

**UNITED NATIONS ECONOMIC COMMISSION
FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN
UNION (EUROSTAT)**

**ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE**

Meeting on the Management of Statistical Information Systems (MSIS 2010)
(Daejeon, Republic of Korea, 26-29 April 2010)

Topic (ii): Software sharing and shared maintenance

R in the Statistical Office: The UNIDO Experience

Invited paper

Prepared by Valentin Todorov (UNIDO)

I. INTRODUCTION

1. The R language ([Ihaka and Gentleman, 2009](#)) is a freely available environment for statistical computing and graphics. It can be used for handling and storage of data and performing of statistical calculations. R provides a number of tools for data analysis, and features excellent graphical capabilities included in a straightforward programming language. R is widely used for statistical analysis in a variety of fields and is backed up by a large number of add-on packages that extend the system.

2. The statistical techniques available in R - linear and non-linear modeling, classical statistical tests, time-series analysis, classification, clustering, robust methods, etc. - as well as its data manipulation and presentation tools make this package an ideal integrated environment for both research and production in the field of official statistics. However its application in the National Statistical Offices and International Organizations is still quite limited. This is mainly due to the wide spread opinion that R is hard to learn compared to the other statistical packages, like SAS and SPSS and it is said to have a very steep learning curve. The lack of a true point-and-click interface adds to this these (meanwhile the development goes on and already several packages are available providing graphical user interface for R). The technical documentation accompanying most of the packages rarely includes syntax examples related to the analytical methods applied in the official statistics.

3. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. R can be obtained as both source and binary (executable) forms from the Comprehensive R Archive Network (CRAN). The source files are available for a wide variety

of UNIX platforms and similar systems (including FreeBSD and Linux) as well as for Windows and MacOS for which are available also precompiled binary distributions of the base system and contributed packages.

4. The most recent release of R is version 2.10.1 (released on 14.December 2009) and pre-release versions of 2.11.0 are in progress.

5. A wide variety of add-on functionality (actually the normal way of extending R) is available from the same web page in the form of contributed R packages, which can be downloaded in source form or installed directly from the R console by using the `install.packages()` function (provided the computer is connected to the Internet). A few examples relevant for national and international statistical organizations are: survey analysis (**survey**, **pps**, **sampling**, **sampfling**), handling of missing data (**VIM**, **mice**, **mi**, **mvnmle**, **mitools**, **EMV**, **mix**, **pan**), time series analysis, robust statistics (**robustbase**, **rrcov**, **robust**). More information on R can be found at the CRAN web site <http://cran.r-project.org>. One could also read a brief overview of R in the paper [Todorov \(2008\)](#).

6. This paper has rather an informative character than a tutorial form. Many aspects of the application of R in the statistical offices are considered but the details are left for another material. A tutorial covering the above areas as well as other relevant for the official statistics with example data sets and code snippets is under preparation.

7. In the following three sections will be considered three typical applications of the programming language and system R, based on examples from the statistical practice of UNIDO. In Section II the abilities of R to import and export data from and to different data formats and different statistical systems is considered. This features render it a very useful tool for facilitating the collaboration in the statistical office. The second example application described in Section III considers the graphical excellence of R as applied to the statistical production process of UNIDO for generating publication quality graphics included in the *International Yearbook of Industrial Statistics*. The graphics together with the related text are type set in L^AT_EX using the dynamically reporting tool implemented in the R package **Sweave**. The third example illustrates the analytical and modeling functions available in R and add-on packages. These are used to implement a nowcasting tool for the Manufacturing Value Added necessary for generating estimates to be published in the *International Yearbook of Industrial Statistics*. Functions from the package for robust statistics **robustbase** are used for this purpose.

II. R AS A MEDIATOR

8. When using a statistical system we must have in mind that this is not done in isolation and the system must be able to communicate with other systems in order to import data for analysis, to export data for further processing (use the right tool for the right work) and to export results for report writing. When collaborating with other researches, they may have already created the necessary data sets in their favorable statistical package or may be they want to access a data set that is already available in another data format. It can also happen that within the same data process at some stage another statistical tool is more appropriate for a particular task but the final result should be produced in the original system. Examples of such task sharing are presented in the subsequent sections III and IV. Even in a small research department like the one at UNIDO many different statistical systems like SAS, Stata, Eviews, Octave, SPSS and R are in use and often the need for collaboration between the tools arises. It is not a rare that a consultant provides the final report on a media in a binary format written by some program like SPSS or Excel or something else.

