

Mirror Outlier Detection; An Overview

Michael Ashbrook

Eurostat, michael.ashbrook@ec.europa.eu

Abstract: Data describing international trade may be arranged into time series of mirroring imports and exports. Where outliers occur in these mirror series, data quality may be checked by the Mirror Outlier Detection (MOD) method. Internally, MOD relies on the robust, non-parametric Median Absolute Deviation algorithm to detect outliers in time series of trade data.

Keywords: Comext, data quality, ESTAT, Eurostat, international trade data, non-parametric, outlier detection, robust, time series,

1. Introduction

After reading or hearing this material, the audience should

- understand how MOD works as a method of checking data quality
- remember how it applies to trade data stored in Eurostat's Comext DB which serves as an example
- be able to decide whether or not MOD can be used to analyse the quality of other data.

This presentation is deliberately restricted to a qualitative analysis of trade data quality because results are available only for 12 months of intra EU trade at this time.. Quantitative results will be made available later when more work has been accomplished. Having said that, the discrepancies found during a typical month vary smoothly in magnitude from a few thousand to a bit over a billion EUROS.

The idea of MOD was independently developed at the same time by Mr Krokow of the Statistisches Bundesamt and Mr Pongas of Eurostat who shared it with the author. . The robust, nonparametric algorithm used for outlier detection in MOD was developed by Mr Ashbrook of Eurostat.

2. The Structure of Trade Data

It is sufficient to think of Comext as a flat file of records with seven fields that interest us for now and disregard the details of implementation. We need answers to seven questions in order to know about trade.

- 1) The first question is whether the transaction was an import or an export. Exports and imports are collectively known as flow.
- 2) Where did this data come from? The country that reports the trade is called the reporter.

- 3) The country with which it has traded is referred to as the partner. Note; that the reporter may be importing or exporting, that makes no difference. The point is that he is reporting the transaction. Ditto for the partner.
- 4) When did the transaction take place? In our case the data is reported per calendar month. The time at which the trade took place is the period.
- 5) Obviously, we want to know just what was traded: product. We are concerned with 8 digit codes that identify circa 17.000 different kinds of products. Our product codes come from the "Combined Nomenclature" known as CN for short.
- 6) We need to know how much money changed hands: value.
- 7) When we ask how much was traded it is often enough to weigh the stuff: quantity.

So, we see that a reasonably complete description of an international trade is made up of at least seven parts.

For MOD, the way to sort the records is first by flow, then by reporter, then by partner, then product and finally by period. That way, for each 4-tuple of [flow, reporter, partner, product] we get a time series of 7-tuples. The 4-tuple of [flow, reporter, partner, product] serves as the identifier of the time series.

3. The MOD Method

Let's take the trade in spelt between the Czech Republic and Germany as an example. We call time series

$TS_1 = [\text{Export, Czech Republic, Germany, Spelt}]$ and

$TS_2 = [\text{Import, Germany, Czech Republic, Spelt}]$

mirror time series because they should have the same values and quantities.

Notice the word "should"? You are right to conclude that the values of a time series and its mirror time series are not always equal in practice. That brings us to simple underlying idea of MOD; if a time series has an outlier, then its mirror should have an outlier for the same month. If it does, we have good reason to trust the data and conclude that there really was a spike in trade for that month.

4. Seven Categories of Trade Data

Most pairs of mirror time series that are free of outliers. Methods other than MOD must be used to verify their quality.

An outlier that is found for the same month in both of the mirror series is called a green outlier. Discovering them can help to spot changes in trade as soon they appear in our data. This data merits our trust.

If an outlier is found only in one of the two mirror series it is called a red outlier. Maybe the reporter made a mistake and entered an outlier that did not exist, or perhaps the partner failed to detect a real outlier. It is up to the two of them to decide what the data

should be, but something needs to be changed. Naturally, at most one of the two values found can be correct.

