AN INTEGRATED APPROACH FOR THE TREATMENT OF OUTLIERS IN SUB-ANNUAL ECONOMIC SURVEYS

Jean-Louis Tambay, Statistics Canada
11-P R.H. Coats Building, Ottawa, ONTARIO, K1A 0T6

KEY WORDS: Levels, outliers, trends, Winsorization.

1. INTRODUCTION

Sub-annual economic surveys, such as the ones conducted by Statistics Canada, are designed to estimate for the levels and the trends of economic activity. Many thus allow for a high proportion of sample overlap between consecutive periods. This compounds the problem of treating outliers - i.e., valid sample observations which have a large, undue influence on the estimates. Some of the methods which are used to treat large values in single-occasion surveys (Hidiroglou and Srinath, 1981) may harm the estimates of change when applied to sample units common to two or more occasions.

To help define a methodology for treating outliers in sub-annual surveys, several possible strategies were compared empirically. This paper presents the results of a study which measured the effects of the strategies on monthly estimates of levels and trends. The strategies treated level outliers, trend outliers, or both in sample data covering a fourteen month period.

The paper has four sections. Section 2 presents the basic elements of the strategies: studied variables, outlier detection rules and treatments. Section 3 describes the empirical study and presents its results. Concluding remarks are given in Section 4.

2. METHODOLOGY

2.1 General Framework

The strategies developed to treat outliers consider the types of sub-annual economic surveys where: trends and levels are of interest; the sample is continuing, with a large overlap between consecutive periods; and where stratification may be used, with estimates produced by applying a sample design weight to the data values. The weights may not be equal for all the units in a stratum. However, they should not change too much between periods.

2.2 Weighted Values in the Outlier Problem

Let the weighted estimator of the total for a given variable be denoted by $\hat{Y}=\Sigma s\ wy$, where w is the sample design weight and y is the variable value for each unit. The summation is done over s, the set of sample units. Unit subscripts have been omitted. If the previous period estimator is given by $\hat{Y}'=\Sigma s'\ w'y'$, where primes denote the previous period sample, weights and values, then the estimate of change over the previous period is

$$\hat{Y} - \hat{Y}' = \Sigma s\ wy - \Sigma s'\ w'y'$$

$$= \Sigma c\ w(y-y') + \Sigma c\ (w-w')y - \Sigma c\ (w-w')(y-y') + \Sigma a\ wy - \Sigma d\ w'y' ,$$

where c denotes the set of common sample units, a denotes additions to the sample and d, deletions from the sample.

Of the three terms involving common sample units, the latter two are usually insignificant relative to the first - unless weights change dramatically between periods. For non-common units, the weighted values wy for sample additions can be analyzed and treated for outliers while little is usually done for units after they leave the sample. For this reason, and because the total estimator also uses the weighted values wy, outlier detection and treatment are applied to the weighted values, wy and $wd = w(y-y')$. As stated above, the weights may not be equal within strata.

2.3 Detection of Outliers

As economic variables usually follow distributions which are highly skewed to the left, level outliers are detected using one-sided thresholds applied to wy. Trend outliers are detected using two-sided thresholds applied to wd. The quartile distance method is used to detect outliers because it is simple, non-parametric and has a high breakdown point (roughly meaning that outlier detection is very robust to the presence of outliers in the data).

In each stratum, the thresholds are calculated from the sample quartiles of wy and wd. If the n stratum units are sorted by increasing value of the study variable, then the definition used for the qth quartile is the $\{q(n+1)/4\}$th observation, where linear interpolation is used if the expression is not an integer. The second quartile is the sample median.

For levels, the upper threshold is at

$$UT = MED(wy) + K1\ \{Q3(wy) - MED(wy)\},$$

for trends, the upper and lower thresholds are at

$$UT = MED(wd) + K2\ \{Q3(wd) - MED(wd)\}, \text{ and}$$
$$LT = MED(wd) - K2\ \{MED(wd) - Q1(wd)\},$$

where $Q1(\cdot)$, $MED(\cdot)$ and $Q3(\cdot)$ denote the first, second and third quartiles for the argument term obtained from the sample. This rule is similar to the rule of fences given in Tukey (1977). The values K1 and K2 are constants used to calibrate the percentages of observations identified as outliers.

Outliers are weighted values which exceed their corresponding thresholds.

