 shows the VA/GO ratio's of Austria from 1990 until 2006. Figure 6 shows the transformed time series as given in Equation 7. The upper line corresponds to the threshold given (30). Comparing this Figure 5 with Figure 6 it is easy to see that the possible erroneous values are somehow weighted by the transformation, i.e. possible outliers are clearly visible as outliers in Figure 6. This can be seen, for example, by looking at ISIC 3000 or 3691. While this method seems to work mostly well, however, one drawback of this editing rule is that some values are declared as outliers because the whole branch have constantly high ratios. For example, values in ISIC 2519 or 273 do not seems to be erroneous when looking at Figure 5, while some of the corresponding values are detected as outliers in Figure 6.

B.2. *Indices Based on WS/EMP.*

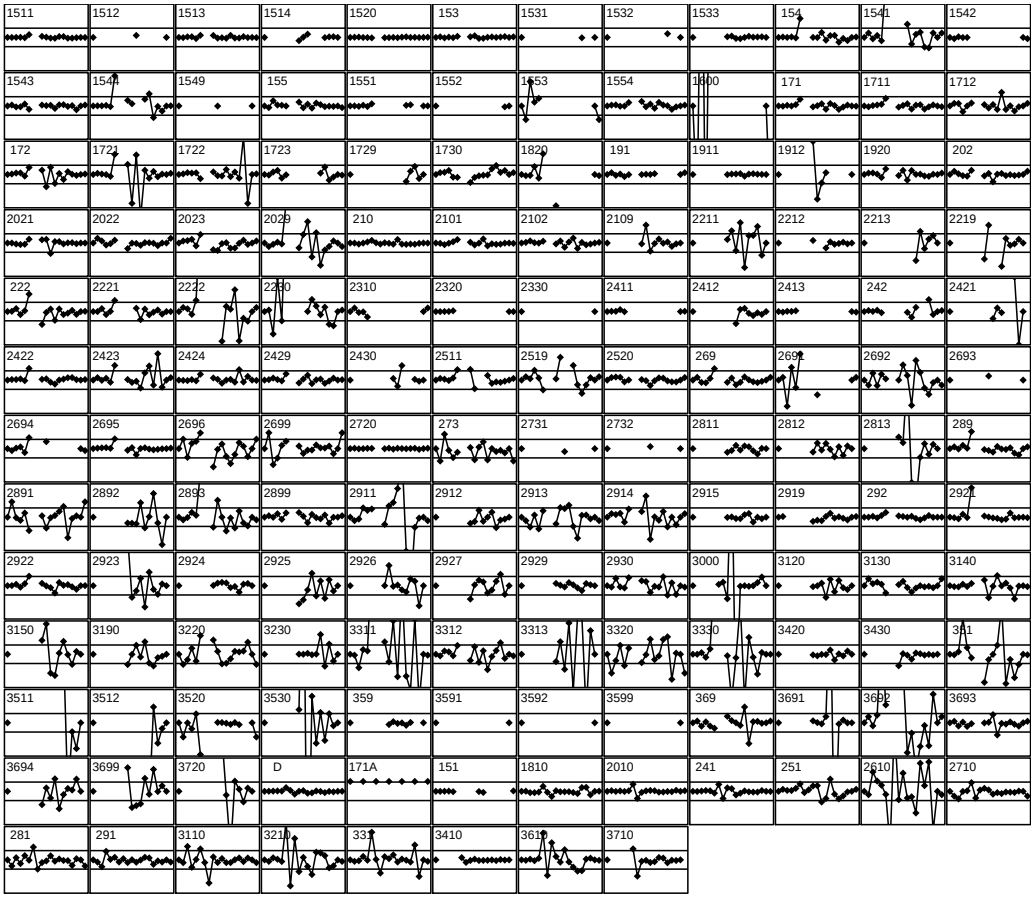


FIGURE 6. Threshold rule on transformed time series of VA/GO given in Equation 7 and Edit 4. Values outside the inner band are possible outliers.

28. For each year and country calculate the ratio

$$y_{ikt} = \frac{WS_{ikt}}{EMP_{ikt}} \tag{8}$$

for all ISIC categories with

- $i = 1, \ldots, M$ refers to the M different ISIC codes
- $k = 1, \ldots, K$ refers to the K countries
- $t = 1, \ldots, T$ refers to the years where the time series is observed and with T being the current time

$$WE_{ikt} = y_{ikt} - \frac{1}{n_o} \sum_{j=1}^{T} y_{.kj} \quad , \tag{9}$$

where "." means that all values of this index are chosen and with n_o the number of observed values of $y_{.kt}$. Equation 9 calculates for each year (and country) the WE index for all ISIC categories. Figure 7 shows the WE index (left upper graphic) and the scaled version (graphic on the right upper hand

