Computational Statistics with Spreadsheets Towards Efficiency, Reproducibility and Security

G. Aydınlı
W. Härdle

E. Neuwirth

CASE
Humboldt-Universität zu Berlin
D-10178 Berlin, Germany

Department of Statistics and DSS
Universität Wien
A-1010 Vienna, Austria

April 22, 2003

Abstract

The use of electronic spreadsheets as the primary software tool for teaching management science modeling techniques and quantitative methods in economics and finance undoubtedly played a key role in the increasing impact of quantitative lectures given in graduate programs. Researchers suggest that the ability to extract data from various sources and embed analytical decision models within larger systems are two of the most valuable skills for business students entering today's IT dominated workplace.

In this paper we will try to contribute to this evolution and furthermore want to argue in favor of spreadsheet applications as appropriate interface solution to matrix oriented statistical languages.

We present two software solutions which are aimed to combine statistical method servers and the spreadsheet application Excel. We will emphasize the productivity gain in terms of transparency and reproducibility available by combining the computational power of a statistical programming environment with the direct manipulation facilities available in spreadsheet programs like Excel. We also want to stimulate the discussion of securing the communication in such a client/server environment.

Keywords

Statistics in education and practice, Spreadsheets, Client / Server based statistical Computing, Security, XploRe, R

1 Introduction

In our understanding modern statistical analysis and efficient method proliferation require standardization, transparency, interactivity, reproducibility and security. In the remainder of this discussion we will try to argue in favor of spreadsheet applications given the mentioned requirements.

As statistics is the science of data driven quantification and decision making, tools which handle data in tables, list or spreadsheets are naturally the preferable device to tackle the problems a statistician (researcher, student, teacher) is faced.

And of course the statistician, mathematician and computer science guild produced an overwhelming variety of languages, environments and applications, to solve the real world problems and with decreasing computing costs a student today has access to computing power which a couple of years ago was reserved to specialized laboratories and data processing centers. We do not want to discuss the pros and cons of one language or environment to the other. Instead we argue that a mixture of different computing approaches might serve the aim better to bring students, teachers and researchers closer to their goal of efficient statistical analysis.

2 Spreadsheets

One of the mainstays in the ever changing information technology is the spreadsheet application, an invention by Dan Bricklin and Bob Frankston, former MIT students, who developed VisiCalc for the Apple II (Liebowitz & Margolis (1999)). The manipulation of figures and functional relations and their conversion respectively representation in charts is the main objective behind the philosophy of spreadsheets, i.e. organizing information into machine readable columns and rows.

Microsofts Excel is one of the most successful applications ever developed (Liebowitz & Margolis (1999)), backed by the increasing acceptance of Microsofts Windows operating system since its 3.0 release in the late 1980s. As one of the first spreadsheets Excel allowed the user to communicate with the application through a graphical user interface and a mouse pointing device instead of a command line syntax, seeding todays standards of pull-down menus, "drag 'n drop" functionality and mouse clicking.

The network consisting of Excel users is remarkable, ranging from home users to academic researchers. Cox (2000) reports more than 120 million licensed Microsoft Office users worldwide with estimated market share ranging from 60% to more than 90% in the relevant segment of spreadsheet applications. Especially in professional businesses Excel therefore seems to postulate a quasi standard for data storage manipulation and visualization. Frequently Excel is also used as a database front-end as it offers various data retrieval methods, e.g. reading ASCII or character files, and gets data from any database system supporting the ODBC standard. From accounting to marketing, polling to enterprise valuation

the areas of application are almost as diverse as the businesses using Excel.

The value of the spreadsheet lies in its Flexibility. It allows one to interactively manipulate data and obtain corresponding graphical representations. In other words spreadsheets offer an interaction model radically different from an "enriched" statistical language like XploRe or R.

3 Statistics and Excel

A vast literature evolved over the last decade on methods of teaching and proliferation of statistics especially in management science. A not to small portion of this literature concentrates on spreadsheets as means of teaching quantitative skills (cf. Cragg & King (1992), Carraway & Clyman (2000), **?**, Ragsdale (2001)). The suggested fields of application reach from introductory statistics, to more elaborate examples like decision support models or Monte Carlo/Markov Chain (MCMC) simulations.