2.4 Treatment of Outliers

Two methods are considered for the treatment of outliers: Winsorization and the Dalén method (Dalén, 1987). For levels, the value y is replaced by y° defined as

$$\text{Winsorization: } y^\circ = y \quad \text{if } wy \le UT,$$
$$= UT/w \quad \text{otherwise;}$$

Dalén Method:

$$y^\circ = y \quad \text{if } wy \le UT,$$
$$= (1/w)y + \{1-(1/w)\}(UT/w), \text{ or}$$
$$UT/w + (1/w)\{y - (UT/w)\} \quad \text{otherwise.}$$

As shown, Winsorization replaces the outlier value y by its threshold, UT/w. This is different from the usual Winsorization which is applied when weights are equal, and which replaces the outlier value by a limiting sample value. Here, UT/w does

not necessarily correspond to a sample value.

The Dalén method can be thought of as a linear combination of the original value and the Winsorized value. Alternatively, it reduces to one the weight for (y - UT/w), the part of the value y which is over the threshold UT/w.

The same can be done for the weighted differences, wd. Once a replacement value d° is obtained, the current value y is replaced by y° = y'+d°. Note that, for trends, the procedure is two-sided on wd, and y° can thus be larger or smaller than y.

These represent the basic elements of the outlier treatment strategies. The strategies will be given after a description of the survey data used in the study.

3. EMPIRICAL STUDY

3.1 Description of Data Used in the Study

The strategies were applied to historical data from the monthly Shipments, Inventories and Orders (SIO) survey at Statistics Canada. Data on monthly manufacturing shipments from August, 1986 to October, 1987 were obtained from sample units in fourteen industries.

The SIO sample is stratified into three size strata within each industry and province. The sample is originally selected as a simple random sample but, as units changing industry are moved to their new industry strata, ends up containing units with varying weights in the same stratum. Rotation of the sample is not used. For study purposes, the sample was treated as a simple random sample from a population which was constant over the study period. Frame population counts were used to provide the new sample weights.

There were 144 take-some strata in the studied industries. Many had small sample sizes, as shown in the table below. The outlier treatments were only applied to take-some strata with 7 or more sample units. The remaining take-some strata accounted for less than 3% of the estimated all-industry total shipments value.

stratum sample size	number of strata	observed new sample weights minimum	maximum
2 - 6	56	1.167	42.50
7 - 14	43	1.125	23.73
15 - 22	21	1.190	17.25
23 - 30	14	1.071	13.68
over 30	10	1.136	9.86

3.2 Outlier Strategies Studied

Fourteen outlier treatment strategies were studied. They were characterized by the method used to replace outliers (Winsorization or Dalén), by the way they were applied to level and trend outliers (Levels Only, Trends Only, Combined and Integrated) and by the overall percentage of weighted values, wy and wd, identified as outliers (2% and 5%).

Outlier rates of 2% were obtained by setting K1=5.99 and K2=20.70 in the threshold equations. Rates of 5% were obtained with K1=3.38 and K2=9.80.

The "Level Only" approach treated only level outliers (wy) and the "Trend Only" approach treated only trend outliers (wd) as explained in Section 2. For trend outliers, the procedure was applied to the oldest data first (i.e., the August to September, 1986 differences) and thus the replaced values affected future trends, as intended. Both approaches replaced only the current values for the month being examined.

The "Combined" approach treated trend outliers first each month, then applied the level outlier procedure to the data treated for trends. If replacing the value of a level outlier caused it to become a trend outlier, then the replacement was not carried out for that value.

The "Integrated" approach treated the data from the month prior to the study period, September, 1986, for level outliers. If a value was changed, then all the other month values for the unit were changed by the same percentage. The trend procedure was then applied to the data. This procedure simulated the effect of treating new sample units which are level outliers by decreasing all their future contributions to the estimates.

Winsorization was used for the Level Only, Trend Only and Combined approaches. The Dalén method was used for all four approaches. These seven treatments, applied at the 2% and 5% outlier rates, gave the fourteen outlier treatment strategies.

The fourteen months of data provided fourteen sets of level estimates and thirteen sets of trends. Only the latest twelve levels and trends were studied, however. The other months were used to "start off" the procedures which treated trend outliers.

3.3 Study Results

The study concentrated on the effects of the outlier treatments on the monthly estimates obtained at the industry level. Estimates of totals ($\hat{Y}$) and month-to-month changes ($\hat{Y}-\hat{Y}'$), and their variances, were calculated using standard equations for stratified simple random samples.

Estimates and variances for the treated data were calculated by substituting replacement data for original data in the standard equations.