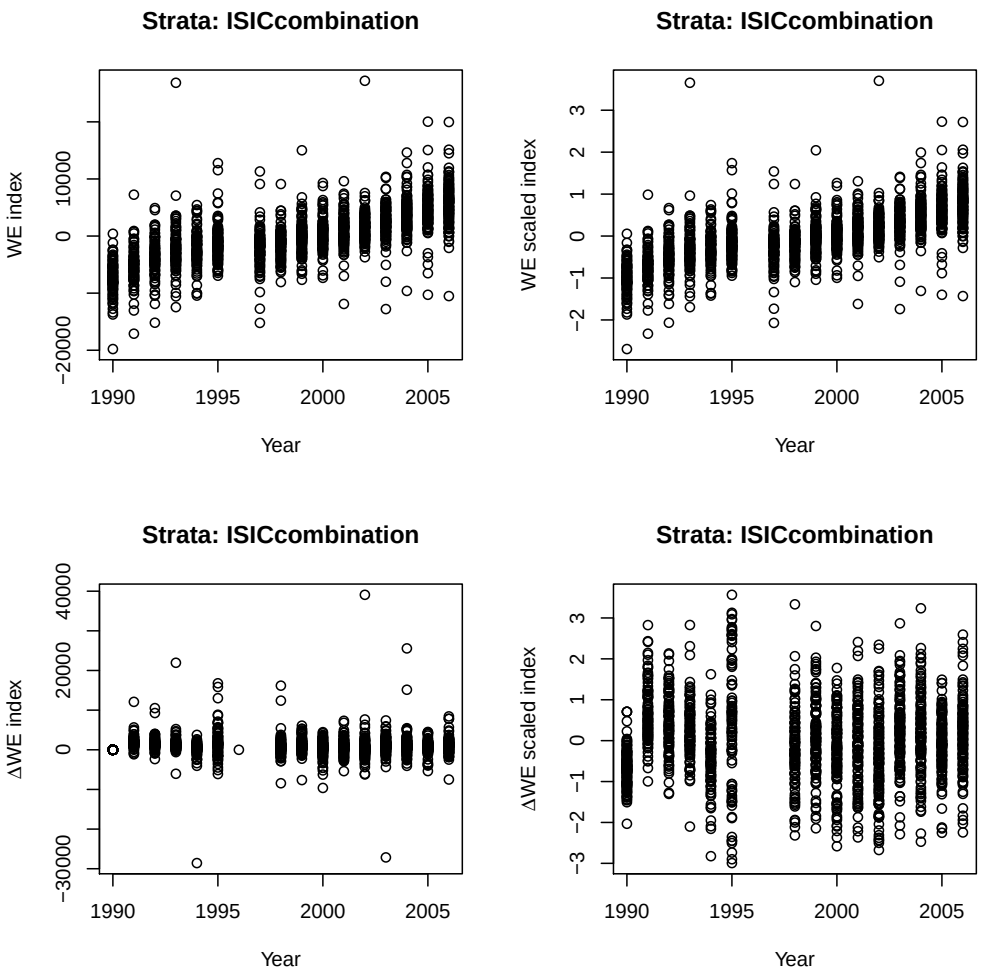


FIGURE 7. WE and Δ WE index.

side), the differenced values of WE, Δ WE (bottom left) and its scaled version (bottom right). Note that the WE index do not consider the trend of the data (see Figure 7) because in Equation 9 the mean is calculated using the information of all years. For ISIC revision 2 many edits were constructed based on the WE and ΔWE index in combination with the coefficient of variation. Few useful editing rules are now reformulated.

Edit 5 *[WE] Flag the item if the absolute value of the scaled WE index is* > 2.5 ***and*** *the absolute value of* Δ *WE scaled index is* > 0.25.

Edit 5 ensures that a value is flaged as erroneous when WE is large and the change over time is large. Originally, at ISIC revision 2 the coefficient of variation was considered as well and both thresholds were changed when the coefficient of variation is above a fixed threshold. Note that such fixed level shifts for the threshold does not make any sense. In spite of that, Edit 5 is based on scaled values and a threshold on such scaled values seems to be more natural to choose. See also Section II/B. For Austria, two observations fail on this edit, namely:

```
     Year ISIC ISICcombination EST EMP       WS
590  1993  351             351   5 376 20979121
1875 2002 2213            2213  51  55  2700000
```

Looking more closely at one of these ISIC categories, namely on ISIC 2213, one could see that one possible erroneous data point is detected by this edit (Figure 8). The regions in grey mark edit rule 5. If a point is for both Figures in the grey area, the point is flagged as erroneous. Another reasonable

