



DISSERTATION

New Computational Tools and Methods for Official Statistics

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der
technischen Wissenschaften unter der Leitung von

Priv.-Doz. Dr.techn. Dipl.-Ing. Matthias Templ, Institut für Stochastik und
Wirtschaftsmathematik (E105)

eingereicht an der Technischen Universität Wien an der Fakultät für Mathematik und
Geoinformation von

Dipl.-Ing. Alexander Kowarik

Matrikelnummer 0225078

Diese Dissertation haben begutachtet:

(Priv.-Doz. Dr.techn. Dipl.-Ing. Matthias
Templ)

(A.Univ.-Prof. Mag. Dr. Andreas
Quatember)

Wien, 27.02.2015

(Dipl.-Ing. Alexander Kowarik)

Erklärung zur Verfassung der Arbeit

Dipl.-Ing. Alexander Kowarik
Eyslergasse 36, 1130 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum)

(Unterschrift Verfasser)

Acknowledgements



First and foremost I want to thank my supervisor Matthias Templ, who has been a good friend and colleague for several years, a great co-author and a superb supervisor in the course of creating this thesis. Besides, I am grateful to the other people who shared the experience of writing papers with me, especially Bernhard Meindl and Angelika Meraner. Furthermore, I would like to show my gratitude to my former boss at the Methods Unit at Statistics Austria, Alois Haslinger, who fostered research activities and shared his knowledge in the area of official statistics with me. I finish with my family, especially my wife Verena, who supports me unconditionally.

Abstract

Statistical computing plays a key role in many aspects of official statistics, e.g. statistical disclosure control, visualisation, imputation and time series analysis. The usage of open source software like R (R Development Core Team, 2014) is of growing importance due to budgetary restrictions in national statistical institutes (NSIs). In addition, software can be used by multiple organisations and users without license costs and therefore the use of R supports cooperations between NSIs, especially on an European level.

NSIs collect a huge amount of confidential data, usually financed by public funds. Therefore it is of increasing importance to release anonymized micro data back to the public and to researchers.. By including sophisticated statistical disclosure control methods in R package **sdcmicro** (Templ *et al.*, 2015, 2012b; Kowarik *et al.*, 2012), NSIs have the possibility to check the disclosure risk of their data sets and afterwards protect the observations with high disclosure risk.

Independently of the data source, it is almost always the case that missing values are included in a data set. These missing values have to be replaced by estimated values (=imputation) before it is possible to apply standard statistical methods. With the R package **VIM** (Templ *et al.*, 2011a) it is easily possible to apply a wide range of imputation methods, such as an iterative stepwise regression imputation approach (see Templ *et al.*, 2011b).

An important step in understanding a specific data set and its quality is visual analysis. With the R package **sparkTable** (Kowarik *et al.*, 2014a) tables presenting quantitative information can be enhanced by including sparklines  and sparkbars  (initially proposed by Tufte, 2001). Sparklines and sparkbars are simple, intense and illustrative graphs, small enough to fit in a single line. Therefore they can easily enrich tables and continuous texts with additional information in a comprehensive visual way.

Seasonal adjustment, a special topic of time series analysis, is of great importance in official statistics to make time-dependent data comparable between different countries or just different points in time. The R package **x12** (Kowarik and Meraner, 2014) provides an interface to the X12-ARIMA software (see e.g. Hood and Monsell, 2010). Moreover an easy to use graphical user interface is available through the R package **x12GUI** (Schopfhauser *et al.*, 2014).

A methodological and computational framework for solving all the mentioned aspects is given in this thesis.

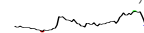

Kurzfassung

Statistische Software spielt eine wichtige Rolle in vielen Bereichen der offiziellen Statistik, wie z.B. statistische Geheimhaltung, Visualisierung, Imputation und Zeitreihenanalyse.

Die Verwendung von Open Source Software, vor allem R (R Development Core Team, 2014), ist von zunehmender Bedeutung auf Grund von budgetären Restriktionen in nationalen Statistikinstituten (NSI) und verstärkter Kooperationen zwischen NSIs, speziell auf europäischer Ebene.

NSIs sammeln eine sehr große Mengen an vertraulichen Daten und die meisten Erhebungen werden durch öffentliche Mittel finanziert. Deshalb steigt der Bedarf und die Nachfrage nach anonymisierten Mikrodatsätzen stetig. Mit den Methoden der statistischen Geheimhaltung im R Paket **sdcmicro** (Templ *et al.*, 2015, 2012b; Kowarik *et al.*, 2012), haben NSIs die Möglichkeit ihre Mikrodatsätze auf Beobachtungen mit hohem Erkennungsrisiko zu überprüfen und anschließend diese Beobachtungen zu schützen.

Unabhängig von der Datenquelle sind in jedem Datensatz fehlende Werte präsent. Da die meisten statistischen Methoden einen vollständigen Datensatz benötigen, müssen diese fehlenden Werte vor deren Anwendung imputiert werden. Mit dem R Paket **VIM** (Templ *et al.*, 2011a) können eine Vielzahl verschiedener Imputationsmethoden angewandt werden, z.B. *Iterative Stepwise Regression Imputation* (siehe Templ *et al.*, 2011b).

Visualisierung ist ein wichtiges Mittel um diverse Eigenschaften, speziell auch die Qualität der Daten zu verstehen. Mit dem R Paket **sparkTable** (Kowarik *et al.*, 2014a) können klassische Tabellen mit numerischen Werten mit Sparklines  und Sparkbars  angereichert und verbessert werden (siehe Tufte, 2001). Sparklines und Sparkbars sind einfache Grafiken mit sehr viel Information auf kleinem Platz. Sie sind klein genug um Platz in einer Zeile bzw. einer Tabellenzelle zu finden.