Three relative measures of the effects of the outlier treatments were obtained. They measure the relative bias of the treatments, the decrease in the variance due to the treatments, and the Mean Square Error for the treated data relative to that for the untreated data. The measures are:

$$\text{Neg.Rel.Bias} = \{\text{Est(Trmt)}-\text{Est(Orig)}\}/\text{StdErr(Orig)},$$
$$\text{Rel.Std.Err.} = \text{StdErr(Trmt)} / \text{StdErr(Orig)}, \text{ and}$$
$$\text{Relative MSE} = (\text{Neg.Rel.Bias})^2 + (\text{Rel.Std.Err.})^2.$$

Figures 1 and 2 show scatterplots of the first two measures on the monthly industry estimates of levels (totals) and trends (month-to-month changes), respectively. The scatterplots are for the seven treatment strategies which use a 2% rate for outliers. The strategies are identified by two characters (WC for Winsorized Combined, etc.) and two digits (02 for 2% outlier rates) placed on each scatterplot. Each plot has 168 observations, corresponding to twelve monthly estimates for each of the fourteen industries.

The plot axes are not all the same. To assist comparisons, standard boxes have been superimposed on each plot. The top and bottom of each box indicate where the treatments did not affect the standard errors, and where they halved them. The side bounds mark where the biases were equal to one-half of the original estimated standard errors, a non-negligible amount.

The unit circle was also traced over each scatterplot as it shows where the Relative Mean Square Error is equal to one. Points outside the circle represent losses of "efficiency" due to the treatments. Points inside represent gains.

The following observations are made:

- Winsorization strongly influences many of the estimates, resulting in many losses of efficiency.
- The Dalén methods fare better in that respect.
- For level estimates, most of the biases are in the same direction. Standard errors rarely rise.
- Strategies DL02, DC02 and DI02 seem to have similar effects on the level estimates. Not unexpectedly, DT02 has a smaller effect.
- For trends, methods DT02 and DI02 give similar effects. DL02 has a smaller effect on the trends, but DC02 has more cases of loss of efficiency.

Figure 3 shows some of the scatterplots obtained with a 5% rate of outliers. The effects of the treatments on the estimates are more pronounced, and more losses of efficiency occur. An exception is the DT method for levels, which is not affected as much by the increased number of outliers treated. This is probably because it is symmetrically applied to the data.

A summary of the scatterplots is given in the table below. It gives, for each method, the average of the 168 calculated Relative Mean Square Error values. Also given are averages of the Relative MSE's calculated at the all-industry levels. These are averaged over the twelve study months.

Averages of the Relative Mean Square Errors

Outlier Treatment Strategy	For Levels		For Trends	
	Industry Average	All-Indus Average	Industry Average	All-Indus Average
WL02	1.35	2.28	1.38	1.10
WT02	1.17	0.99	1.20	0.92
WC02	1.21	1.35	1.42	0.99
DL02	0.95	1.32	1.03	0.95
DT02	0.99	0.98	0.97	0.86
DC02	0.95	1.20	1.08	0.84
DI02	0.98	1.38	0.96	0.87
WL05	1.92	6.21	1.54	1.13
WT05	1.26	1.05	1.29	0.78
WC05	1.40	2.10	1.54	0.91
DL05	1.04	3.02	1.04	0.88
DT05	0.99	1.00	0.93	0.62
DC05	1.01	1.69	1.07	0.80
DI05	1.06	2.81	0.91	0.65

The industry averages confirm the observations made earlier: that Winsorization is not as efficient, for the rates of outliers studied, as the Dalén method; that method DT02 has a smaller effect on the level estimates than the other Dalén methods - and the effect is close to that of DT05; and that other treatments are not as efficient when the rate of outliers is at 5%.

Of interest is the difference between the industry averages and the all-industry averages of the Relative MSE's. For levels, it is large and positive for those approaches which treat level outliers. This is because, unlike the "Trend Only" approach, the all-industry bias is an aggregate of industry biases which are mostly in the same direction. In fact, the all-industry results for levels do not seem to promote the use of any treatment method.

For trends, it is the all-industry averages which are lower. This reduction is largely due to the Relative Standard Error component of the Relative MSE. It decreases substantially at the all-industry level. The fact that for trends all the methods have biases in both directions also contributes to the smaller all-industry Relative MSE's. Biases tend to cancel each other out more often, resulting in a smaller aggregate bias.

