RatSWD Working Paper Series

www.ratswd.de

RatSWD.
German Data Forum

203

Morpheus – Remote access to micro data with a quality measure

Dr. Jörg Höhne and Julia Höninger

July 2012

Federal Ministry of Education and Research

Working Paper Series of the German Data Forum (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

Contact: S C I V E R O Press | German Data Forum (RatSWD) | Mohrenstr. 58 | 10117 Berlin | office@ratswd.de

Morpheus – Remote access to micro data with a quality measure

Dr. Jörg Höhne and Julia Höninger

State Statistical Institute Berlin-Brandenburg
Alt-Friedrichsfelde 60, 10315 Berlin, Germany

Joerg.Hoehne@statistik-bbb.de, Julia.Hoeninger@statistik-bbb.de

Abstract

Morpheus is a novel approach in providing remote access to micro data of official statistics. Researchers work on anonymous micro data files with common statistical software packages and get their results back in real time. Additionally, a measure of goodness of fit will be provided for every single result. Therefore researchers can work with the anonymous results as they can have confidence that they would have obtained the same or very similar results with the original data. All statistical analyses and all commands are allowed. Furthermore, users can browse through the anonymous data which is very helpful when developing program syntax and not possible in most other systems of remote access. Research data centres would greatly benefit from such a system as well, as the cumbersome manual disclosure control would be eliminated. All results would be safe and automatically returned to the researcher. This system would respect the special requirements to micro data access even in Germany where laws are especially strict.

1 Introduction

In many countries, National Statistical Offices and other data providers have established Research Data Centres (RDC) to provide access to micro data for the scientific community. To protect the confidentiality of micro data, data is anonymised depending on the mode of data access. Scientific-use-files which can be analysed off-site, that means outside the statistical office at the researcher's work station, have a high level of anonymity and hence a lower level of information content or a lower quality. Analysing micro data on-site, meaning the data does not leave the buildings of the statistical offices, allows providing a higher level of information content. Yet researchers then have to travel to safe centres within the statistical offices or send programme syntax to the RDC staff without having direct access to the data. This way of data access is usually referred to as remote data processing and until today involves manual actions by RDC staff while researchers have to wait several days for their results.

Scientists prefer in a best case scenario to run analysis from their own computer, get results back in real time and have data available with the original level of information content and quality. The traditional ways of data access (SUF, safe centre and remote data processing) have to make compromises on at least one of these aspects. Therefore this paper suggests a new system for data analysis called Morpheus which allows for all three features at once. This work is part of the project an "Informational Infrastructure for the E-Science Age (infinitE)" which is funded by the Federal Ministry of Education and Research (Brandt and Zwick 2009).

The general idea of Morpheus is as follows: The researcher analyses an anonymous data file that is stored on a server within the statistical office or some other data provider[1]. All types of analysis are allowed and results will be returned in real time. Additional to the results with the anonymous data, a measure of goodness of fit will be reported that allows the researcher to decide whether the interpretation made with anonymous data is the same as with the original data. As long as all measures of fit show the good quality of the anonymous data set, researchers can work with the system "Morpheus".

2 Morpheus – Advantages and Challenges

The Morpheus system is a new approach in providing access to micro data and returning results in real time. Therefore, it impacts the whole process of data access. In this article we want to demonstrate that this system would be of great advantage to both users and data providers. In this paper the benefits for both groups of stakeholders will be highlighted and the challenges will be named. To do so, it will first be described how the system works in general.

[1] The term "statistical office" will be used synonymously for any potential data provider.

Morpheus consists of three components as depicted in graph 1. In a first step, a data user works on an anonymous data file. The data file will be stored on a server, and he can work on the server with one of the common statistical packages (SPSS, SAS, Stata or R). To simplify the discussion in the following, we focus on the program Stata, as was done with the initial development of a prototype. The user should ideally have a normal access to the statistical package with all functionalities available that the program offers, that is especially with an enabled possibility to view the micro data. As the micro data file is anonymous, there are no concerns of data confidentiality. Furthermore there should be no additional rules apart from the established ones when working in a safe centre (as is best practice see Office for National Statistics 2008). Syntax should be well documented and counts have to be produced for every analysis. Working with the anonymous file should be comfortable for the data analyst. All analyses are allowed, as the data is anonymous and would in other circumstances even be admitted for off site use. No analysis can possibly constitute a disclosure risk.