With add-on modules as the *Scenario-Manager*, the *Solver* and the *Analysis Tool-Pack* and 81 (!) built-in statistical functions Microsoft enhanced Excel for statistical analysis (see Table 1).

1. one/two way ANOVA
2. Correlation
3. Covariance
4. Descriptive Statistics
5. Exponential Smoothing
6. Fourier Analysis
7. Two-Sample F-Test
8. Histogram
9. Moving Average
10. various Two Sample T-Tests
11. Random Number Generation
12. Rank and Percentile
13. Regression
14. Sampling

Table 1. Microsoft Excel XP statistical tools.

Thus Excel seems to be well suited to accomplish the usual tasks of statistical analysis given that the basic operations of Excel are known to the user. In particular these are (cf. Monka & Voß (2002)):

- Data input and storage
- Data correction
- tabular and graphical representation
- statistical calculations

- usage of Excel's statistical function

It may be not far-fetched to assume that anyone who has worked with Windows PCs is capable of using Excel and its basic features in a short amount of time. Especially in a teaching environment it should be expectable that students are familiar with Excel. Thus we might conclude that our first recommendation for efficient statistics is fulfilled: standardization. Of course only if one is willing to accept a proprietary quasi-standard which Microsoft gained through market power. Anyway since market frictions and anti-trust issues are not of concern within this research, it is enough for us to state that there is a sufficient large amount of installations and users of Excel.

Nevertheless statisticians should be careful in exploiting the statistical features of Excel. Excel has never been designed to be a full blown statistical package. Therefore we cannot expect functionality similar to professional statistical programs hence we have to accept a lack of advanced statistical methods like time series or panel data analysis or neural networks.

Features within Excel which should be used with due care in statistical analysis are (cf. Cryer (2001)):

- Computing Algorithms
- Graphics
- Treatment of Missing Data
- Random number generators
- Regression
- Help Screens

Because of the known deficiencies of Excel in the field of statistics, the literature even suggests not to use Excel at all (cf. McCullough & Wilson (1999), Knüsel (1998), Cook et al. (1999), Simonoff (2002)). We would not go that far: Since numerical and methodological impreciseness can be circumvent by redirecting the numerical computations to a statistical back end. (We will discuss involved technical aspects later.) Thus through combining the beneficial features of spreadsheets especially the direct manipulation and graphical interaction abilities together with having powerful statistical methods available directly within Excel we may turn this application into a well known and convenient frontend to high-end statistical engines. For a discussion of the benefits of using spreadsheets to convey mathematical and statistical concepts cf. Neuwirth (1996), Neuwirth (1997), Neuwirth (1998), Neuwirth (2000).

4 Drawbacks of Excel

As mentioned in the previous section, Excel reveals some serious detriments concerning numerical accuracy. In our context of efficient statistical analysis

we can furthermore conclude that transparency and reproducibility are hardly guaranteed within Excel. This accounts not only for the numerical impreciseness but even worse for the methodological approach since used algorithms are completely hidden from the user. Hence one is not able to fully understand or reproduce of what happened in the background while Excel was calculating.

To achieve reproducibility and transparency in statistical research one essential necessity is the disclosure of used methods and algorithms. Both suggested environments XploRe and R usually provide their methods as open-source. Within XploRe the so called *Quantlib* architecture acts as method-archive for logically grouped algorithms. E.g. the `Times` library contains various *Quantlets* (programs written in the XploRe programming language) for time series analysis. These can be easily viewed in a text viewer. Furthermore XploRe provides a collection of reference data sets, which can be used to verify statistical results.