Saisonale Zeitreihenbereinigung als Teilgebiet der Zeitreihenanalyse ist von großer Bedeutung in der offiziellen Statistik, z.B. um zeitabhängige Daten vergleichbar zwischen verschiedenen Ländern zu machen. Das R Paket **x12** (Kowarik and Meraner, 2014) dient als Schnittstelle zu der Software X12-ARIMA (siehe z.B. Hood and Monsell, 2010). Außerdem ist eine grafische Oberfläche in dem R Paket **x12GUI** (Schopfhauser *et al.*, 2014) verfügbar.

Die methodischen und programmiertechnischen Aspekte der genannten Gebiete werden in dieser Dissertation erörtert.

Contents

1	Introduction	1
1.1	The Production of Official Statistics	2
1.2	Imputation	6
1.3	Seasonal Adjustment	7
1.4	Statistical Disclosure Control	8
1.5	Visualisation	10
2	Iterative Stepwise Regression	13
3	Imputation with the R package VIM	15
4	Seasonal Adjustment with x12 and x12GUI	17
5	Statistical Disclosure Control for Micro-data Using the R Package sdcMicro	19
6	Inclusion of New Methods Into sdcMicro	21
7	sparkTable: Generating Graphical Tables for Websites and Documents with R	23
8	Geographical Information and Traditional Plots	25
	Bibliography	27

Introduction

The term *official statistics* describes statistics produced and published by national statistical institutes (NSI) or other governmental bodies. In recent years and decades a rapid development in the area of official statistics took place - from simple stocktaking actions to a modern statistical production process.

Many different areas of society and economy are covered by official statistics, and many different methodological issues arise. The quality of statistics produced by NSIs is of great importance, because the results have significant influence on policy makers and public opinion. Figures produced by NSIs are used for benchmarking policy changes or for finding and explaining regional differences.

The *General Statistical Business Process Model* (Vale, 2013) provides a good overview of the statistical production process. It was developed by numerous international organisations and NSIs under the leadership of the United Nations Economic Commission for Europe (UNECE). A summary of the model is given in the next Section (1.1). Of course a lot of the steps involve processing data in an automatic manner and the application of advanced statistical methods.

The field of statistical computing plays a key role in providing the necessary tools for the statistical production process. Statistical computing refers to the usage of tools and methods from the field of scientific computing in the area of statistics. For over a decade the dominating development platform for statistical computing in the academic world has been R (R Development Core Team, 2014). In recent years, the usage of R outside of academic institutions increased strongly. Today, R is used in many companies and governmental institutions. R is a programming language and environment specially designed for data analysis and the production of statistical graphs, it is freely available and open source. Several NSIs are already using R in their production process and the growing interest is nourished by budgetary constraints, since most of the alternative software solutions are closed source and expensive.

The two class system in R - **S3** and **S4** - make it possible to program in an object-oriented manner. **S4** can be seen as the successor to **S3** and it provides a cleaner and

stricter interface with features such as prototypes, inheritance and validation (Chambers, 2008).

As a high-level programming language, R can be slow in some scenarios. In those cases it is possible to call functions programmed in different programming languages such as C, Fortran, Java and C++ to enhance performance. For C++ a very comfortable solution exists with the R package **Rcpp** and the corresponding C++ libraries (Eddelbuettel and François, 2011). Another R package that can help to speed up data processing, is **data.table** (Dowle *et al.*, 2013). A **data.table** is an extension of the **data.frame** class and allows for very fast aggregation, ordering and groupwise operations. Both packages - **Rcpp** and **data.table** - are used in several of the tools presented in this thesis.

In the following, several steps of the statistical production process to which this thesis contributed with methods and tools are pointed out. A broader overview of R tools suitable for the statistical production process is given by Todorov and Templ (2012) and Todorov (2010). In this publication some features of the R packages **VIM** (see Chapter 3) and **sdcMicro** (see Chapter 5) are also described and advertised.

1.1 The Production of Official Statistics

The General Statistical Business Process Model is divided into nine phases on its first level. The nine phases are described below and can be seen together with the second level of the model in Figure 1.1.

Phase 1 (Specify needs): This first phase is used to define the needs in a specific area, this is mostly driven by non-statistical reasoning. The first step in this phase is to clearly identify needs by consulting with major stakeholder. Secondly, the statistical output appropriate to meet the demand is defined and the variables and concepts are defined in an abstract manner. After checking already existing data sources, if they cover or partly cover a new topic, the business case including budgetary expectations is prepared.

Phase 2 (Design): This phase can be described as research and development phase. All necessary background information is gathered and a concept for the major parts of the statistical process is defined. A detailed definition of the statistical output has to be created and necessary new tools for dissemination have to be developed. If the output contains a time series, it is necessary to consider possible seasonal adjusted series as additional output. New tools for dealing with seasonal adjustment should be developed in this phase.

Additionally, the development of necessary IT tools for visualisation (as major part of dissemination) and disclosure control should be completed. The R packages **sparkTable** (Kowarik *et al.*, 2014a) and **sdcMicro** (Templ *et al.*, 2012b) with its methods for visualisation and disclosure control are direct results of this development process.

Collection methods and instruments need to be defined and the variables or respectively the survey questions have to be designed. In case of a sample survey, the sampling frame and the sampling scheme have to be defined.

Also in this phase the necessary methods and tools for processing the data are developed, one step of processing is imputation, therefore the R package **VIM** (Templ *et al.*, 2011a) is of importance here.

Phase 3 (Build): The results from the Design phase are now built and enhanced until they are ready for use in the production process. The specific tools are built with the help of a variety of different IT tools, among them data warehouse solutions, software for statistical analysis and visualisation and tools for dissemination, like web tools and publication software.