4. CONCLUDING REMARKS

The study has demonstrated that there is a potential for exploring further approaches which treat both level and trend outliers in sub-annual surveys. The method chosen to combine the two may depend on factors such as the relative importances of level and change estimates, or of industry level and all-industry level aggregates, and, of course, on operational factors. Procedures to treat trend outliers may be more difficult to incorporate into a survey.

Just as the methods complement each other, each carries risks. For example, the trend outlier procedure may carry for many months the repercussions of a very unusual value in one month by moving future values in the direction of the outlier. Good outlier diagnostics and safeguards will always be needed.

REFERENCES

Barnett, V. and Lewis, T. (1978). **Outliers in Statistical Data**. New York: John Wiley.

Berthelot, J.-M. and Hidiroglou, M.A. (1986). Statistical Editing and Imputation for Periodic Business Surveys. **Survey Methodology**, 12, 73-83.

Dalén, J. (1987). Practical Estimators of a Population Total Which Reduce the Impact of Large Observations. R & D Report. Statistics Sweden.

Hidiroglou, M.H. and Srinath, K.P. (1981). Some Estimators of a Population Total From Simple Random Samples Containing Large Units. **Journal of the American Statistical Association**, 76, 690-695.

Tukey, J.W. (1977). **Exploratory Data Analysis**. Reading, MA: Addison-Wesley.

ACKNOWLEDGMENTS

The author would like to thank Hyunshik Lee and Jack Gambino of Statistics Canada for their many helpful comments and suggestions.