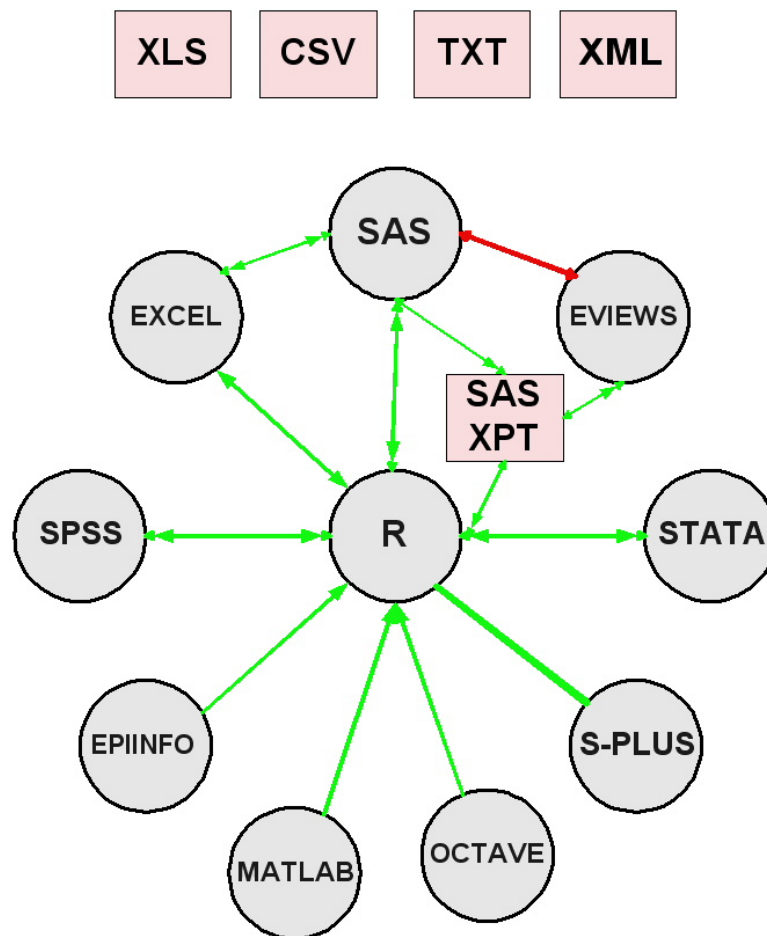


FIGURE 1. R interfaces to other statistical systems and data formats.

9. In most of these cases a tool which could "mediate" among the other programs and formats would be very useful. To our experience R provides the most flexible import and export interfaces to most of the available statistical and other packages. A rich variety of facilities for data import and export as well as for communication with databases, other statistical systems and programming languages are available either in R itself or through packages available from CRAN. In Figure 1 are shown most of the interfaces available in R. We do not claim comprehensiveness of the presented relations between the different systems but rather illustrate our experience in exchanging data between systems (for example the red arrow between SAS and EViews means that the link between the two packages based on a SAS ODBC driver is to our knowledge broken). The easiest data format to import into R is a simple text file but reading XML, spreadsheet-like data, e.g. from Excel is also possible. The recommended package **foreign** provides import facilities for reading data in the format of the statistical packages Minitab, SAS, S-Plus, SPSS, Stata, SYSTAT and Octave as well as export capability for writing Stata files, while the package **matlab** provides emulation for MATLAB. Further details can be found in the documentation of the packages **foreign** and **Hmisc**.