Phase 4 (Collect): In case of a sample survey the sampling frame is created and the sample is then selected according to the sampling design defined in Phase 1. The collection is part of this phase. The collection might be the completion of questionnaires in face-to-face interviews, telephone interviews or online surveys. However, the collection can also be to receive data for the whole population, e.g. tax data for enterprises or employment status for every person from a central employment registry.

Phase 5 (Process): In this phase statistical computing plays a key role, because a lot of the steps include the application of statistical methods. Integrating several data sources into one coherent data set is a subprocess, where merging and also statistical matching is applied.

The imputation process is part of this phase, a general introduction into this topic and a summary of the contribution of the two papers (see Chapter 3 and Chapter 2) can be found in Section 1.2.

The calculation of survey weights are of great importance in processing survey data. Calibration is one method, which is very commonly used in NSIs.

Phase 6 (Analyse): In this phase the outputs - defined in phase 2 - are produced. When time-dependent data is part of the output, seasonal adjustment is regularly applied. In Section 1.3 an introduction to these methods is given while the development of tools in this area of research is in focus in Chapter 4. The results with regards to their quality and plausibility are validated in this phase and, for example, sampling errors are computed.

Disclosure control is applied before finalising the results. This can be done directly on the micro data or on tabular data. An introduction to the field of statistical disclosure control for micro data can be found in Section 1.4, where also a description of Chapter 5 and 6, both related to the area of statistical disclosure control, is given.

Results of the statistical process are interpreted and explained in detail by subject matter specialists and statisticians, before they are prepared as final outputs.

Phase 7 (Disseminate): In this phase the statistical output is released to the public. This can be in the form of tables, articles, micro data-sets and visualisations. A short introduction to visualisation techniques in the area of official statistics and the contribution of Chapter 7 and 8 are given in Section 1.5. All downstream actions, e.g., promotional efforts and user service and support are realised in this phase.

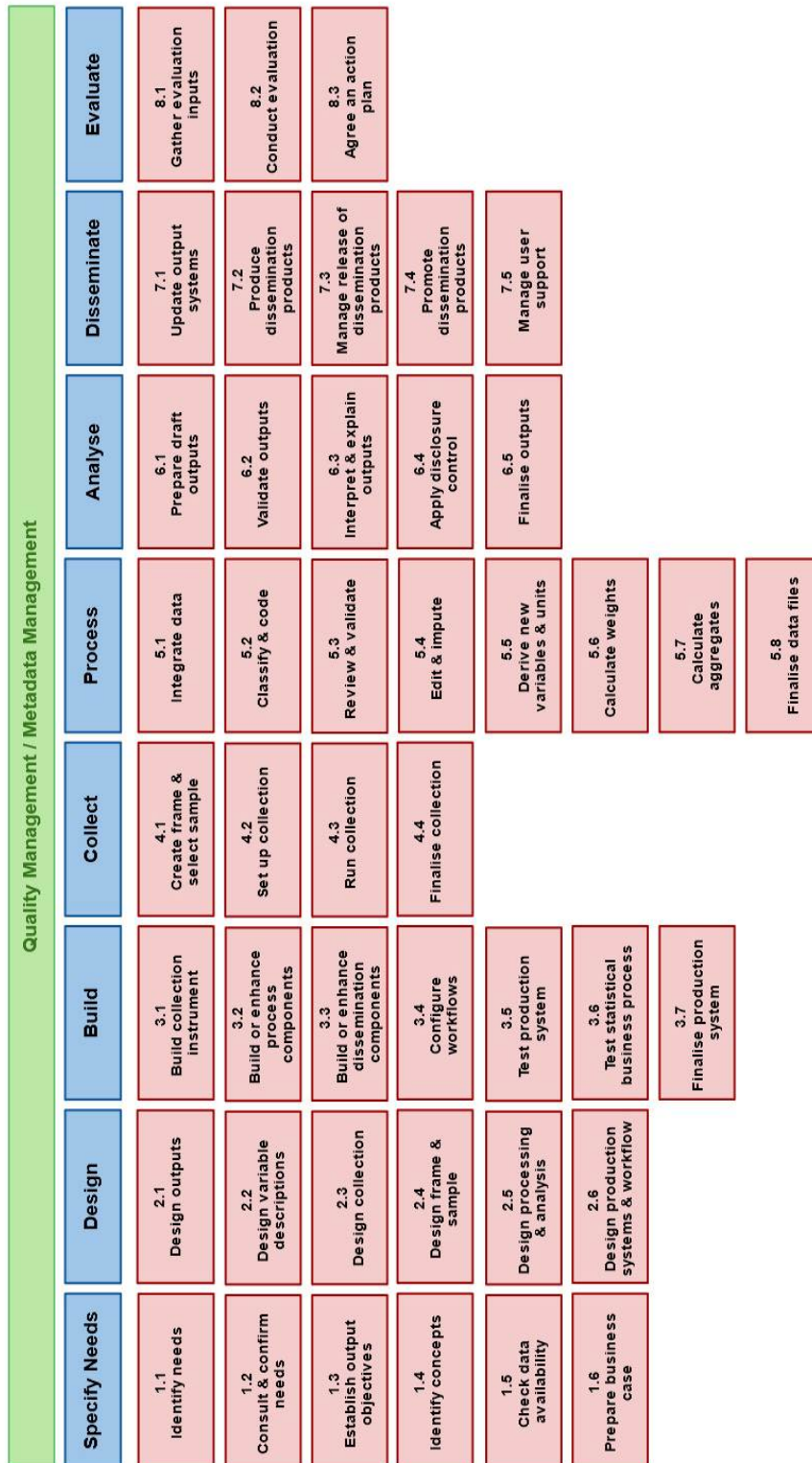


Figure 1.1: The General Statistical Business Process Model (Vale, 2013)

1.2 Imputation

Almost every data set used in official statistics includes missing items after the collection and before pre-processing the data. More generally this is also true for a lot of fields where statistical analysis are applied. Missing values can occur, among other things, due to non-response or measurement errors, where implausible values are simply deleted. The process of filling these “holes” in the data set is called imputation. The mechanism behind the missing data strongly influences *the choice of imputation method*. Little and Rubin (1987) introduced three different scenarios:

- **MCAR:** (missing completely at random), where the probability of the occurrence of a missing value is independent of observed values or the missing value itself;
- **MAR:** (missing at random), where the probability of a missing value is only dependent on the observed values;
- **MNAR:** (missing not at random), where the probability of a missing value is dependent on the missing value itself.