Besides the openness of algorithmic solutions the documentation is crucial. Hence R and XploRe come with large documentation in various formats. The HTML based *Auto Pilot Support System* (APSS) and tutorials of XploRe provide the user with the syntactical help he needs in order to use the regarding function. Moreover these contain the exact reference of the algorithm and an example of how this method has to be used. The user has the ability to document custom algorithms with help templates. For an in-depth introduction to XploRe the reader is referred to Härdle, Hlavka & Klinke (2000) and Härdle, Klinke & Müller (1999)

5 Techniques for Statistical Add-ins

5.1 COM-Component Object Model

After giving the framework it is now time to have a closer look at the technical side of implementation. We referred to productivity gain via the mixture of computing approaches. Both tools utilize similar techniques but have different implementations concerning the architectural background of the statistical application. MD*ReX and RExcel exploit a component based approach within a client/server environment and automatization. For an exhaustive overview on component based programming cf. Griffel (1998). The architectural building block of Microsoft Office is the Component Object Model, short COM, which in the meanwhile became part of the operating system. COM aims at interoperability on component level, i.e. to enable *components* or objects to communicate with each other regardless of the language (cross-language) and the process in which the components *live* (cross-process) either on a local machine or via a network on a remote machine. The COM architecture allows to embed code libraries into any application implementing a COM client interface hence COM simplifies the creation of plug-and-play components. The outcome is readily apparent for the Office suite of applications with an extensible set of third-party tools e.g. for typesetting in LaTeX and creation of PDF-documents from Microsoft's word processor and of course the huge-amount of add-in packages for

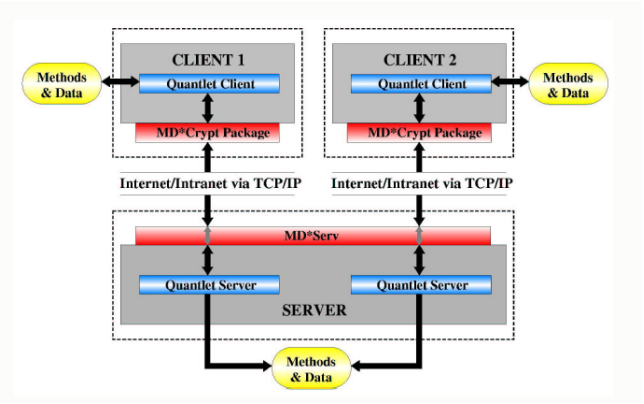

Figure 1: XploRe Client/Server Architecture

Excel. To achieve this goal one has to develop "glue" code in the client application. COM add-ins like MD*ReX allow to use one shared "glue" code library across Office applications.

5.2 Client/Server computing

Our aim is to provide an environment which on the one hand offers an intuitive and easy approach to statistical / quantitative methods and on the other hand does not give up accuracy, transparency and efficiency.

XploRe offers a highly efficient architecture to achieve our goal in the context of spreadsheets. Based on XploRe's Middleware approach we propose a multi-level architecture based on the computing paradigm of client / server applications and add-ins (see Figure 1); the COM and COM add-in approach together with XploRe's middleware implement a client server model. The statistical engine hereby functions as method repository, providing scalability of computing power for applications e.g. multi asset option hedging (Fengler & Schwendner (2003)), Value at Risk (Härdle, Kleinow & Stahl (2002)) or teaching (Aydınlı et al. (2003)). This architecture also allows for easy exchangeability of the statistical engine. Härdle, Kleinow & Tschernig (2001) show how to exploit web enabled client / server techniques in statistical computing.

Kleinow & Lehmann (2002) give an detailed description of the XploRe client/server model. Feuerhake (2001) presents the underlying communication protocol MD*Crypt, which has been implemented as custom COM version for MD*ReX. The technical background for R and the implementation in RExcel is described in Baier (2003) and Baier & Neuwirth (2003).

6 Tools

As described above the office functionality can be enhanced by problem specific add in packages. Both presented client models provide additional statistical intelligence as well as menu driven GUIs and worksheet functions.

In this setup of spreadsheet and statistical engine we differentiate two user interaction models in-sheet functions (a.k.a. worksheet functions) and menu based functionality and three user profiles Methods developer, Sophisticated methods user and Naïve user. We will try to make this approaches clear in the context of the regarding tool.