The second component in Morpheus is the corresponding original micro data file that is stored on the server as well. Any analysis executed on the anonymous file runs concurrently on the original data file. However, the user of the Morpheus system does not see the original results, they are only used as an input for the third component in Morpheus. While the user sees the results from the anonymous file, he will be given a measure of goodness of fit, a quality indicator for his anonymous results. This indicator will be the difference between the result with the anonymous data and the original data. To avoid a direct inference of the original result only the absolute deviation will be given. Section 3 will deal with all possible disclosure risks arising from supplementing the anonymous risk free results with the quality indicator.

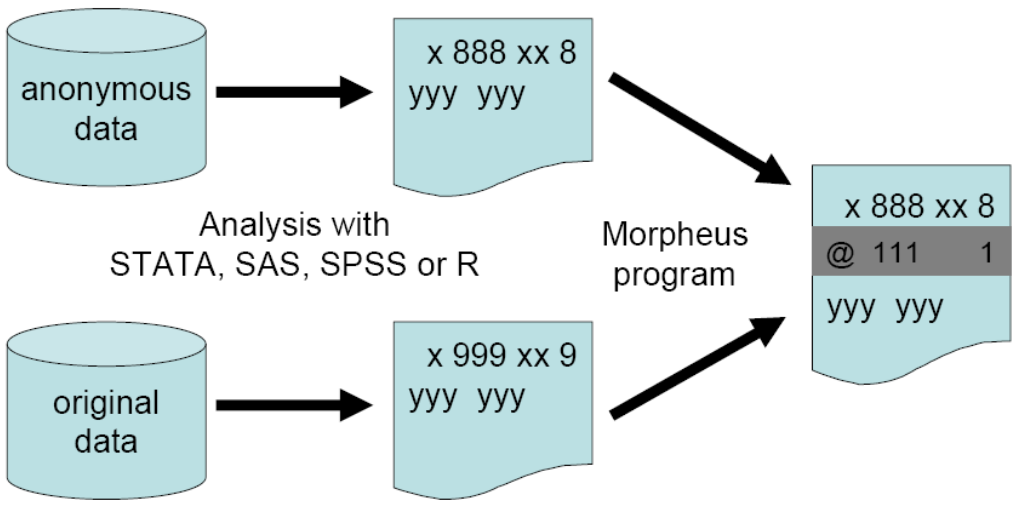

Fig 1. Overview of Morpheus, own illustration.

The quality indicator is calculated for any statistical result that the user is requesting from the statistical software. It is written in a separate line underneath the numerical result. The new line is marked by an @-sign on the first position as this is a sign that is typically not part of the programming language or result display. Additionally, the line will be colored in a grey shade.

It is planned that Morpheus is used for intermediate results and that researchers can request results based on the original data set for publications that will still be checked manually. This will increase the acceptance of Morpheus and still reduce the burden of manual work in the RDCs drastically as in a typical research project the majority of results are intermediate results that are not included in the final publication.

2.1 Advantages for users

The Morpheus system is very user-friendly, as the user receives results from his statistical analysis in real time. One of the most powerful features of Morpheus is that all statistical analysis and commands are allowed. This is not standard in remote data access as most systems either only allow or exclude a list of commands (O'Keefe et al. 2009, Lucero et al. 2009, National Center for Health Statistics 2010). Most researchers combine descriptive and inference statistics. Usually they do not calculate confidence intervals for descriptive statistics, neither for sample surveys nor full surveys. In this aspect the suggested system differs substantially from the verification servers proposed by Oganian et al. (2009). With Morpheus no amendments in the familiar way of working with statistical programs are necessary.

Furthermore, Morpheus offers many advantages over the established ways of data access offered by most RDCs today. When using remote data processing, users cannot have a look at the micro data and usually have to wait several days until their program is started by a member of RDC staff and checked manually for disclosure risks. With Morpheus users do not have to wait. Using data access in safe centres includes high travel and subsistence costs and even in safe centers, data still has to be anonymised to a lower degree. With scientific-use-files analysts always have to subordinate that the data providers did a good job when anonymising the micro data, while having no means of verifying that the anonymisation methods have been applied correctly (Alexander et al. 2010). With Morpheus no travel costs have to be invested. The data user can have a look at the micro data in the data browser to see the structure of the data. And he will get the results back in real time with an adequate measure whether the results can be trusted. Morpheus can therefore also serve as a tool to increase the confidence in anonymisation methods (Oganian et al. 2009).