In the first scenario (*MCAR*) it would be possible to delete all observations with missing values without introducing a bias, but this would shrink the sample size and therefore increase the sampling error. If instead of a sample survey, a census is used, the deletion of observations with missing values would lead to undercoverage of the population. In the second (*MAR*) and third scenario (*MNAR*), the deletion of observations with missing values would lead to a biased estimate. Only in the first two scenarios it is possible to impute missing values in a reasonable manner.

A popular method, especially in previous years, has been the hot-deck imputation method, where missing values of a data set are imputed by using values of donor observations from the same data set. The name hot-deck originated from the hot deck of punch cards, in a time when punch cards were used to store the observation. Cold-deck imputation refers to using donor values from other data sets, most of the time the same survey in previous periods. An overview of different hot-deck methods, where the crucial issue is how to choose the donor value, can be found in Ford (1983). Alternative methods include model-based methods, for example to use linear regression models, estimated on the observed part of the data set to impute missing numeric values. A broad overview of imputation methods can be found in Durrant (2005) and in Chapter 3.

Imputation with the R package VIM

The paper presented in Chapter 3 is co-authored with Matthias Templ, it was submitted to the *Journal of Statistical Software* and is currently in the review process of the journal. The article presents different imputation methods, which are included in the R package **VIM** (Templ *et al.*, 2011a, 2014). The methods are presented with regards to the theory behind them and its implementation. The implemented methods are hot-deck imputation, k nearest neighbor imputation, regression imputation and iterative robust

model-based imputation. The last method is described in details in Chapter 2. The Section about k nearest neighbor imputation presents an extension of the Gower distance (Gower, 1971) with respect to the covered types of variables in the distance calculation. Using this general distance measure, variables of different scale (continuous, categorical, ordinal, count and semi-continuous) can be present in the data set. The computational speed and the imputation error of the different imputation methods is compared in a simulation study. Furthermore, all methods are explained by using close-to-reality examples. **VIM** is a widely used R package and provides a customizable and optimized implementation of several advanced imputation methods.

The thesis author developed and implemented the hot-deck and k -nearest neighbor imputation methods of the R package **VIM**, developed the extension of the Gower distance and contributed to the development and implementation of the iterative model-base stepwise regression imputation method.

Iterative Stepwise Regression Imputation Using Standard and Robust Methods

Chapter 2 presents a paper published in *Computational Statistics & Data Analysis* (Templ *et al.*, 2011b), which is co-authored with Matthias Templ and Peter Filzmoser. A new imputation algorithm called **IRMI** (Templ *et al.*, 2011b), which stands for iterative robust model-based imputation, is presented in theory and the R implementation is described. **IRMI** can handle mixed types of variables in the data set and is a data-driven automatic algorithm. In a simulation study **IRMI** is compared to a software called **IVEWARE**, which is a popular software in NSIs for imputation. One big advantage of **IRMI** in comparison with **IVEWARE** is the robustness, since usually data does not originate from a normal distribution, which is assumed in many model-based approaches. Robust methods allow deviation from the assumptions and can compensate for outlying observation in the data set. The method is now available in the R package **VIM**, see Chapter 3.

The thesis author made significant improvements to the IRMI algorithm proposed by the first author of the paper and its implementation in R. Additionally, he mainly designed and carried out the simulation studies.

1.3 Seasonal Adjustment

In official statistics, a lot of statistical outputs are produced in a regular periodic manner, therefore it is quite straightforward to produce time series for important indicators. A time series can be seen as a composition of several underlying components. Examples for such components are a seasonal component, a trend component, an outlier component, a trading day component and an irregular component. These components are often not clearly visible when looking at the time series directly. However when looking at the spectra of a time series, they show specific patterns and can be quite obvious (for details on spectral analysis, see for example Brockwell and Davis (2009) or Box *et al.* (2008)).

If the change of an indicator in comparison to the previous month or a previous year is estimated, it is often of importance to remove the seasonal component and the trading day component and compare the remaining part of the time series. With an European or even international perspective in mind, it is necessary to make data comparable. Two methodological frameworks are widely used in NSIs, **X-12-ARIMA** (see e.g., Findley *et al.*, 1998) and **TRAMO/SEATS** (see e.g., Maravall, 2003). The first one was developed by the United States Census Bureau and uses a non-parametric filter-based approach to decompose the series into its components. The second one, developed by the Bank of Spain, uses an ARIMA (see e.g., Brockwell and Davis, 2009) model-based approach for seasonal adjustment.

Seasonal Adjustment with the R packages **x12** and **x12GUI**

Chapter 4 presents a paper published in the *Journal of Statistical Software* (Kowarik *et al.*, 2014b), which is co-authored with Angelika Meraner, Matthias Templ and Daniel Schopfhauser. The presented R package **x12** (Kowarik *et al.*, 2014b; Kowarik and Meraner, 2014) is a flexible command line framework for the usage of the **X-12-ARIMA** method. Batch processing and additional analytical graphics are implemented as well. The second R package, **x12GUI** (Kowarik *et al.*, 2014b; Schopfhauser *et al.*, 2014), provides a graphical user interface, which makes the methods available to users without much R knowledge. Without deeper knowledge of the syntax and parameters of **X-12-ARIMA**, the **x12** R package offers the possibility to use R for managing time series data, for setting **X-12-ARIMA** parameters and for presenting diagnostics in a approachable manner. The application to a large number of time series is also straightforward using the R package **x12**.