6.1 MD*ReX

MD*ReX combines Excel and XploRe filling the gap between the every day work with office applications and the need for accurate, powerful and reliable computational statistics. The tool seamlessly fits into a well known environment, has the same look and feel as MS Office applications and enables the user to combine the powerful graphical engine of Excel with the high-grade statistics skills of XploRe.

Technically speaking MD*ReX is a dynamic link library registered as in-process COM Server which is linked into the MS Office Suite. As mentioned the Java based MD*Crypt package has been implemented in a custom COM version on the Windows platform. The COM Add-in further implements `IDTExtensibility2`, an abstract interface to create add-ins for applications which support COM

The basic graphical interface is depicted in Figure 2. This toolbar can be intuitively operated by anyone who is familiar to Excel. As mentioned above it seems desirable to account for various user profiles: the methods **developer** (e.g. teacher) who needs direct access to the statistical engine. In MD*ReX this is achieved through a command line utility in the toolbar (see Figure 2). The **sophisticated methods user** (e.g. graduate Student) seeking for a macro editor to develop his own functions. In MD*ReX this is the XploRe Direct utility (see Figure 3).

And of course the **naïve user** (e.g. undergraduate student) who is accustomed to a menu driven interface with dialogues and menu options. These can be accomplished via custom menus in MD*ReX as with the Time Series module depicted in Figure 4.

Figure 4 and 5 demonstrate the interaction modes described above, namely worksheet functions and menu based functions.

6.2 RExcel

RExel offers similar functionalities (see Figure 6). Since the paradigm of a dialogue based window application is obeyed by both clients, the user can easily create interactive graphics or sophisticated statistical analyses. For further details we refer the reader to Baier & Neuwirth (2003).