Users can work with these results as long as the quality indicators are good. This decision will be left to the user: he can decide whether the indicators are sufficiently good or whether he prefers waiting for the manual

disclosure control on the original results. As an example consider a researcher who calculates the rate of change in some variable, for example income. When income increased by 10% (with a quality measure of ±2% points), he can confidently interpret that the income increased. On the other hand, if the output shows an increase of 10% (with a quality measure of ±20% points) no definite conclusion can be drawn and the researcher will probably not accept Morpheus for his study.

A disadvantage for users is the unfamiliar way of presentation of results. In the common log files produced by Stata, the additional grey lines marked with an @-sign are at first unusual. However, we are confident that users will get used to the new standard as it offers huge advantages over current methods.

Morpheus could be used by all data providers. In Germany it would constitute a notable step forward as data protection laws are especially strict. It has been argued that even displaying micro data on a monitor screen in a modus of remote desktop is a transmission of data (according to juridical expertises presented at a meeting of the "Working Group on Future Data Access" by the German Data Forum (RatSWD) on February 17th 2010). Furthermore only output that provides a certain basic level of anonymity can be displayed on a monitor. The whole process has to secure that no violation is possible.

2.2 Advantages and challenges for data providers

The big advantage for data providers is that researchers can work for the overwhelming part of their analyses independently with the Morpheus system. No manual disclosure control is necessary. Currently it is planned that for the very final results that are published, results based on the original data would be checked manually and sent to the researcher. However, this is not necessarily the way it has to be.

The biggest disadvantage is that anonymous data files for every statistic have to be created. It would be an investment to anonymise each statistic that statistical offices or other data providers offer access to. Yet, usually it is possible that anonymisation methods applied and the corresponding parameters used can be re-used when the next wave of a survey or statistic is to be included into Morpheus. Most data providers are already active in finding the best anonymisation methods for their data as they usually are also offering scientific-use-files or data structure files that are needed for remote data processing. Morpheus works independently of the anonymisation method used in creating the anonymous data file. The only condition on the anonymisation technique is that the dimensions have to be preserved, i.e. the number of observations and the categories in categorical variables. Therefore, coarsening is ruled out as anonymisation method, but for example swapping, stochastic noise or imputation (for an overview see Hundepool et al. 2010) can be applied.

3 Stochastic modification of the distance

With Morpheus, all analyses are allowed as they are executed on an anonymous data file in the first instance. On the other hand, the additional information of the goodness of fit could lead to new disclosure risks. Therefore it has been analyzed whether the additional information constitutes new disclosure risks.

As original results might disclose confidential information, it has to be prevented that users can calculate the exact value of the original result. To avoid the possibility of computing the true result with the anonymous result and the distance between true and anonymous result, we do not publish whether the original result is lower or greater than the result based on anonymous data. Adding or subtracting the distance from the anonymous result that is published leads to two potential original results. To increase ambiguity and uncertainty, a stochastic modification of the absolute distance is suggested.

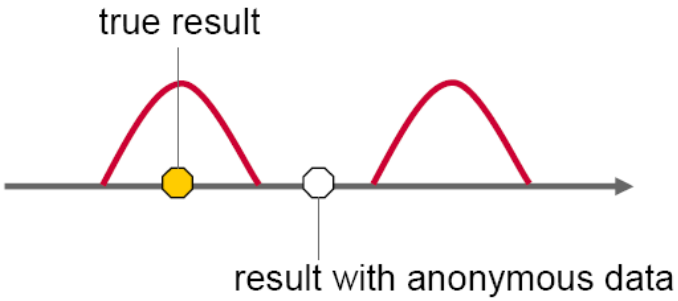

Fig 2a. Variant of modifying the distance stochastically: unbiased point estimate, own illustration.

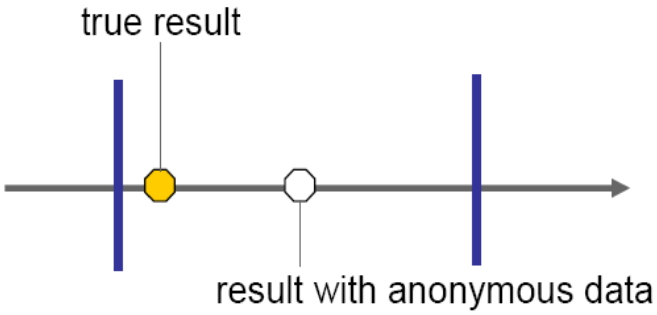

Fig 2b. Variant of modifying the distance stochastically: maximum distance, own illustration.