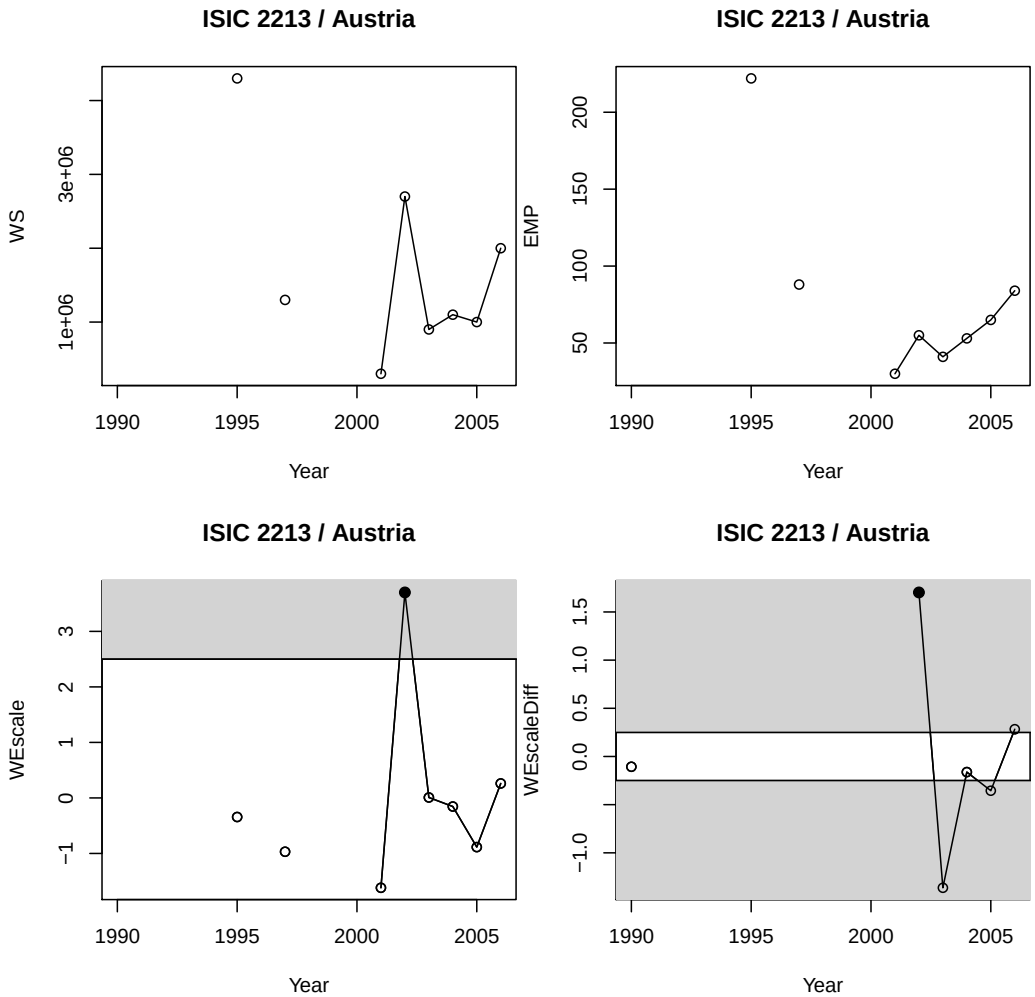


FIGURE 8. Original time series of ISIC 2213 of Austria and their corresponding scaled WE and scaled Δ WE indices.

index is to calculate WS/EMP for each year (and country) to its mean value whereas the mean value is based on every year and not for ISIC categories as in Edit 5, i.e. the stratification of the data is now based on years and not on ISIC categories.

$$WE_{ikt}^{(years)} = y_{ikt} - \frac{1}{M}\sum_{j=1}^{M} y_{jkt} \quad , \tag{10}$$

where "." means that all values of this index are chosen and with M the number of ISIC categories. Equation 10 calculates for each ISIC category (and country) the WE index for all years separately.

Edit 6 *[WE (years)] Flag the item if the absolute value of the scaled $WE^{(years)}$ index is > 2.5* ***and*** *the absolute value of the scaled Δ $WE^{(years)}$ index is > 0.25.*

Since the index is calculated for each year, there is no trend problem (see Figure 7) in Edit 5. This edit is again of type deviation from initial values in Section II/B and used de-trending as formalized in Section II/D. Using the data for Austria the ISIC category 2213 fails on this edit while ISIC category 351 passes the edit.

B.3. *VA/EMP Index.*

29. For each year and country calculate the ratio

$$y_{ikt} = \frac{VA_{ikt}}{EMP_{ikt}} \tag{11}$$

for all ISIC categories with

- $i = 1, \ldots, M$ refers to the M different ISIC codes
- $k = 1, \ldots, K$ refers to the K countries
- $t = 1, \ldots, T$ refers to the years where the time series is observed and with T beeing the current time

$$VE_{ikt} = y_{ikt} - \frac{1}{n_o} \sum_{j=1}^{T} y_{.kj} \quad , \tag{12}$$

where "." means that all values of this index are chosen and with n_o the number of observed values of $y_{.kt}$. Equation 12 calculates for each year (and country) the VE index for all ISIC categories.