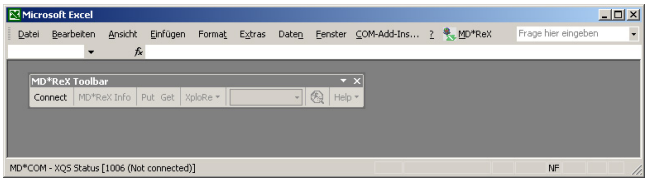

Figure 2: MD*ReX Toolbar in Excel

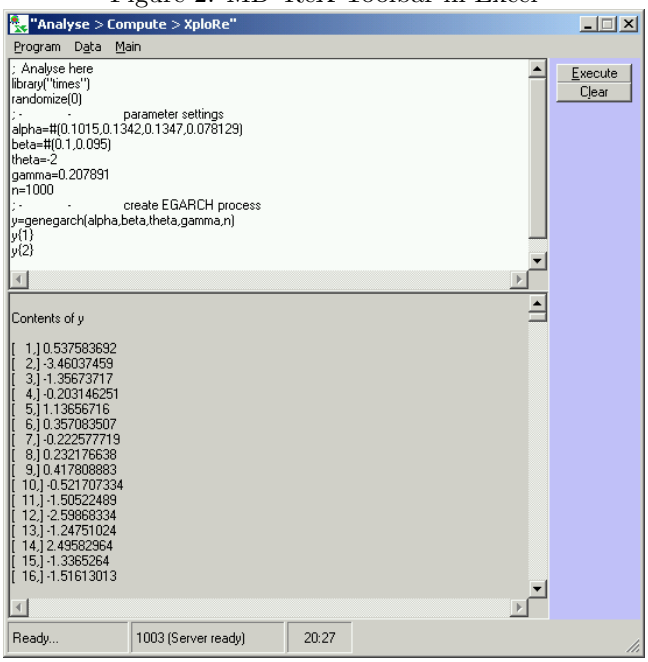

Figure 3: XploRe Direct Editor in MD*ReX

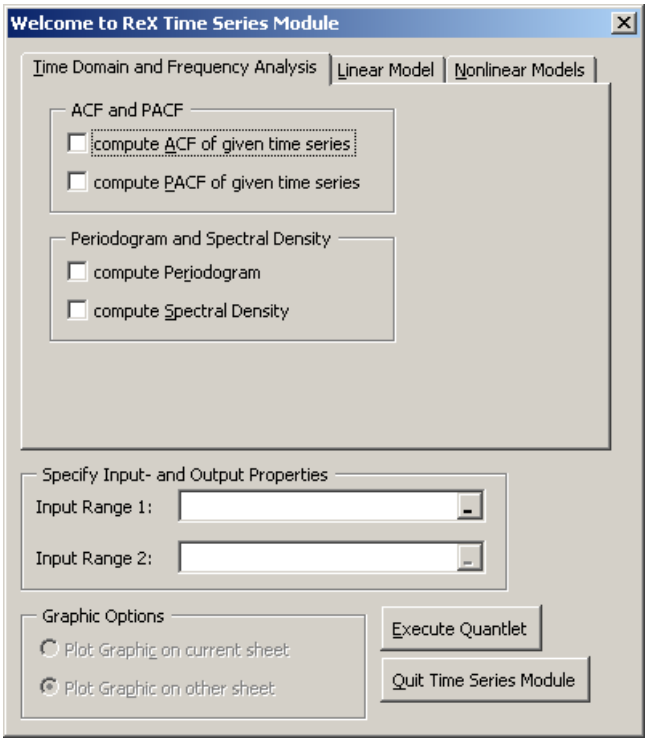

Figure 4: MD*ReX Menu for Time Series Analysis

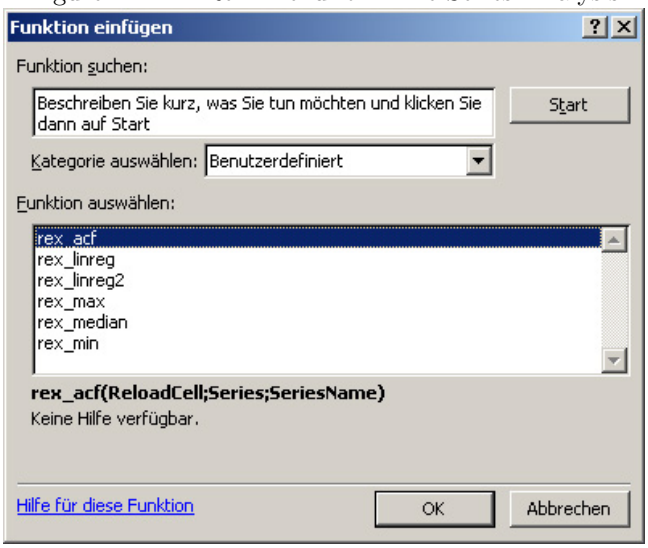

Figure 5: Worksheet Functions generated by MD*ReX

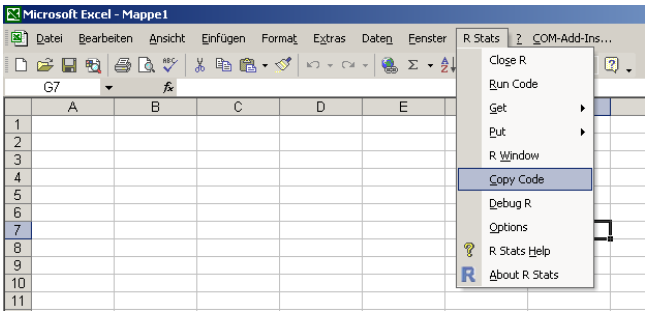

Figure 6: RExcel Toolbar

7 Applications

To better illustrate the usage it might be helpful to show some applications in both tools. The mentioned `Times` library in XploRe might be a good starting point. Assuming that a data file with the passenger volume of an airline is given as Excel worksheet we would like to apply some non-trivial seasonal time-series analysis also known as Seasonal ARIMA. Via the `Put` and `Get` menu options we are able to exchange data with the XploRe Quantlet Server. An automatic object mapping transparently shows which objects exist at the server and which Excel ranges correspond to them (see Figure??)

Issuing the following *Quantlet* at the command line or the Editor in MD*ReX returns the Autocorrelation and the partial Autocorrelationfunction of the Passenger data set

```
library("times")
x = tdiff(log(passengers))   ; first ordinary differences
x = tdiff(g,12)     ; first seasonal differences

pax_acf=acf(x,37)
pax_pacf=pacf(x,36)
```

which then can be displayed in usual Excel charts.