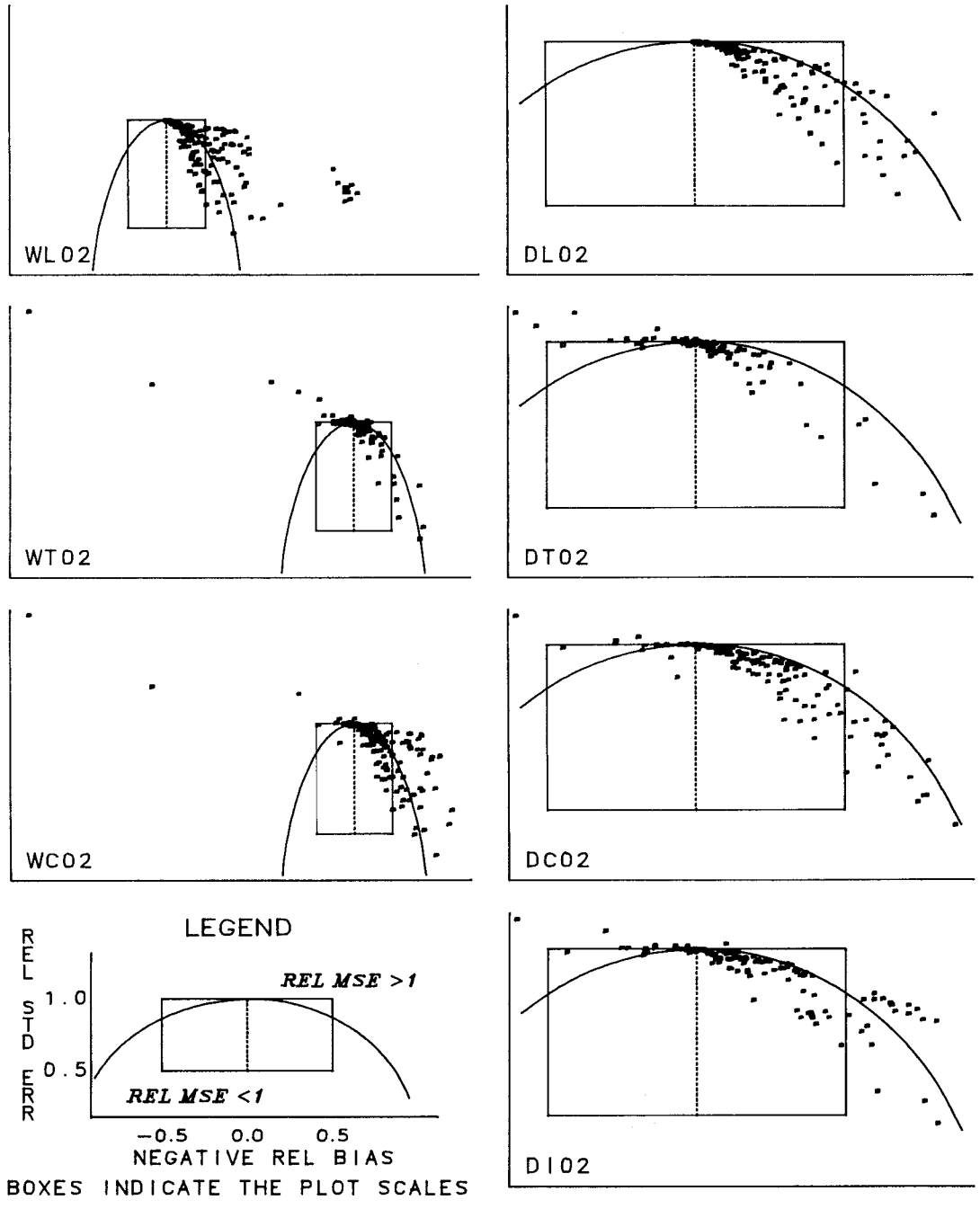

FIGURE 1 - EFFECT OF THE OUTLIER TREATMENTS ON THE MONTHLY ESTIMATES OF LEVELS.

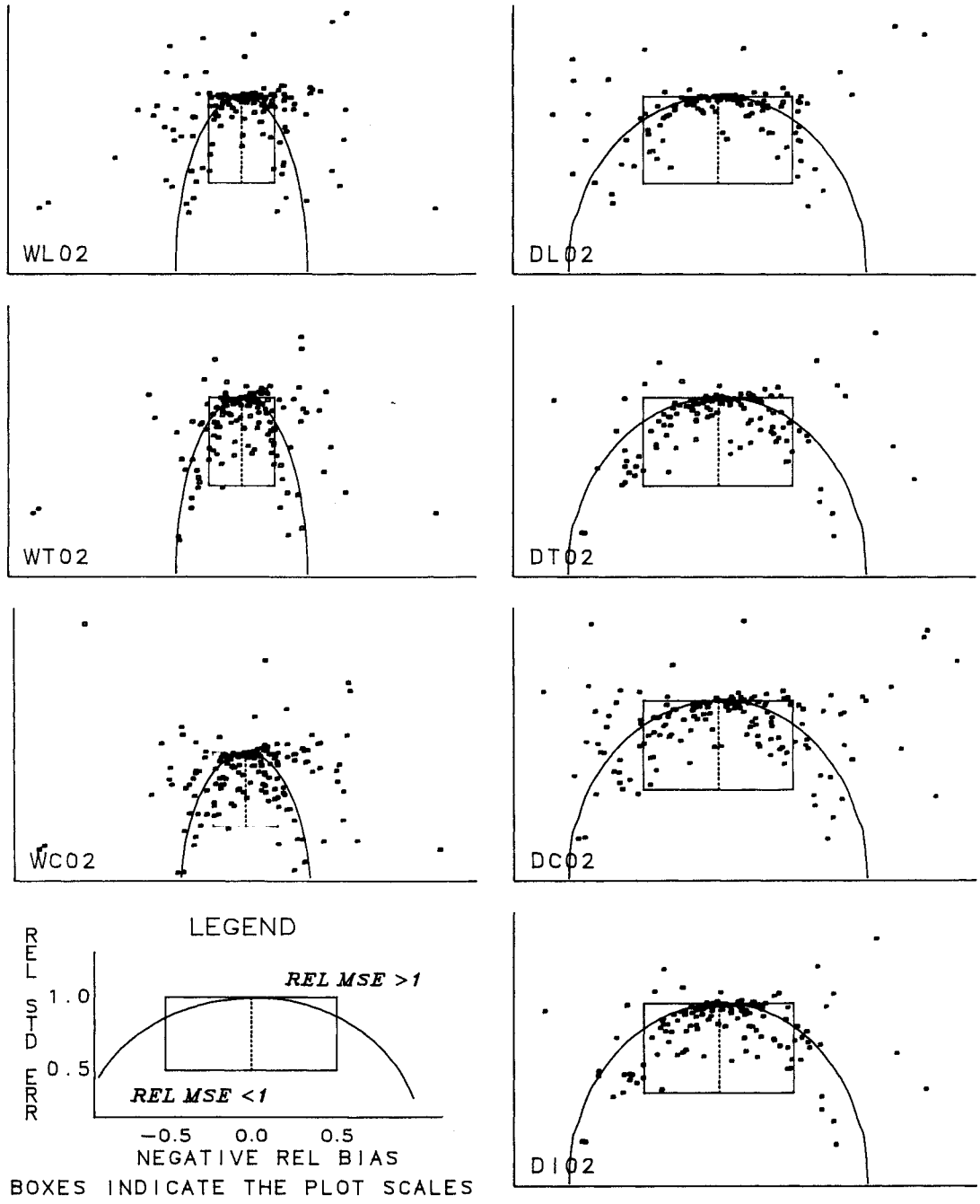

FIGURE 2 - EFFECT OF THE OUTLIER TREATMENTS ON THE MONTHLY ESTIMATES OF TRENDS.