10. Working with large data sets could be a problem in R (if the data do not fit in the RAM of the computer) but the interface to RDBMS could help in such cases. Another limitation is that R does not easily support concurrent access to data, i.e. if more than one user is accessing, and perhaps updating, the same data, the changes made by one user will not be visible to the others. This could

also be solved by using the interface to relational databases. There are several packages available on CRAN for communication with RDBMSs, providing different levels of abstraction. All have functions to select data within the database via SQL queries, and to retrieve the result as a whole, as a data frame or in pieces (usually as groups of rows). Most packages are tied to a particular database - **ROracle**, **RMySQL**, **RSQLite**, **RmSQL**, **RPgSQL**, while the package RODB provides a generic access to any ODBC capable relational database. Another way to solve the large data set problem is to use the new (August 2007) package **filehash** - it implements a simple key-value style database where character string keys are associated with data values that are stored on the disk and the provided utilities allow to treat the database much like the familiar in R environments and lists.

III. R AS A GRAPHICS ENGINE

11. A natural way to visualize data are graphs and plots and it hardly needs saying that their use is common even in non-technical documents. Usually in order to produce plots with sufficient output quality to match a well-typeset document much work is required (and rarely performed). Without going into details about graphical data visualization, for which neither the time nor the space is available within this work, we will refer to [Tuft \(2001\)](#) who states that publication quality displays should be both informative and aesthetically pleasing and lists several important aspects of data presentation which contribute to graphical excellence. Some of these aspects are:

- present many numbers in a small space;
- encourage the eye to compare different pieces of data.

12. One of the most important strengths of R is the ease with which simple exploratory graphics as well as well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control. R provides the standard statistical plots which are usual form most of the statistical packages like scatterplots, boxplots, histograms, barplots and piecharts as well as basic 3D plots which can be produced by a single function call. A fundamental feature of the R graphics is that graphical elements can be added sequentially to a plot in order to obtain the final result. In addition to the traditional statistical plots R has an implementation of the *trellis* plots through the package `lattice` (see [Sarkar, 2008](#)). *Lattice* is a powerful and elegant high level data visualization system that is sufficient for most everyday graphics needs, yet flexible enough to be easily extended to handle demands of any research.

The best way to visualize the potential of R when it comes to producing publication quality graphics is to type `demo(graphics)` at the R prompt and then navigate through a list of example plots. Another exciting example of the R graphics capabilities is the R Graph Gallery which aims to present many different graphics fully created with the programming environment R. Graphs are gathered in a MySQL database and browsable through PHP.

R can produce graphics in many formats, including:

- On screen
- PS and PDF files for including in \LaTeX and pdfLaTeX or for direct distribution
- PNG or JPEG bitmap formats for the WEB
- On Windows, metafiles can be created to be used in Word, PowerPoint, and similar programs

For example to save a plot to a PDF file the following commands are necessary.

```
> pdf(file="graph10.pdf")
> plot(graph10)
> dev.off()
```

Alternatively one could plot the graphical result on screen and than call the function `savePlot()` to save it as shown in the next example:

```
> plot(graph10)
> savePlot(file="graph10", type="pdf")
```

An excellent reference to R Graphics is the book of Paul Murrell ([Murrell, 2005](#)), a member of the R Core Development Team who has not only been the main author of the grid package but has also been responsible for several recent enhancements to the underlying R graphics engine.

13. The *International Yearbook of Industrial Statistics* published by UNIDO - see <http://www.unido.org/index.php?id=o3544> - is a unique and comprehensive source of information, the only international publication providing economists, planners, policymakers and business people with worldwide statistics on current performance and trends in the manufacturing sector. The Yearbook is designed to facilitate international comparisons relating to manufacturing activity and industrial development and performance. It provides data which can be used to analyze patterns of growth and related long term trends, structural change and industrial performance in individual industries. Statistics on employment patterns, wages, consumption and gross output and other key indicators are also presented. In its part I the yearbook contains summary tables for the manufacturing sector (1.1) and for the manufacturing branches (1.2) and the second part consists of country tables. Recently a new section was added to Part I containing analysis of the major trends of growth and distribution of manufacturing in the world. This section is illustrated with many graphical displays based on the summary tables in 1.1. An example is shown in Fig. 2.