The thesis author developed the R package **x12** with its abstraction layer around **X-12-ARIMA** and the graphical features. Additionally, he developed the concept for the graphical user interface in the R package **x12GUI** and is main author of the text in the paper.

1.4 Statistical Disclosure Control

Sensible information on people, enterprises or more generally on individual units is present in data from all kinds of sources. The awareness of data privacy has increased a lot in recent years. On the other hand, the demand for publicly available micro data sets is increasing strongly, because many detailed analysis are only possible by applying statistical methods directly to the micro data set. In several countries including Austria, national law prohibits the distribution of data set whenever a re-identification of a person or an enterprise is possible. A first and easy step to reduce the re-identification risk of a unit is to remove all direct identifiers, e.g., ids, names and addresses. A next step can be to achieve anonymity according to the concept of k -anonymity (Sweeney, 2002), where at least k units have the same combination of values in predefined key variables. Examples for key variables in social surveys are age, education, region, citizenship and profession.

It is not a trivial task to define the key variables for a given data sets. Common values for k are 2 and 3.

Several risk estimation methods based on the frequencies of the value combinations (cross tabulation) of the key variables exists (see e.g., Skinner and Holmes, 1998; Franconi and Poletti, 2004). The concept of measuring risk for categorical variables can be extended to numerical variables, where a risk computation is also important, because for example a very high value in *turnover* can make the re-identification of an enterprise possible. This estimation can be done model-based (see e.g., Rinott and Shlomo, 2006), so that the risk is a function of the numerical key variables.

Various methods exist for handling observations with high re-identification risk, from simple methods such as recoding (e.g., merging 5-year age classes to 10-year age classes) to more sophisticated methods such as optimal local suppression, which tries to suppress as few as possible values, to achieve k -anonymity (Samarati and Sweeney, 1998). A comparison of different disclosure control methods can be found in Domingo-Ferrer and Torra (2001) and Matthews and Harel (2011). A possible alternative could be the generation of synthetic data sets with properties close to the real data set (see e.g., Drechsler, 2011).

Statistical Disclosure Control for Micro-data Using the R Package `sdcMicro`

Chapter 5 presents a paper (Templ *et al.*, 2015), which is accepted by the *Journal of Statistical Software* and is co-authored with Matthias Templ and Bernhard Meindl. Several different disclosure control methods are described in theory, e.g., local suppression, post randomisation and micro-aggregation. The implementation of the methods is described and their application is demonstrated on a close-to-reality household survey data set. The R package `sdcMicro` (Templ, 2012) provide an easy-to-use R interface to the most popular methods in the area of statistical disclosure control. The statistical disclosure control methods and their results are available in an exploratory, interactive and user-friendly manner. The reporting facility summarizes the anonymizations and their effect on the quality and risk of the data. The package is used for the anonymization of data in several NSIs.

The thesis author contributed significantly to the implementation of the methods in the R package `sdcMicro`, especially with regards to optimising the computational performance. Furthermore he extended the micro-aggregation algorithm for the use on categorical data by using an extension of the Gower distance (Gower, 1971) and aggregation functions suitable for categorical data.

Testing of IHSN C++ Code and Inclusion of New Methods Into `sdcMicro`

Chapter 6 presents a paper published in the peer-reviewed proceedings of the *Privacy in Statistical Databases* conference (Kowarik *et al.*, 2012), which is co-authored with Matthias Templ, Bernhard Meindl, Francois Fonteneau and Bernd Prantner. The integration of existing fast C++ implementation of micro data perturbation methods in the R package `sdcMicro` is described. Furthermore, the methods are described and the com-

putational speed is compared in several simulations. The methods now run fast enough to be applied on very large data sets up to several hundred thousand observations.

The thesis author integrated the **C++** methods into **sdcMicro** using the R package **Rcpp** (Eddelbuettel and François, 2011) and further enhanced the speed of the **C++** implementation.

1.5 Visualisation

In the long history of official statistics, the dominant form of data dissemination were paper-based publications with many pages containing huge tables. Of course, nowadays the importance and the acceptance of paper publications is decreasing and simultaneously the need of visualisation techniques to present the data in a concise and comprehensible way is increasing. Data visualisation has a long history and tradition, which started at least about 300 years ago (see e.g., Friendly, 2006). Today, data visualisation is a still growing research field, where different visualisation techniques are developed and tested with principles of graphical perception in mind (see e.g., Cleveland and McGill, 1987). In several years, paper publications of statistical results will probably be abandoned and digital dissemination online will be the dominant communication channel. This medium works especially well with visualisation and it can be used to not just show static graphs, but to present information in an interactive and visual way. One small example of how the addition of a visual element can help gaining additional insights, in this case the trends of time series, is shown in the comparison between a tabular presentation of population numbers in Figure 1.2 and the same table with an additional sparkline (see Chapter 7 for details). Visual analysis is not just important in the dissemination of results, it is also of great importance during the processing stage, for example to assess the quality of imputed values (see e.g., Templ *et al.*, 2012a).