At first it was planned to multiply the absolute distance with a stochastic factor. One could draw a stochastic value u1 from a normal distribution with expectation 1 and variance σ as in equation 1:

$$d' = d * u_1 \quad \text{where } u_1 \sim N(1, \sigma) \tag{1}$$

This way a single value for the distance could be displayed. It will be unbiased as in half of the cases the deviation is increased and in the other half diminished. The expectation value would yield the distance measured between anonymous and original result. This variant of stochastic modification is displayed in Graph 2a. Yet this variant has a significant disadvantage: Every time the stochastic modification reduces the measured deviation, the anonymous result will appear to be of a better quality than it really is. This is unacceptable as users can no longer correctly identify whether working with the anonymous file is equivalent to the original data.
Therefore another way of stochastic modification of the absolute distance is suggested. The distance is multiplied with a stochastic factor u2 drawn from a uniform distribution on the interval [1; x] where x>1.

$$d' = d * u_2 \quad \text{where } u_2 \sim U[1; x] \tag{2}$$

The exact value of x should not be published as it could be used by data intruders. The distance between the original and the anonymous result will remain unaltered or be increased. The new measure of quality can be interpreted as the maximum distance of the original result from the anonymous result. This variant is displayed in Graph 2b. The user will be given the anonymous result and the maximum distance, depicted as two blue lines as the user does not know whether the original result is lower or greater than the anonymous one. The original result may possibly be any point in the interval set by the two blue lines.

An important aspect in generating the stochastic factor is that the factor is always the same as long the same analysis is repeated at different points in time or with different commands. Every time the same type of analysis is repeated over the same group of observations, the stochastic modification has to be implemented with the same stochastic factor to avoid that a sufficient repetition of the analyses reveals the underlying distribution of stochastic factors.

The stochastic modification shall also serve to avoid that logical restrictions can be used to infer the direction of the deviation. One could think of variables that can never be negative as for example number of employees. Is the distance greater than the absolute value of the anonymous result, one could think that the original result must be bigger than the anonymous one. Yet this logical deduction might be a result of the stochastic modification of the distance measure and the true result is indeed smaller than the anonymous one. Hence the stochastic modification of the absolute deviation between anonymous and original result protects the anonymity of the results and the micro data used in the analysis.

4 Technical Implementation

A prototype of Morpheus that can process Stata programs has already been developed. This first version of Morpheus produces output with the anonymous results and the quality measure in a separate row underneath the results. The newly inserted line is marked by an @ and has a grey background. Morpheus output can be read like normal Stata output while the quality measure can be read as follows: the result with the original data deviates at most by the amount of the quality measure from the anonymous result.

```
. tabstat expshare2000, stats(N mean sd p25 p50 p75)
    variable |         N      mean        sd       p25       p50       p75
-------------+--------------------------------------------------------------
 expshare2000 |     48305  14.71019  22.10718         0  1.776615  22.77032
@             |         0   0.06056   0.22858         0  0.011463   0.11264
-------------------------------------------------------------------------------
```

Fig 3a. Example of a Morpheus output - descriptives, own illustration.

Graph 3 displays an extract of an example output. Graph 3a shows some descriptives on the variable export share in the year 2000. Graph 3b shows the results of a fixed effects regression of labor productivity in logs on an export dummy, firm size measured as workers and workers squared and a dummy for human capital intensity. The regression is inspired by work on determinants of export activity in the manufacturing sector in Germany by Professor Joachim Wagner (i.e. Fryges and Wagner 2008).

```
.   xreg lnapro export pers perssq hc, fe r
-------------------------------------------------------------------------------
             |               Robust
      lnapro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
@            |                                                  0
-------------+-----------------------------------------------------------------
      export |   .0882477   .0076721    11.50   0.000     .0732102    .1032851
@            |  0.0002211  0.0000887     0.15       0    0.0004296   0.0000251
        pers |  -.0000162   2.99e-06    -5.42   0.000     -.000022   -.0000103
@            |  0.0000066   9.11E-07     3.32   0.010     0.000004   0.0000089
      perssq |   1.37e-11   3.28e-12     4.17   0.000     7.24e-12    2.01e-11
@            |   1.85E-12    2.4E-12      2.1   0.032      7.1E-12    3.86E-12
          hc |   .0001662   .0000164    10.15   0.000     .0001341    .0001983
@            |  0.0000163  0.0000040     2.49       0    0.0000235   0.0000068
       _cons |   8.696176   .0413789   210.16   0.000     8.615073    8.777278
@            |   0.025095   0.0081233    35.25       0    0.011238    0.036079
-------------+-----------------------------------------------------------------
```

Fig 3b. Example of a Morpheus output - regression, own illustration.