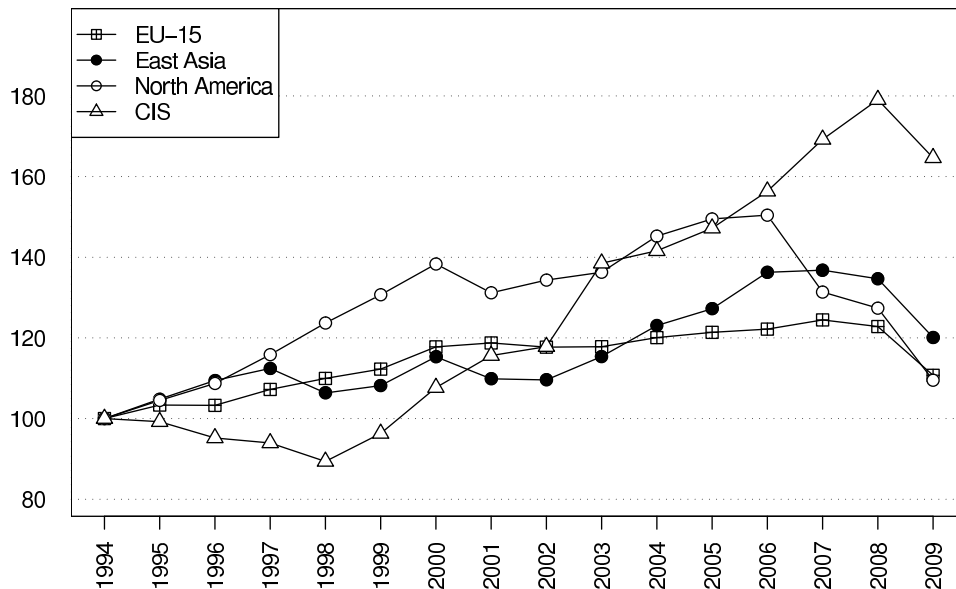


FIGURE 2. Growth of MVA in industrialized regions, at constant 2000 USD (1994=100)

Most of the industrialized countries have suffered a severe decline in their industrial production in the last two years or so. The worst affected region has been North America, where according to UNIDO estimates, manufacturing output has fallen by 20 per cent since 2007. CIS countries, which recovered just at the beginning of this decade from the turmoil of the 1990s and became one of the fastest growing regions, have also suffered the set back from the recent financial crisis.

14. The production line for the yearbook receives as an input the 'raw' data from the UNIDO database but also from other external sources and generates as an output a ready-to-publish PDF file which is submitted to the publishing house. Most of the components of the system are developed in SAS (historically, migrated from the Mainframe) or in .Net (all new development, interfacing to the SAS-based subsystem and generating the output using CrystalReports). The new section consists of several pages of graphics and text typeset in two columns in landscape format. The software for the production of this output should fulfill several key requirements:

- (1) To be able to create publication quality graphics. Although this is possible to do with other packages too, R provides the most flexible solution.
- (2) To interface easily with the other components of the production lines, i.e. with SAS and with the Sybase database. This types of interfaces were described in more detail in Section II.
- (3) To comply with the submission guidelines of the publisher - one of the most important issues is that the final document must contain only embedded fonts. For this purpose the function `embedFonts()` is used.
- (4) To provide means for easy text and image placement. Whenever the data are changed the document should preferably be automatically regenerated.
- (5) To use the same fonts in figure labels as in the main document - this is desirable for reasons of consistency and aesthetics. Sometimes possible to match or to approximate the document font from within the data analysis program (e.g. SAS) when the figure is saved, this would not be ideal because the document fonts themselves might not be constant and the graphics quality will never equal that of which L^AT_EX is capable.
- (6) To be easy to maintain and extend. Every year new graphs are added, other are removed and the textual explanations are rewritten to correspond to the new display.

15. A suitable tool that allows to embed the code for complete data analysis in documents is the R package **Sweave** (see Leisch, 2002). The purpose of this tool is to create dynamic reports, which can be updated automatically if data or analysis change. The necessary programming code for obtaining of the graph is contained in the master document and is written in R while the text is written in L^AT_EX. The idea behind **Sweave** is as follows: A document will be created and at certain positions in this document will be placed commands which calculate statistics, create tables or draw graphics, which will be included into the document. Whenever necessary (e.g. the input data has changed or the text of the document has changed) R will be run on the document and the output will be a ready, always up-to-date document containing all the required results, tables and graphics. The commands for R will be issued using the package **Sweave** with a syntax resembling the *Noweb* syntax (see Ramsey, 1998) in blocks (*Chunks*) which consist of three main parts: (i) begin markup '`«... »`', (ii) command part which contains normal R code and (iii) end markup '`@`'. The output report can be automatically updated if data or analysis change, which allows for truly reproducible document generation.