The most prominent charts, produced by NSIs, are pie charts, bar charts and line charts (see e.g., Bosch and de Jonge, 2008). Comparing results of different groups in the population or following results over time is one of the big advantage of the visual presentation of data. A geographical dimension is often present in data from official statistics, e.g., the GDP of a specific region. In such cases, presenting the data in their geographical context, for example on a map, gives additional insights (see e.g., Few, 2009).

sparkTable: Generating Graphical Tables for Websites and Documents with R

Chapter 7 presents a paper, which is submitted in revised form after major revision to the *RJournal* and is co-authored with Bernhard Meindl and Matthias Templ. It introduces a simple way to enhance classical statistical tables with spark-type graphics (see e.g., Tufte, 2001). Several different kind of graphics are available. The presented R package **sparkTable** (Kowarik *et al.*, 2014a) introduces an object-oriented framework to generate graphical tables for websites, presentations and documents in a simple and

Bevölkerung zu Jahresbeginn seit 1981 nach Geschlecht bzw. breiten Altersgruppen (Absolutwerte)

Jahr	Insgesamt	Nach Geschlecht		Männer auf 1.000 Frauen	Nach Altersgruppen			
		Männer	Frauen		0 bis 19 Jahre	20 bis 64 Jahre	65 Jahre und älter	dar.: 75 Jahre und älter
1981	7.553.326	3.570.172	3.983.154	896	2.184.224	4.212.971	1.156.131	454.278
1982	7.584.094	3.590.286	3.993.808	899	2.159.778	4.292.823	1.131.493	465.300
1983	7.564.185	3.582.589	3.981.596	900	2.115.305	4.348.057	1.100.823	473.838
1984	7.559.635	3.583.422	3.976.213	901	2.070.767	4.415.758	1.073.110	480.749
1985	7.563.233	3.588.116	3.975.117	903	2.028.352	4.465.937	1.068.944	491.279
1986	7.566.736	3.594.380	3.972.356	905	1.988.702	4.499.348	1.078.686	500.239
1987	7.572.852	3.602.199	3.970.653	907	1.950.892	4.528.383	1.093.577	508.013
1988	7.576.319	3.608.710	3.967.609	910	1.911.761	4.553.802	1.110.756	519.409
1989	7.594.315	3.623.136	3.971.179	912	1.879.112	4.589.333	1.125.870	527.740
1990	7.644.818	3.654.915	3.989.903	916	1.862.258	4.642.719	1.139.841	534.306
1991	7.710.882	3.696.200	4.014.682	921	1.856.653	4.700.847	1.153.382	526.559
1992	7.798.899	3.746.551	4.052.348	925	1.864.333	4.770.187	1.164.379	511.086
1993	7.882.519	3.793.245	4.089.274	928	1.876.578	4.831.640	1.174.301	494.349
1994	7.928.746	3.820.889	4.107.857	930	1.880.290	4.862.793	1.185.663	479.964
1995	7.943.489	3.831.200	4.112.289	932	1.875.112	4.871.503	1.196.874	481.743
1996	7.953.067	3.836.950	4.116.117	932	1.871.831	4.873.219	1.208.017	494.972
1997	7.964.966	3.844.019	4.120.947	933	1.870.818	4.877.700	1.216.448	511.436
1998	7.971.116	3.848.305	4.122.811	933	1.866.873	4.880.028	1.224.215	528.564
1999	7.982.461	3.856.029	4.126.432	934	1.862.619	4.890.127	1.229.715	545.049
2000	8.002.186	3.868.331	4.133.855	936	1.857.356	4.911.163	1.233.667	559.914
2001	8.020.946	3.881.104	4.139.842	938	1.844.074	4.938.856	1.238.016	575.493
2002	8.063.640	3.906.734	4.156.906	940	1.827.823	4.986.599	1.249.218	593.437
2003	8.100.273	3.929.599	4.170.674	942	1.819.450	5.030.344	1.250.479	601.901
2004	8.142.573	3.952.600	4.189.973	943	1.813.186	5.068.488	1.260.899	612.140
2005	8.201.359	3.984.866	4.216.493	945	1.809.717	5.083.697	1.307.945	625.028
2006	8.254.298	4.014.344	4.239.954	947	1.803.687	5.093.024	1.357.587	638.263
2007	8.282.984	4.030.062	4.252.922	948	1.790.880	5.093.505	1.398.599	648.843
2008	8.318.592	4.048.633	4.269.959	948	1.777.869	5.115.684	1.425.039	658.531
2009	8.355.260	4.068.047	4.287.213	949	1.763.948	5.140.425	1.450.887	665.415

Q: STATISTIK AUSTRIA, Statistik des Bevölkerungsstandes.- Revidierte Ergebnisse für 2002 bis 2008. Erstellt am: 27.05.2009.

Figure 1.2: Population numbers in different groups in Austria presented in a table (Source: Statistics Austria).

flexible manner. Eduard R. Tufte’s quote “Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space” is the motivation for graphical tables, since a lot of information can be put into a table. Graphical Tables make it is possible to compare and track changes in time series and show the distribution of the data set. The R package **sparkTable** provides the possibility to interactivity create a graphical table, this allows non-experts in R to create graphical tables enriched by all kind of sparklines.

The thesis author contributed significantly in the process of creating a concept for the presented work and implementing the necessary tools and additionally created and implemented the concept of geographical tables, which are an extension of the checkerplot (see Chapter 8).