Edit 7 *[VE] Flag the item if the absolute value of the scaled VE index is* > 2.5 ***and*** *the absolute value of* Δ *VE scaled index is* > 0.25.

B.4. *EMP Index.*

30. Let Δ EMP the differ Let $\Delta(x_i kt) = x_{ik(t+1)} - x_{ikt}$ for $t = 1, \ldots, T-1$ the differenced time series of EMP for each country and ISIC category.

31. The edits for INDSTAT2, Rev 3 data was based on threshold of differences based on the largeness of a current value, i.e. the differences between two adjacent years could be higher when one of the values are already large. If the values are low, also the differences in years should be low. This was categorised in 5-6 steps, for example, if one value is larger than 10000 the other value (adjacent years) has to be larger than 1000 ($\Delta x \leq -9000$), but if one value is 0 the other one does not be larger than 1000 ($\Delta x \leq 1000$). This leads to a curious rejection region (where points are flagged as possible errors) as visible in Figure 9 (grey area). The following new rule smooths the boarders of the rejection region.

Edit 8 *[EMP] For each time series (year* $\times$ *ISIC category* $\times$ *country), flag the item if*

$$\begin{cases} E_0 < 0 & \\ E_1 > 1000 & \text{if } E_0 = 0 \\ E_1 > 10000 + E_0\sqrt{E_0}/10 & \text{if } E_0 \in (0, 1000] \\ E_1 > 10000 + \sqrt{1000} * 10 + \sqrt{x} \cdot 150 & \text{if } E_0 \in (1000, 10000] \\ E_1 > \sqrt{E_0} * 150 + E_0 \cdot 1.1 & \text{if } E_0 \geq 10000 \end{cases}$$

and vice versa with E_0 *and* E_1 *interchanged.*

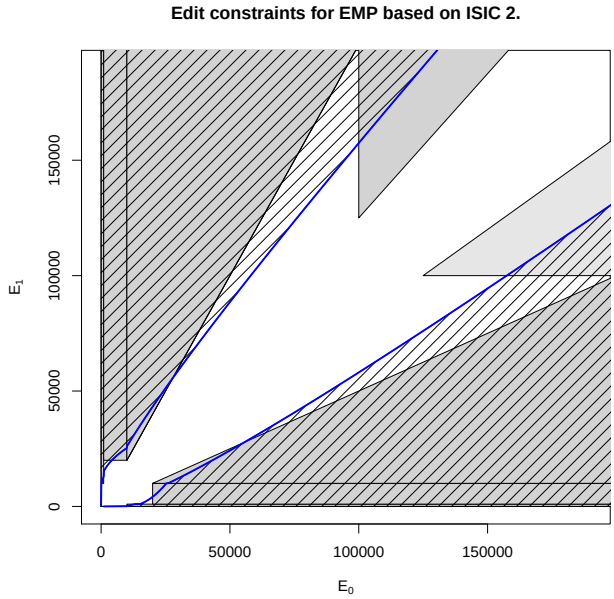


FIGURE 9. Newly proposed edits (shaded area) versus deterministic edits from the formerly implemented rules for INDSTAT2, Rev. 3 (grey area).

C. An Editing Rule Based on Residuals of Modelled Time Series

32. The following editing rule evaluates the residuals received from a fitted time series. The theory is described in Section II/D.

Edit 9 *[White Noise] For each country and ISIC category. If a time serie contains less than 2 missing values:*

(1) *De-trend the time series by differencing, i.e.* $\Delta(x_t) = x_{t+1} - x_t$ *for* $t = 1, \ldots, T-1$
(2) *Fit an AR(1) model to the de-trended time serie*
(3) *Scale the residuals by diving them with the standard deviation of the residuals*
(4) *Flag a value of the time serie if the absolute value of the residual is larger than* 2.5.

If the time serie contains more than 2 missing values skip the second step.

Figure 10 shows the outlier detection using de-trending and scaling of the time series VA/GO for all ISIC combination categories of Austria. Points in red are possible erronous values.