IV. NOWCASTING MVA FOR CROSSCOUNTRY COMPARISON

16. The Research and Statistics Branch of the United Nations Industrial Development Organization (UNIDO) is responsible for implementing the international mandate of the Organization in the field of industrial statistics. It maintains a unique industrial statistics database and updates it regularly with data collected from the national statistical offices (NSO). A separate database at macro level is also maintained primarily for compilation of statistics related to manufacturing value added (MVA) such as its growth rate and share in gross domestic product (GDP) for various countries and regions. These figures are published in the International Yearbook of Industrial Statistics and posted on the statistical pages of the UNIDO web site. For current economic analysis it is crucial that the

Yearbook presents MVA data for the most recent years. Because of a time-gap of at least one year between the latest year for which data are available and the year for which MVA data must be reported in the Yearbook, nowcasting methods are used to fill in the missing data up to the current year. For this purpose a parsimonious methodology was proposed exploiting the relationship between MVA and GDP, together with the availability of reliable estimates of GDP growth rates from external sources, to produce reliable nowcasts of MVA.

17. The standard OLS estimator may be biased because of a violation of the assumption of exogeneity of the regressors and with respect to the error term and because of the presence of outliers in the data. The time series plots of the growth of *GDP* and *MVA* indicate the presence of outliers for some countries. Fig. 3 illustrates this for the 1991-2007 data for Poland. In the transition years 1991 and 1992 the MVA and GDP growth rates are extreme. In 1991 and 1992 the MVA (GDP) growth rates equal -16.3% (-7.0%) and 80.2% (2.6%), respectively. For all other years the MVA (GDP) growth rates are between -0.6% (1.2%) and 13.8% (7.1%). The estimation of the regression models using OLS is known to be problematic in the presence of outliers. As can be seen in Fig. 4, the 1992 observation tilts the OLS slope estimate to its position and yields a distorted estimate of the regression line fitting the bulk of the data. The OLS estimator thus does not satisfy the requirement that the influence of single observations on the nowcast should be small. For this reason, we also consider a robust alternative to the OLS estimator, namely the MM estimator.



FIGURE 3. Time series plot of the 1990-2007 annual GDP and MVA growth rates for Poland - to illustrate the presence of additive outliers in the data.

18. The robust MM estimator is a two-step estimator. First, it estimates the parameter vector that minimizes the sum of the 50% smallest squared residuals. This 50% Least Trimmed Squares (LTS) estimate then serves as the starting value for the M-estimation, where a loss function is minimized that downweights outliers. The MM estimator has a high efficiency under the linear regression model with normally distributed errors. Because it is initialized at the LTS estimates, it is also highly robust to outliers (see e.g. Maronna et al., 2006, Chapter 5). In Fig. 4 the OLS and robust MM regression estimates are compared. We see that, in contrast with the OLS estimator, the robust MM estimate is rather insensitive to the outlying observations and produces an accurate fit of the bulk of the data.

19. The computations are performed using the package **robustbse** (see [Rousseeuw et al., 2009](#)) which was developed to provide “essential robust statistics” within R available in a single package and to provide tools that allow analyzing data with robust methods. This includes regression methodology including model selections and multivariate statistics and the goal is to cover the book [Maronna et al. \(2006\)](#).

20. The nowcast accuracy comparison made showed that the approach using (i) the econometric model that specifies the conditional expectation of the yearly MVA growth rate as a linear function of the contemporaneous GDP growth rate and (ii) a robust estimation method has the best performance of all considered methods. Further details about the methodology and comparison of the accuracy of the nowcasts can be found in [Boudt et al. \(2009\)](#).

V. CONCLUSION AND OUTLOOK

21. There is an increasing demand for statistical tools which combine the ease of use of the traditional software packages with the availability of the newest analytical methods which is only possible through the flexibility provided by a statistical programming language such as R. We discussed briefly the versatility of the R programming environment and how it is possible to apply this environment in the statistical offices. This was illustrated by examples from the statistical production process of UNIDO where certain steps were either migrated or newly developed in R - the areas of data integration, automatic generation of publication quality graphics for dissemination of statistical data and nowcasting methods to fill in the missing data.