Bevölkerung zu Jahresbeginn seit 1981 nach Geschlecht bzw. breiten Altersgruppen (Überblick)

	1981-2009	1981	2009	Minimum	Maximum
Insgesamt		7.5533.26	8.355.260	7.553.326	8.355.260
Männer		3.570.172	4.068.047	3.570.172	4.068.047
Frauen		3.983.154	4.287.213	3.967.609	4.287.213
Männer auf 1.000 Frauen		896	949	896	949
0-19 Jahre		2.184.224	1.763.948	1.763.948	2.184.224
20-64 Jahre		4.212.971	5.140.425	4.212.971	5.140.425
65+ Jahre		1.156.131	1.450.887	1.068.944	1.450.887
75+ Jahre		454.278	665.415	454.278	665.415

Bevölkerung zu Jahresbeginn seit 1981 nach Geschlecht bzw. breiten Altersgruppen (Absolutwerte)

Jahr	Geschlecht			Männer auf 1.000 Frauen	Altersgruppen			
	Insgesamt	Männer	Frauen		0-19 Jahre	20-64 Jahre	65+ Jahre	75+ Jahre
2009	8.355.260	4.068.047	4.287.213	949	1.763.948	5.140.425	1.450.887	665.415
2008	8.318.592	4.048.633	4.269.959	948	1.777.869	5.115.684	1.425.039	658.531
2007	8.282.984	4.030.062	4.252.922	948	1.790.880	5.093.505	1.398.599	648.843
2006	8.254.298	4.014.344	4.239.954	947	1.803.687	5.093.024	1.357.587	638.263
2005	8.201.359	3.984.866	4.216.493	945	1.809.717	5.083.697	1.307.945	625.028
2004	8.142.573	3.952.600	4.189.973	943	1.813.186	5.068.488	1.260.899	612.140
...

Figure 1.3: Population numbers in different groups in Austria with a small visual aid.

Combining Geographical Information and Traditional Plots: The Checkerplot

Chapter 8 presents a paper, which is published in the *International Journal of Geographical Information Science* and is co-authored with Matthias Templ, Beat Hulliger and Karin Fürst (Templ *et al.*, 2013). The checkerplot combines the easy readability of trellis plots with the geographical context of the information which helps users in their orientation and interpretation. This is done by displaying the plots on a grid and arranging them in an optimal manner. The grid resembles an approximation of the spatial information. A linear programming problem is formulated to minimise the distortion between the geographical coordinates and the location on the grid. The functionality for arranging the grid points and for rendering the checkerplot are implemented in the R package **sparkTable** (Kowarik *et al.*, 2014a).

The thesis author developed and implemented the presented optimisation technique using the methods of linear programming. Furthermore, he contributed to the implementation of the checkerplot in R.

Iterative Stepwise Regression Imputation Using Standard and Robust Methods

The paper was published in *Computational Statistics & Data Analysis* (Templ *et al.*, 2011b) and is co-authored with Matthias Templ and Peter Filzmoser.

CHAPTER 3

Imputation with the R package
VIM

The paper, which is co-authored with Matthias Templ, was submitted to the *Journal of Statistical Software* and is currently in the review process of the journal.

Seasonal Adjustment with the R-packages x12 and x12GUI

The paper is co-authored with Angelika Meraner, Matthias Templ and Daniel Schopfhauser and it is published in the *Journal of Statistical Software* (Kowarik *et al.*, 2014b).

Statistical Disclosure Control for Micro-data Using the R Package sdcMicro

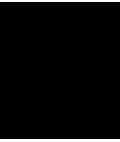
The paper, which is co-authored with Matthias Templ and Bernhard Meindl, is accepted by the *Journal of Statistical Software* (Templ *et al.*, 2015).

Testing of IHSN C++ Code and Inclusion of New Methods Into sdcMicro

The paper (co-authored with Matthias Templ, Bernhard Meindl, Francois Fonteneau and Bernd Prantner) is published in the peer-reviewed proceedings of the *Privacy in Statistical Databases* conference (Kowarik *et al.*, 2012).

sparkTable: Generating Graphical Tables for Websites and Documents with R

The paper (co-authored with Bernhard Meindl and Matthias Templ) is submitted in revised form after major revision to the *RJournal*.



Combining Geographical Information and Traditional Plots: The Checkerplot

The paper is published in *International Journal of Geographical Information Science* (Templ *et al.*, 2013) and is co-authored with Matthias Templ, Beat Hulliger and Karin Fürst.

Bibliography

- Bosch O, de Jonge E (2008). “Visualising Official Statistics.” *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*, **25**(3), 103–116.
- Box G, Jenkins G, Reinsel G (2008). *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York.
- Brockwell P, Davis R (2009). *Time Series: Theory and Methods*. Springer Series in Statistics. Springer, New York.
- Chambers J (2008). *Software for Data Analysis: Programming with R*. Statistics and Computing Series. Springer-Verlag New York.
- Cleveland W, McGill R (1987). “Graphical Perception: The Visual Decoding of Quantitative Information on Graphical Displays of Data.” *Journal of the Royal Statistical Society. Series A (General)*, **150**(3), 192–229.
- Domingo-Ferrer J, Torra V (2001). “A Quantitative Comparison of Disclosure Control Methods for Microdata.” In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pp. 111–134.
- Dowle M, Short T, Lianoglou S with contributions from Srinivasan A, Saporta R (2013). *data.table: Extension of data.frame for Fast Indexing, Fast Ordered Joins, Fast Assignment, Fast Grouping and List Columns*. R package version 1.8.10, URL <http://CRAN.R-project.org/package=data.table>.
- Drechsler J (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. Lecture Notes in Statistics. Volume 201. Springer-Verlag, New York.
- Durrant G (2005). “Imputation Methods for Handling Item-Nonresponse in the Social Sciences: a Methodological Review.” *Ncrm methods review papers*, Southampton Statistical Sciences Research Institute (S3RI), University of Southampton.
- Eddelbuettel D, François R (2011). “Rcpp: Seamless R and C++ Integration.” *Journal of Statistical Software*, **40**(8), 1–18.
- Few S (2009). “Introduction to Geographical Data Visualization.” *Visual Business Intelligence*, **6**(8), 1–11.