D. Robust Filtering

33. The following editing rule corresponds to filtering methods described in Section II/C.

Edit 10 *[robust filtering] Outlier declaration by function* `robust.filter` *with repeated median filter using window width 9 for any time series with more than 14 observed values. Remaining missing values have to be interpolated beforehand.*

Figure 11 shows the result of repeated median filters applied on each time series (see also Section II/C) of Austria. Reducing the window width, the more values are declared as outliers. When comparing the results to Figure 5 the outlier detection using robust filtering performs well. For example, outliers in

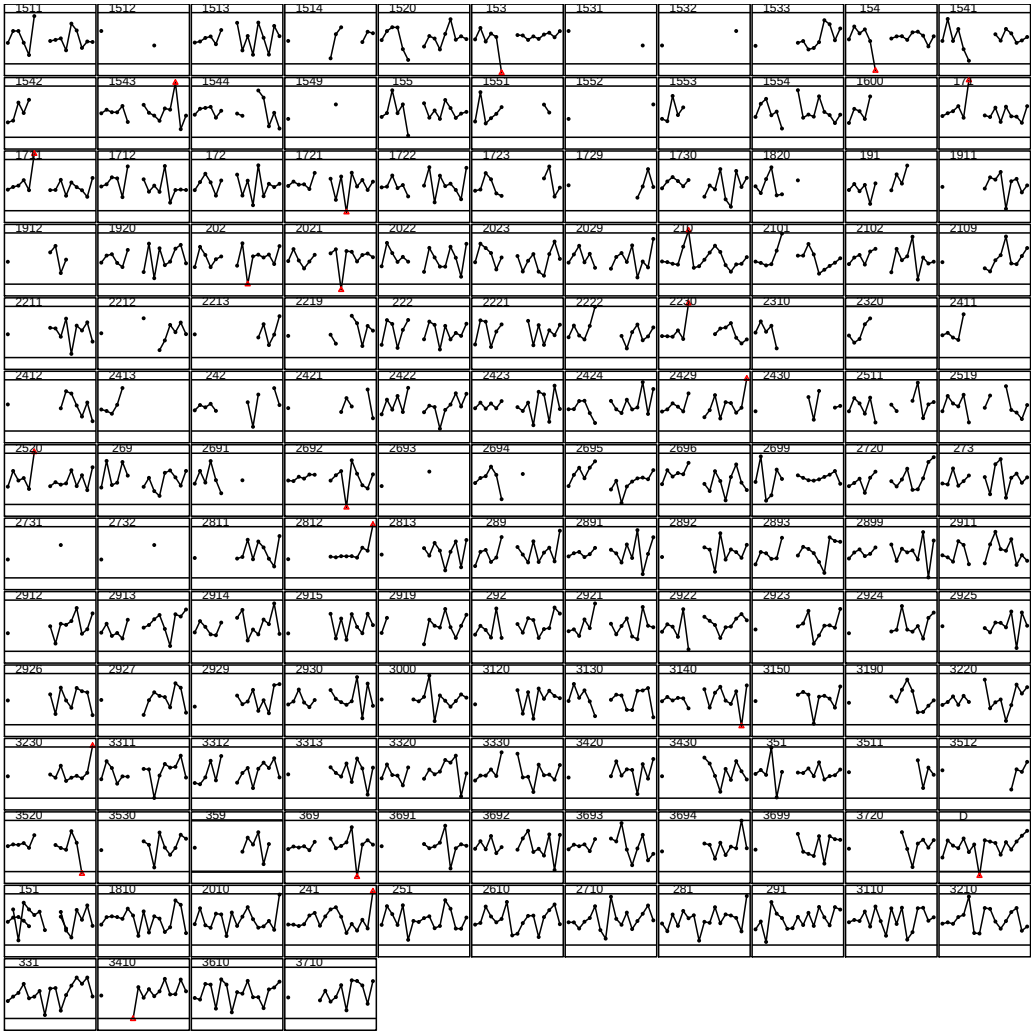


FIGURE 10. De-trended and scaled time series of VA/GO for each 3-digit and 4-digit ISIC code.

ISIC 1541 and 3330 are correctly identified. However, some of them are not recognized (for example, the outlier in ISIC 3000) and some values may declared as outliers even they seems do not behave strangely (see, for example, ISIC 2029). The recommendation is to only apply this method when the time series includes almost no missing values or to skip the robust filtering approach for outlier detection. At least, further research is necessary to adapt these kind of methods to INDSTAT data.

E. Advanced Soft Editing Rules Based on Regression

34. Figure 12 shows the result of the error detection using the differential model (see also Section II/E). All extreme values are detected. In comparison to the additive outlier model in Figure 13 less observations are marked as outliers. For example, in ISIC 2022 no outliers are detected by the differential approach but are marked as outliers with the additive outlier model. For this reason, the application of the differential model is the more conservative choice while the additive outlier model tend to mark also few non-outliers as outliers. This leads to the last editing rules:

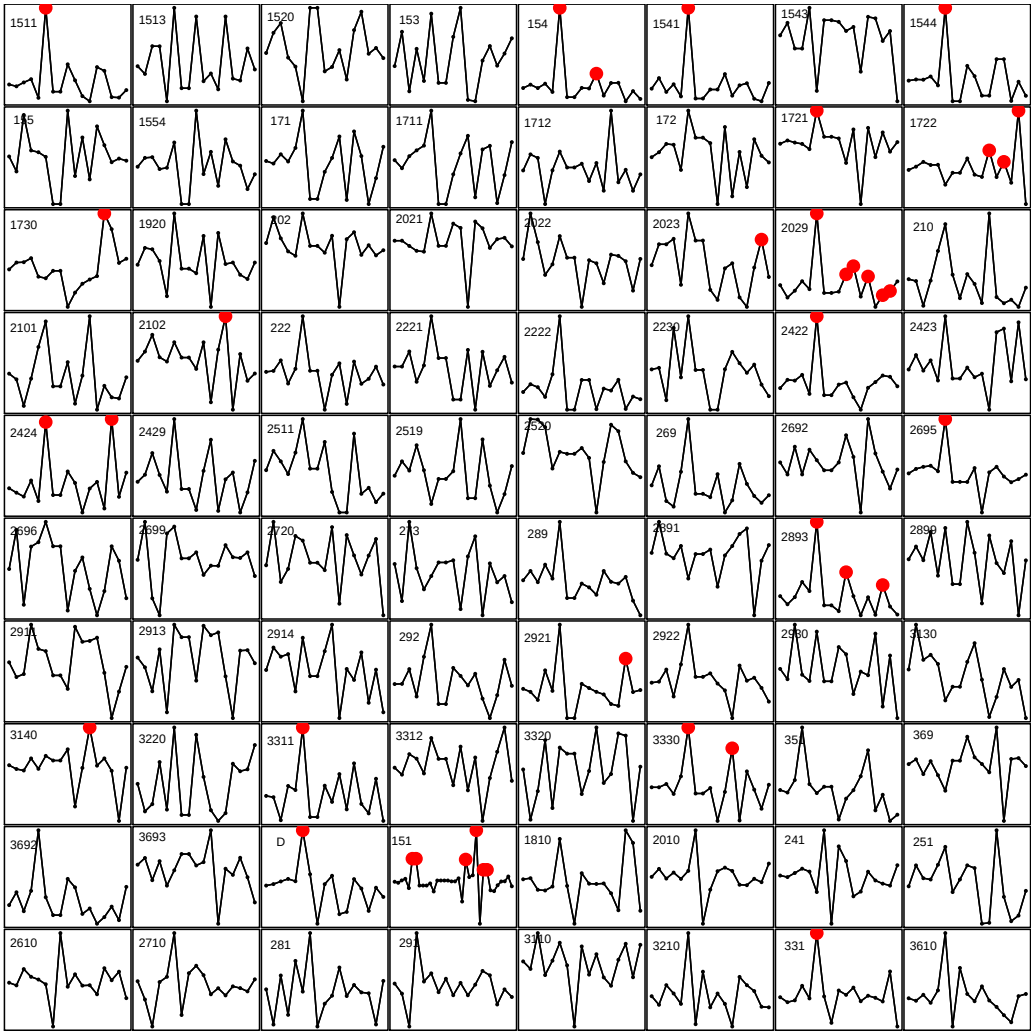


FIGURE 11. Repeated median filters using window width 9 of time series (VA/GO) including more than 14 observed values. Outliers are drawn with large symbols.

Edit 11 *[differential model] For all time series with more than 9 observed values mark those values which are detected by the differential model. Remaining missing values has to be imputed/interpolated beforehand.*

Edit 12 *[additive outlier model] For all time series with more than 9 observations mark item which are detected by the additive outlier model. Remaining missing values has to be imputed/interpolated beforehand.*

IV. CONCLUSIONS AND OUTLOOK

35. In this paper we presented the currently used in UNIDO screening methods and proposed new ones. The methods discussed vary from logical relationships in the data (balanced edits, e.g.