Another example is provided is in the context of financial statistics. In one is interested in copulas in the context of VaR analysis the following might be done to generate a 3D-plot of a given copula for a given θ.

```
library("VaR")
library("times")

theta=2
uv=#(0,0)
huv=#(0.05,0.05)
nuv=#(30,30)
```

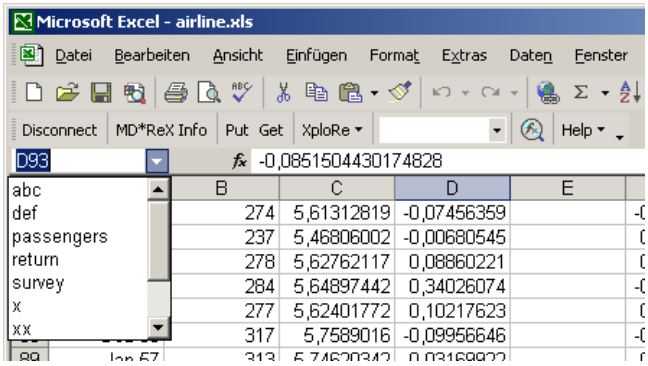

Figure 7: Name Mapping in Excel

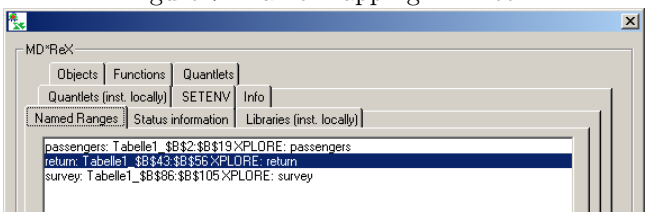

Figure 8: Name Mapping in MD*ReX

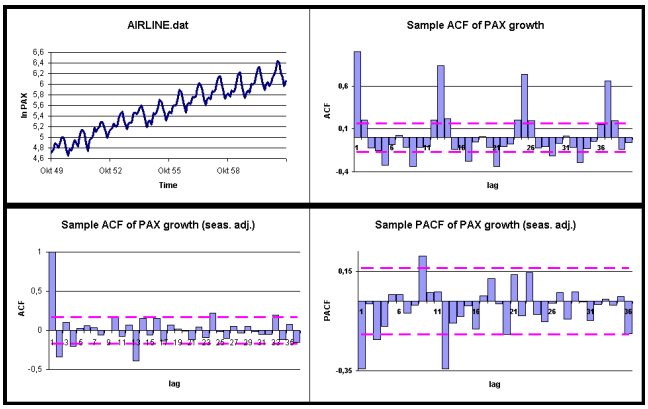

Figure 9: SARIMA Analysis of Airline.dat in MD*ReX

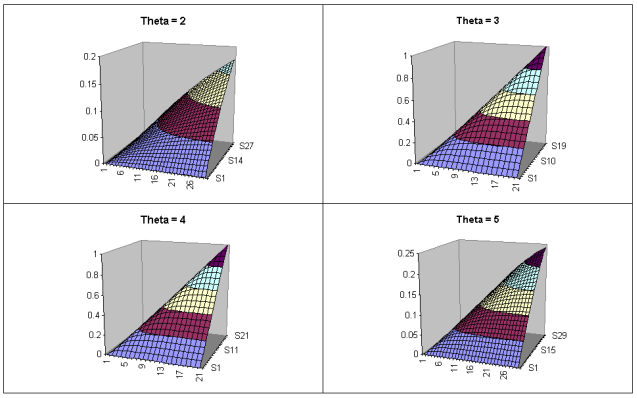

Figure 10: Gumbel-Hougaard copula for different θ

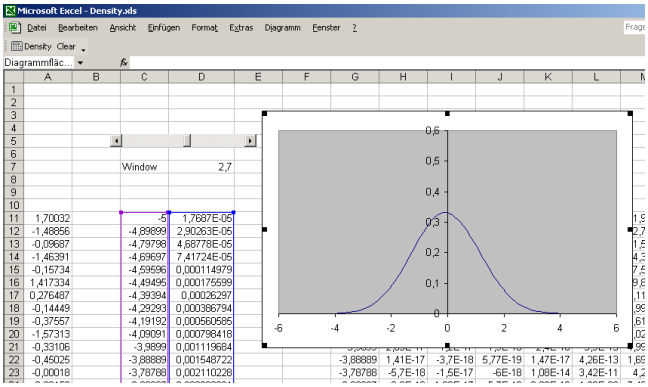

Figure 11: RExcel with slider

```
guv=grid(uv,huv,nuv)
cuv=VaRcopula(guv, theta, 0, 4)
rexcuv1=reshape(cuv,#(30,30))
rexcuv1
```

More examples of this kind are found in Aydınlı (2001).

An application providing a kernel density estimator using RExcel, where R is used to compute the density function for many different window widths and Excel is used to present a animated graph where a slider is used to control the window width is shown in Figure 11.

8 Security

Data transfer in distributed environments induces security questions. Especially the data transmission via several hosts, which is the common way in internet traffic, implies the risk of snooping. Hence we have to deal with a lack in privacy if our transmission scheme is like that. This holds true for MD*ReX which utilizes a custom TCP/IP protocol for its communication to any XploRe Server.

A remedy to this problem might be the use of encrypted data transmission. In case of MD*ReX and its underlying protocol MD*Crypt this might be implemented via the Java Cryptography Extension (JCE). This means to send or receive encrypted streams of data over either network or filesystem. The en- respectively decryption of data read or written through streams can then be implemented via key based secure socket connections. Generally one might think of an encrypted exchange of keys in the handshake process and then use symmetric keys for the input/output streams to achieve higher throughput.

The COM architecture of RExcel provides authentication but no encryption. This is due to the RPC based security mechanisms of COM. Security might be achieved via secure Virtual Private Networks.