- Findley DF, Monsell BC, Bell WR, Otto MC, Chen BC (1998). “New Capabilities and Methods of the X-12-ARIMA Seasonal-Adjustment Program.” *Journal of Business & Economic Statistics*, **16**(2), 127–152.
- Ford B (1983). “An Overview of Hot-Deck Procedures.” *Incomplete Data in Sample Surveys*, **2**(Part IV), 185–207.
- Franconi L, Polettini S (2004). “Individual Risk Estimation in μ -Argus: a Review.” In J In: Domingo-Ferrer (ed.), *Privacy in Statistical Databases, Lecture Notes in Computer Science*, pp. 262–272. Springer.
- Friendly M (2006). “A Brief History of Data Visualization.” In C Chen, W Härdle, A Unwin (eds.), *Handbook of Computational Statistics: Data Visualization*, volume III. Springer-Verlag, Heidelberg.
- Gower JC (1971). “A General Coefficient of Similarity and Some of Its Properties.” *Biometrics*, **27**(4), 857–871.
- Hood CCH, Monsell B (2010). “Getting Started with X-12-ARIMA, Using the Command Prompt on Your PC.” *Washington, DC: US Census Bureau*.
- Kowarik A, Meindl B, Templ M (2014a). *sparkTable: Sparklines and Graphical Tables for Tex and Html*. R package version 0.11.0, URL <http://CRAN.R-project.org/package=sparkTable>.
- Kowarik A, Meraner A (2014). *x12: X12 - Wrapper Function and Structure for Batch Processing*. R package version 1.3-0, URL <http://CRAN.R-project.org/package=x12>.
- Kowarik A, Meraner A, Templ M, Schopfhauser D (2014b). “Seasonal Adjustment with the R-packages **x12** and **x12GUI**.” *Journal of Statistical Software*, **62**(2), 1–21.
- Kowarik A, Templ M, Meindl B, Fonteneau F, Prantner B (2012). “Testing of IHSN C++ Code and Inclusion of New Methods into sdcMicro.” In *Privacy in Statistical Databases*, pp. 63–77. Springer.
- Little R, Rubin D (1987). *Statistical Analysis with Missing Data*. Wiley, New York.
- Maravall A (2003). “Notes on Programs TRAMO and SEATS: Part I.” *Banco de España*.
- Matthews G, Harel O (2011). “Data Confidentiality: A Review of Methods for Statistical Disclosure Limitation and Methods for Assessing Privacy.” *Statistics Surveys*, **5**, 1–71.
- R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.

- Rinott Y, Shlomo N (2006). “A Generalized Negative Binomial Smoothing Model for Sample Disclosure Risk Estimation.” In *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer-Verlag*, pp. 82–93.
- Samarati P, Sweeney L (1998). “Protecting Privacy When Disclosing Information: k -anonymity and its Enforcement Through Generalization and Suppression.” *Technical Report SRI-CSL-98-04*, SRI International.
- Schopfhauser D, Kowarik A, Meraner A (2014). **x12GUI: X12 - Graphical User Interface**. R package version 0.10-0, URL <http://CRAN.R-project.org/package=x12GUI>.
- Skinner C, Holmes D (1998). “Estimating the Re-identification Risk Per Record in Microdata.” *Journal of Official Statistics*, **14**, 361–372.
- Sweeney L (2002). “ k -anonymity: A Model for Protecting Privacy.” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **10**(5), 557–570.
- Templ M (2012). *sdcMicro: Statistical Disclosure Control methods for the generation of public- and scientific-use files. Manual and Package*. R package version 3.1.1, URL <http://CRAN.R-project.org/package=sdcMicro>.
- Templ M, Alfons A, Filzmoser P (2012a). “Exploring Incomplete Data Using Visualization Techniques.” *Advances in Data Analysis and Classification*, **6**(1), 29–47.
- Templ M, Alfons A, Kowarik A, Prantner B (2014). **VIM: Visualization and Imputation of Missing Values**. R package version 4.2.0, URL <http://CRAN.R-project.org/package=VIM>.
- Templ M, Hulliger B, Kowarik A, Fürst K (2013). “Combining Geographical Information and Traditional Plots: the Checkerplot.” *International Journal of Geographical Information Science*, **27**(4), 685–698.
- Templ M, Kowarik A, Filzmoser P (2011a). “Imputation of Complex Data With R-Package VIM: Traditional and New Methods Based on Robust Estimation.” *Work Session on Statistical Data Editing, Conference of European Statisticians*.
- Templ M, Kowarik A, Filzmoser P (2011b). “Iterative Stepwise Regression Imputation Using Standard and Robust Methods.” *Computational Statistics & Data Analysis*, **55**(10), 2793–2806.
- Templ M, Kowarik A, Meindl B (2012b). “sdcMicro: Statistical Disclosure Control Methods for the Generation of Public-and Scientific-use Files.” *Manual and Package*.
- Templ M, Kowarik A, Meindl B (2015). “Statistical Disclosure Control for Micro-data Using the R Package **sdcMicro**.” *Journal of Statistical Software*. Accepted for publication.

- Todorov V (2010). “R in the Statistical Office: The UNIDO Experience.” *Working Paper 03/2010 1*, United Nations Industrial Development.
- Todorov V, Templ M (2012). “R in the Statistical Office: Part 2.” *Working paper. In press. 2*, United Nations Industrial Development.
- Tufte E (2001). *The Visual Display of Quantitative Information*. Graphics Press. ISBN 978-0-9613921-4-7. Second edition.
- Vale S (2013). “Generic Statistical Business Process Model (Version 5.0).” *UNECE/Eurostat*.

CURRICULUM VITAE

Personal Data

Born June 30, 1984 in Vienna Austria
Nationality Austria
Marital Status Married

Education

until 2002 Secondary school GRG XIII Wenzgasse
2002-2008 Master in technical mathematics at the Vienna University of Technology

Career History and Current Employment Status

since 2002 Web development as a freelancer
2008 - 2014 Methodological expert at Statistics Austria
since 2010 Founder and consultant of data-analysis OG
since 2014 Head of the methods unit at Statistics Austria