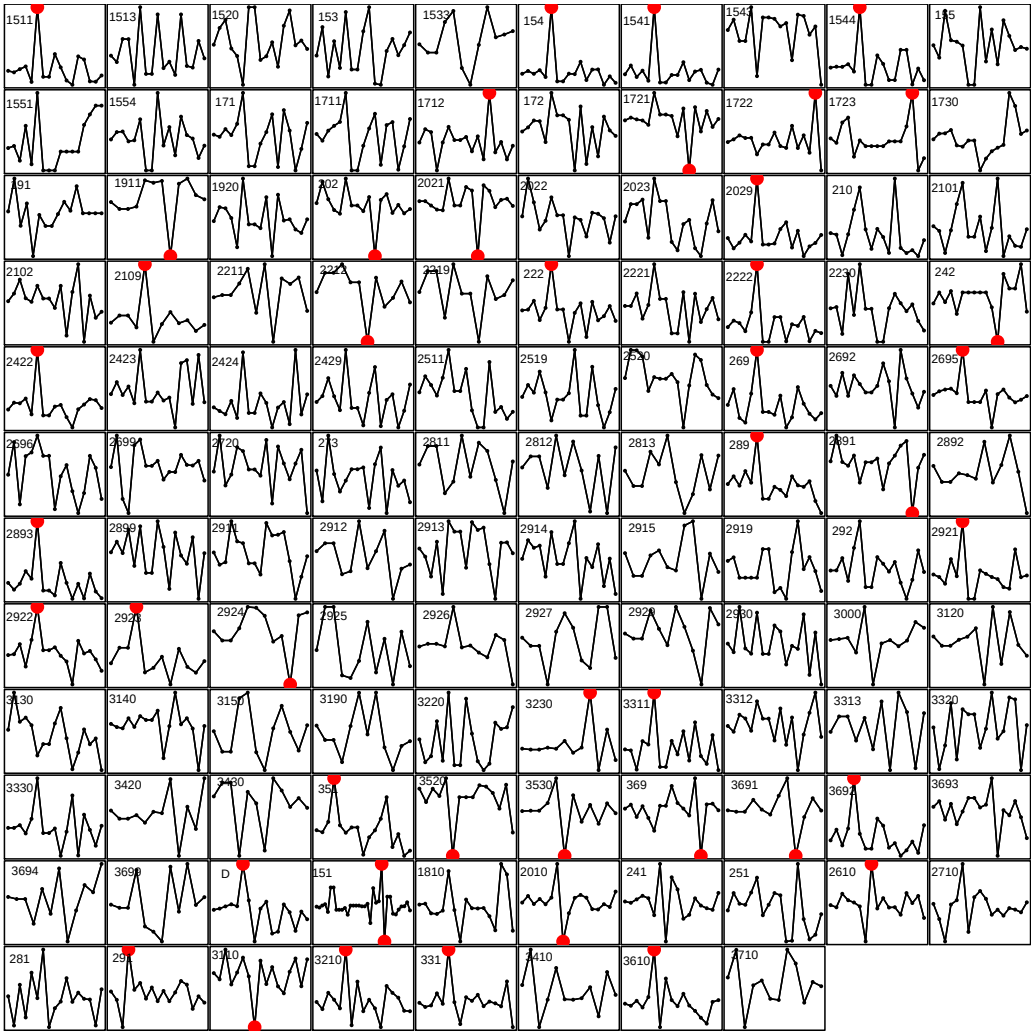


FIGURE 12. Differential model approach for time series (VA/GO) including more than 9 observed values. Outliers are drawn with large symbols.

wages and salaries cannot be negative, number of female employees cannot be larger than number of employees, etc.), to highly sophisticated outlier detection procedures.

36. The presented editing rules for soft ratio edits in most of the cases do not consider the possible trend in the data. This can be repaired by transforming the time series into white noise. It is shown that taking differences is often sufficient, but also an ARIMA(1,1,0) model is applicable. These methods are easier to implement as the editing rules are based on dummy regression models. The length of the time series is very important and only the most simple methods work with short time series while the methods based on innovation models are not applicable in such cases. It turned out that the robust filtering approach does not work very well for detecting possible errors since it declares too many values as outliers.

37. Further work is necessary to investigate the usefulness of the methods for all countries and if possible to adjust them, since the time series vary too much from country to country. Another problem in some countries is the presence of missing values. For all soft editing methods it is crucial how to fix the thresholds and parameters in such way that not too much values are declared to be