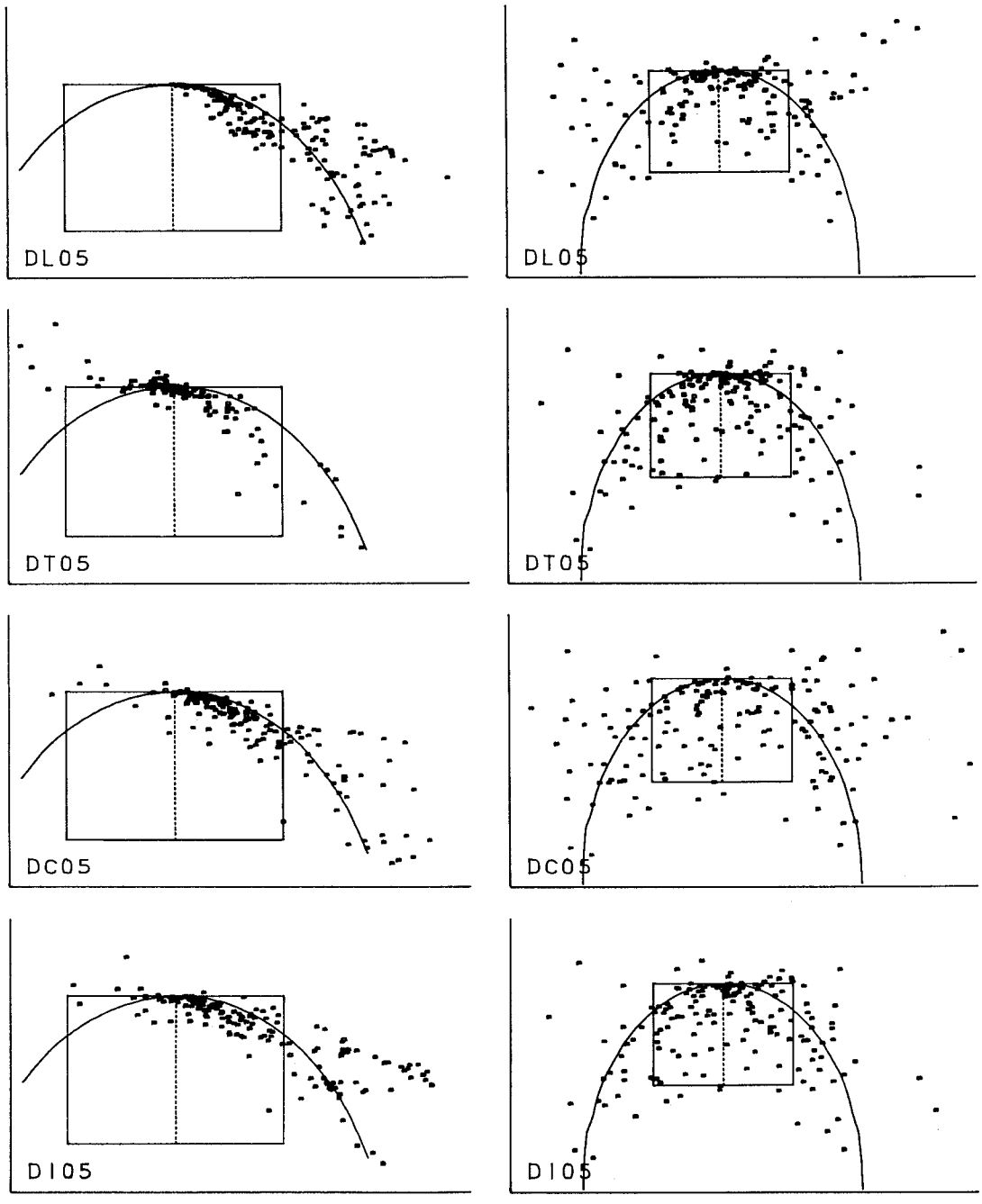

FIGURE 3 - EFFECT OF THE OUTLIER TREATMENTS ON THE MONTHLY ESTIMATES OF LEVELS (LEFT) AND TRENDS (RIGHT). LEGEND IS THE SAME AS FIGURES 1 AND 2.