There is currently some research conducted in securing the MD*Crypt communication between XploRe Client and XploRe server via Secure Sockets. However this is work in progress. The same is applicable to the security issues of RExcel.

9 Summary and Conclusions

We shown the possibility to integrate Excel with different statistical environments such as XploRe and R. This approach enables the user to access statistical methods under any Office application using statistical add-ins. The aim is to provide a workspace in a familiar and intuitive environment and guarantee transparency and reproducibility. The need for security and encryption is induced via the communication methods the statistical add-ins use. In case of MD*ReX the TCP/IP communication is at risk of potential snooping. The RPC based COM Security in RExcel allows for user authentication but encryption is not yet available. There is still place for future research especially concerning security and encryption issues. But the COM add-in approach seems to be a promising technology in creating a cross-language and cross-application framework for computational statistics in office applications.

References

Aydınlı, G. (2001). *Net Based Spreadsheets in Quantitative Finance*, in W. Härdle, T. Kleinow, G. Stahl: Applied Quantitative Finance, Springer Verlag, Heidelberg.

Aydınlı, G., Härdle, W., Kleinow, T. & Sofyan, H.(2001). *ReX: Linking XploRe to Excel*, Computational Statistics & Data Analysis, Vol. 12, p. 27-37.

Aydınlı, G., Härdle, W. & Rönz, B.(2003). *E-Learning / E-Teaching in Statistics*, Forthcoming IASE Satellite Conference 2003.

Baier, T. (2003). *R: Windows Component Services Integrating R and Excel on the COM layer*, DSC 2003, Working Papers (draft versions), online available at `http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/`.

Baier, T. & Neuwirth, E. (2003). *High-Level Interface Between R and Excel*, DSC 2003, Working Papers (draft versions), online available at `http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/`.

Carraway, R.L. & Clyman, D.R. (2000). *Integrating Spreadsheets into a case-based MBA Quantitative Methods Course: Real Managers make real decisions*, INFORMS Transactions on Education, 1:1(38-46).