Surprisingly, perhaps, things get even worse than red when mirror time series both have outliers, but in opposite directions. This is the worst case of data quality. Fortunately, these violet mirror outliers are the rarest of all.

But things can also get better. Each Member State has the right to "censure" the trade data that Eurostat publishes in order to protect someone's legitimate interests. But it has no authority to censure its partners' trade data. This sometimes leads to the case of one time series with an outlier, but no mirror time series at all. We call these unmatched outliers pink. As far as MOD is concerned, there is no problem with pink data because the asymmetry is the result of deliberate policy.

In the next case, only one of the two mirror time series can be found in Comext, the other is missing. But here neither the reporter nor the partner censure the data for the product in question. This is a typical black outlier. Often it turns out that the partner has classified the product with one particular 8-digit code while the reporter has chosen a different one. (It is hardly plausible that someone has made the same typographical error 36 times in a row.) Black is a special case of red, with one of the two values being blank. Again, something is wrong with the data.

Sometimes we find a record in Comext that has a value above zero but a quantity equal to zero. If this data were accurate, there would be people in Europe who pay good money (sometimes more than 20 million EUROS) for nothing at all. That is nonsense; we are looking at a list of mistakes. Playfully, we have baptized these the lightweight products.

5. Causes of Mirror Outliers

- 1) Honest human errors in entering the values
- 2) Deliberate fraud, especially VAT fraud
- 3) Time lags in recording between reporter and partner
- 4) Confusion about reporter and partner when transit countries come between them
- 5) Confusion about currencies
- 6) Confusion about units of measurement for the quantity
- 7) Disagreements about the correct product code

6. The MAD Algorithm

The algorithm is applied to each time series of values and of quantities that is found in Comext. It is also applied to the time series of unit values calculated from the values and quantities.

- 1) Without loss of generality, let TS be the time series of values V_1 through V_{36} .
- 2) The median M of TS is calculated.

- 3) For $i = 1$ through 36, $AD_i = |M - AD_i|$
- 4) The median MA of $\{AD_1, \dots, AD_{36}\}$ is calculated.
- 5) Those V_j with $AD_j > 5 * MA$ are selected as outliers.

With circa 76.000.000 records for a total of 20 GB of input data for each month of intra EU trade data, a computationally efficient method is needed. MAD is easily seen to be robust and non-parametric.

7. Conclusion

- MOD works as a method of checking data quality by systematically testing the assumption that a time series and its mirror time series should be equal even when outliers are present.
- Like all real world data bases, Comext is not perfect; up to thirty percent of the records may be somewhat questionable, but there are rarely more than fifty discrepancies of over one hundred million EUROS for a given month of intra EU trade data. For the time being, Eurostat forwards mirror outlier reports to the 27 agencies that supply the raw data for Comext because they have the exclusive right to modify it.
- Any data that can be arranged into mirror time series is open to MOD; examples could be balance of payments, flows of migration, or any other time dependent flows of something or other that should obey a "conservation law".

References

- Eurostat, 2006: Statistics on the trading of goods – User guide. Methods and Nomenclatures, Luxembourg: Office for Official Publications of the European Communities, pp. 20 – 22.
- Ricardo A. Maronna, Douglas R. Martin, Victor J. Yohai, 2006: Robust Statistics: Theory and Methods, John Wiley & Sons, Ltd.
- Oskar Morgenstern, On the Accuracy of Economic Observations, 1963 Princeton University Press
- Jerzy Rozanski, Alexander Yeats, On the (in) Accuracy of Economic Observations: An Assessment of Trends in the Reliability of International Trade Statistics, 1994, Elsevier Science
- P. Sprent, Applied Nonparametric Statistical Methods, 1989, Chapman & Hall
- Nassim Nicholas Taleb, The Black Swan, 2007, Random House
- Tommy Wright, Statistical Methods and the Improvement of Data Quality, 1983, Academic Press, Inc.