The quality measure is for all results quite well: the descriptive anonymous statistics deviate minimally from the corresponding original results. The original mean and the percentiles deviate by less than 1% from the anonymous results. In the regression the same levels of significance and the same order of magnitude in the coefficients are displayed for almost all regressors. All exogenous variables are significant at a 1% significance level both in the anonymous and the original data, except the regressor workers squared that is only significant at the 5% significance level in the original data. The magnitudes of the coefficients are well preserved, changes in signs do not occur. A researcher receiving these results back, could develop his analysis and the specifications of his inferential analysis with Morpheus and even write a draft of a paper based on these unambiguous intermediate Morpheus results.

5 Transferability to Other Statistical Programs

The most popular programs for micro data analysis in Research Data Centres in most countries are SPSS, SAS, Stata and R. While focusing on the statistical package Stata, output from the statistical programs SPSS, SAS and R can generally be processed as well. SPSS output files have to be converted to ASCII files.

6 Conclusion

This paper demonstrated that Morpheus is a new innovative way of remote micro data access. The big advantages for users are the return of their statistical results in real time, the possibility to have a look at the micro data and the fact that all commands and statistical analysis are permitted. To enjoy these benefits, the analyses are executed on anonymous micro data but the researcher receives a measure of quality for each statistical result.

Data providers have to invest to generate an anonymous file for each micro data set. Yet usually, data providers possess profound knowledge in anonymisation techniques as they generally offer some data sets as scientific-use-files or generate dummy files to help developing program syntax for remote processing.

What is left to be done is to develop an access system. A server infrastructure with an interface that handles jobs automatically and can securely identify registered users is needed. However, the technical solution can be adapted from existing systems like LISSY (Coder and Cigrang 2003). Morpheus can be installed on a web server that can be accessed via remote desktop and secured through certificates and passwords or smart cards. Even computing time does not necessarily have to double when powerful servers are used that can manage two programs at once.

Even though a first prototype of Morpheus is already developed, not all details are studied exhaustively yet and Morpheus is work in progress. One thing left to examine is for example the question: when does a distance of 0 constitute a risk and how could it be modified stochastically?

References

Alexander, J.T., M. Davern and B. Stevenson (2010): Inaccurate Age and Sex Data in the Census PUMS Files: Evidence and Implications. NBER Working Paper 15703. Cambridge, Massachusetts.

Brandt, M. and M. Zwick (2009): Improvement of the Informational Infrastructure – on the Way to Remote Data Access in Germany. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Bilbao, Spain, 2-4 December 2009, Working Paper No. 16.

Coder, J. and M. Cigrang (2003): LISSY Remote Access System, Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Bilbao, Spain, 2-4 December 2009, Working Paper No. 7.

Fryges, H. and J. Wagner (2008): Exports and Productivity Growth – First Evidence from a Continuous Treatment Approach. Review of World Economics 144 (2008), 4, 695-722.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Naylor, E. Schulte Nordholt, G. Seri and P.-P. de Wolf (2010): Handbook on Statistical Disclosure Control. Version 1.2, available from http://neon.vb.cbs.nl/casc/ SDC_Handbook.pdf.

Lucero, J., Singh, L. and L. Zayatz (2009): Recent Work on the Microdata Analysis System at the Census Bureau. Research Report Series (Statistics #2009-09) Statistical Research Division, U.S. Census Bureau, Washington, D.C.

National Center for Health Statistics - Research Data Center (2010): Disclosure Manual. Website accessed 14th February 2011, www.cdc.gov/rdc/Data/B4/Dis-closureManual.pdf and www.cdc.gov/rdc/Data/B2/SASSUDAAN Restrictions.pdf.

O'Keefe, C.M. and N.M. Good (2009): Regression Output from a Remote Analysis Server. Data & Knowledge Engineering 68: 1175-1186.

Office for National Statistics (2008): How to access the VML (Virtual Microdata Laboratory). Website accessed 14th February 2011, http://www.ons.gov.uk/about/who-we-are/our-services/vml/accessing-the-vml/how-to-access-the-vml/index.html.

Oganian, A., Reiter, J. P. and Karr, A. F. (2009): Verification servers: Enabling analysts to assess the quality of inferences from public use data. Computational Statistics and Data Analysis 53(4): 1475-1482.