22. Future development will cover survey data analysis, detection of outliers in survey data and imputation of missing values in multivariate data.

References

- Kris Boudt, Valentin Todorov, and Shyam Upadhyaya. Nowcasting manufacturing value added for cross-country comparison. *Statistical Journal of the IAOS: Journal of the International Association of Official Statistics*, 26:15–20, 2009.
- R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 2009.
- Friedrich Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 - Proceedings in computational statistics*, pages 575–580. Physica Verlag, Heidelberg, 2002.
- R. A. Maronna, D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, New York, 2006.
- Paul Murrell. *R Graphics*. Chapman and Hall/CRC, 1 edition, 2005.
- N. Ramsey. Noweb man page, 1998. [edu / nr / noweb](#). version 2. 9 a.
- Peter Rousseeuw, Christophe Croux, Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, Tobias Verbeke, and Martin Maechler. *robustbase: Basic Robust Statistics*, 2009. URL <http://CRAN.R-project.org/package=robustbase>. R package version 0.5-0-1.
- Deepayan Sarker. *Lattice: Multivariate Data Visualization with R*. Springer, New York, 2008. ISBN 978-0-387-75968-5.
- Valentin Todorov. R: An open source statistical environment, 2008. URL <http://www.unece.org/stats/documents/2008.04.msis.htm>. presented at the Meeting on the Management of Statistical

- Information Systems (MSIS 2008) (Luxembourg, 7-9 April 2008).
- Valentin Todorov and Peter Filzmoser. An object oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47, 2009. URL <http://www.jstatsoft.org/v32/i03/>.
- E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut, 2001.

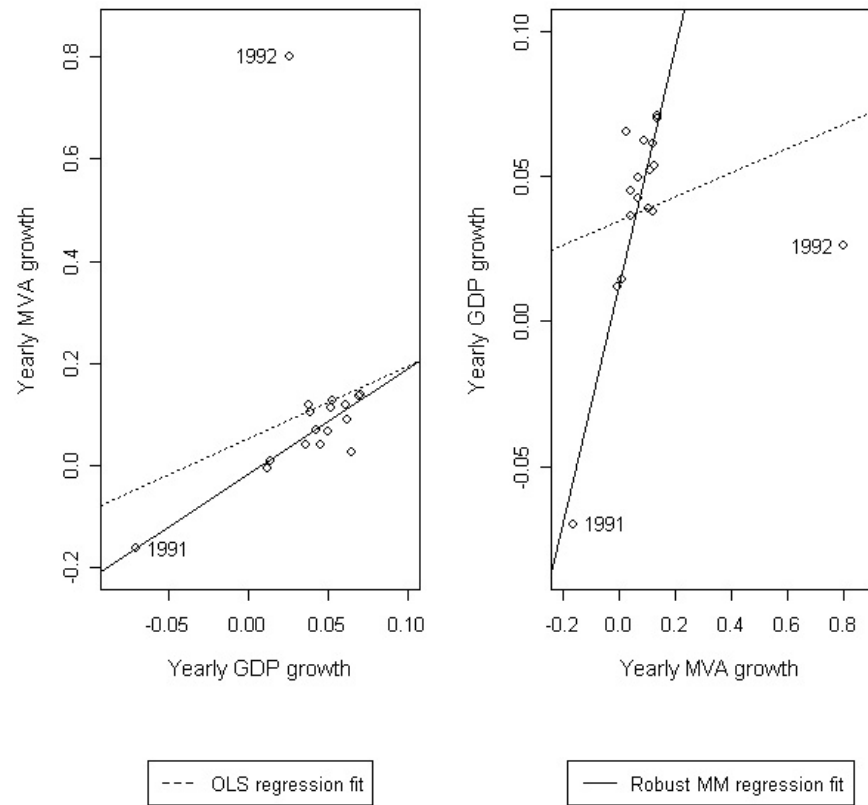


FIGURE 4. Scatter plots of the 1991-2007 annual GDP and MVA growth rates for Poland, together with the OLS and robust MM regression fit.