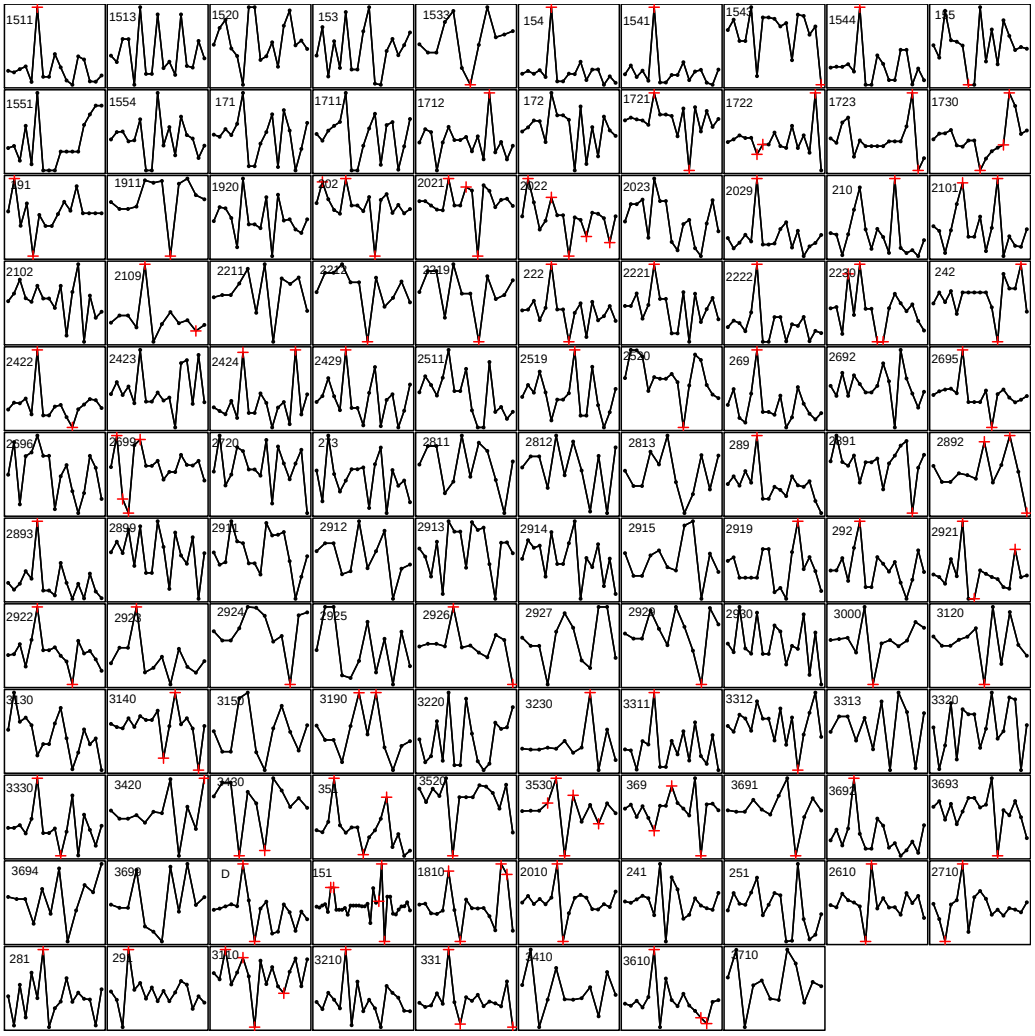


FIGURE 13. Additive outlier model approach for time series (VA/GO) including more than 9 observed values. Outliers are marked with crosses.

erroneous and on the other hand to detect reliably real errors in the data. This could be done by case studies involving pre-defined errors (detected by manual screening by a subject matter experts), then applying the soft editing methods on the same data sets and comparing the results.

ACKNOWLEDGEMENTS

The views expressed herein are those of the authors and do not necessarily reflect the views of the United Nations Industrial Development Organization (UNIDO).

References

G.E.P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons Inc, New Jersey, 2008. 4rd edition.

I. Chang, G.C. Tiao, and C. Chen. Estimation of time series parameters in the presence of outliers. *Technometrics*, 30:193–204, May 1988. ISSN 0040-1706. doi: 10.2307/1270165.

W. S. Cleveland. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

T. de Waal. An overview of statistical data editing. Discussion paper 08018, Statistics Netherland, 2008.

R. Fried. Robust Filtering of Time Series with Trends. *Journal of Nonparametric Statistics*, 16: 313–328, 2004.

R. Fried and U. Gather. On Rank Tests for Shift Detection in Time Series. *Computational Statistics and Data Analysis*, 52:212–233, 2007.

U. Gather and R. Fried. Methods and algorithms for robust filtering. In *Proceedings in Computational Statistics, Physika-Verlag, Heidelberg*, pages 159–170, 2007.

B. Hulliger and D. Lussmann. Bewertung der Nachhaltigkeits- und Umwelt-Indikatoren. Technical report, Institute for Competitiveness and Communication, University of Applied Sciences, Northwestern Switzerland, 2008.

A. Kowarik, A. Meraner, D. Schopfhauser, and M. Templ. Interactive adjustment and outlier detection of time dependent data in r. In *UNECE Work Session on Statistical Data Editing; Oslo, Norway*, page 11, 2012. to appear.

R.D. Martin and J.E. Zeh. *Robust Estimation of Autoregressive Models*, pages 228–254. Hayward, CA: Institute of Mathematical Statistics, 1980.

M. Templ. Correlation between indicators over time in thematic maps. *Austrian Journal of Statistics. Festschrift zum 65. Geburtstag von Rudi Dutter*, pages 67–80, 2012.

M. Templ, A. Alfons, P. Filzmoser, B. Hulliger, and D. Lussman. Visualisation tools. Research Project Report WP8 – D8.2, FP7-SSH-2007-217322 AMELI, 2011. URL http://ameli.surveystatistics.net.

Time Series Staff. X-12-ARIMA reference manual, version 0.2. Technical report, Statistical Research Division, U.S. Census Bureau, 2009. URL http://www.census.gov/srd/www/x12a/.

V.J. Yohai. High breakdown-point and high efficiency estimates for regresssion. *The Annals of Statistics*, 15:642–656, 1987.