Cook, H. R., Cox, M. G. & Dainton, M. P. (1999). *Testing Spreadsheets and other Packages in Meterology, Testing the intrinsic Functions of Excel*, Report to the National Measurement System Policy Unit, Dept. of Trade and Industry, NPL Report CISE 27/99.

Cox, B.F.(2000). *Spreadsheets: Bigger and Better*, online available at `http://http://www.informationweek.com/798/financial.htm`.

Cragg, P. B. & King, M. (1992). *A Review and Research Agenda for Spreadsheet Based Systems in End-User Computing*, mimeo.

Cryer, J. D. (2001). *Problems With Using Microsoft Excel for Statistics*, Joint Statistical Meetings, August 2001, Atlanta, GA.

Fengler, M.R. & Schwendner, P. (2003). *Correlation Risk Premia for Multi-Asset Equity Options*, SFB373, Discussion Paper, online available at `http://sfb.wiwi.hu-berlin.de/`.

Feuerhake, J. (2001). *MD*CRYPT - The XQS/XQC protocol*, online available at `http://md-crypt.com/`.

Griffel, F. (1998). *Componentware Konzepte und Techniken eines Softwareparadigmas*, Dpunkt-Verlag, Heidelberg.

Härdle, W., Kleinow, T. & Stahl, G. (2002). *Applied Quantitative Finance*, Springer Verlag, Heidelberg.

Härdle, W., Klinke, S. & Müller, M. (1999). *XploRe Learning Guide*, Springer Verlag, Heidelberg.

Härdle, W., Hlavka, Z. & Klinke, S. (2000). *XploRe Application Guide*, Springer Verlag, Heidelberg.

Härdle, W., Kleinow, T. & Tschernig, R. (2001). *Web Quantlets for Time Series Analysis*, The Annals of Mathematical Statistics, 53(1), 179-188.

Kleinow, T., & Lehman, H. (2001). *Client/Server based Statistical Computing*, forthcoming in Computational Statistics Special Issue, Springer Verlag, Heidelberg .

Knüsel, L. (1998). *On the Accuracy of Statistical Distributions in Microsoft Excel*, The Statistical Software Newsletter in Computational Statistics & Data Analysis, 26, 375-379.

Liebowitz, S.J. & Margolis, S.E. (1999). *Winners, Losers, and Microsoft, Competition and Antitrust in High Technology*, Independent Institute.

Monka, M. & Voß, W. (2002). *Statistik am PC*, Hanser Verlag, Munich.

McCullough, B.D. & Wilson, B. (1999). *On the Accuracy of Statistical Procedures in Microsoft Excel*, Computational Statistics & Data Analysis, 31, 27-37.

Neuwirth, E. (1996). *Visualizing structural and formal relationships with spreadsheets*, in DiSessa et al., *The Design of Computational Media to Support Exploratory Learning*, Springer-Verlag.

Neuwirth, E. (1997). *Spreadsheets as Tools in Mathematical Modeling and Numerical Mathematics.*, in G. Filby: *Spreadsheets in Science and Engineering*, Springer-Verlag.

Neuwirth, E. (1998). *Spreadsheets: just smart calculators or a new paradigm for thinking about mathematics structure?*, in D. Tinsley and D. Johnson (eds): *Information and Communications Technology in School Mathematics*, Chapman-Hall.

Neuwirth, E. (2000). *Spreadsheets as Tools for Statistical Computing and Statistics Education*,in J.G. Bethlehem and P.G.M. Van Der Heijden (eds): *Compstat 2000*, Proceedings in Computational Statistics, Spinger-Verlag.

Neuwirth, E. & Baier, T. (2001). *Embedding R in Standard Software, and the other way round*, Proceedings of the 2nd International Workshop on Distributed Statistical Computing, March 15-17, Vienna, in K. Hornik & F. Leisch (eds), ISSN 1609-395X.

Ragsdale, C.T. (2001). *Teaching Management Science with Spreadsheets: From Decision Models to Decision Support*, INFORMS Transactions on Education, 1:2(68-74).

Simonoff, J. S. (2002). *Statistical Analyis using Microsoft Excel*, mimeo.