Höhne, Jörg; Höninger, Julia

Working Paper

Das Verfahren Morpheus – Auf dem Weg zu Remote Access

RatSWD Working Paper, No. 205

Provided in Cooperation with:
German Data Forum (RatSWD)

Suggested Citation: Höhne, Jörg; Höninger, Julia (2012) : Das Verfahren Morpheus – Auf dem Weg zu Remote Access, RatSWD Working Paper, No. 205, Rat für Sozial- und Wirtschaftsdaten (RatSWD), Berlin

This Version is available at:
https://hdl.handle.net/10419/75358

RatSWD Working Paper Series

www.ratswd.de

RatSWD.
Rat für Sozial- und WirtschaftsDaten

205

Das Verfahren Morpheus – Auf dem Weg zu Remote Access

Dr. Jörg Höhne and Julia Höninger

September 2012

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Working Paper Series des Rates für Sozial- und Wirtschaftsdaten (RatSWD)

Die *RatSWD Working Papers* Reihe startete Ende 2007. Seit 2009 werden in dieser Publikationsreihe nur noch konzeptionelle und historische Arbeiten, die sich mit der Gestaltung der statistischen Infrastruktur und der Forschungsinfrastruktur in den Sozial-, Verhaltens- und Wirtschaftswissenschaften beschäftigen, publiziert. Dies sind insbesondere Papiere zur Gestaltung der Amtlichen Statistik, der Ressortforschung und der akademisch getragenen Forschungsinfrastruktur sowie Beiträge, die Arbeit des RatSWD selbst betreffend. Auch Papiere, die sich auf die oben genannten Bereiche außerhalb Deutschlands und auf supranationale Aspekte beziehen, sind besonders willkommen.

RatSWD Working Papers sind nicht-exklusiv, d. h. einer Veröffentlichung an anderen Orten steht nichts im Wege. Alle Arbeiten können und sollen auch in fachlich, institutionell und örtlich spezialisierten Reihen erscheinen. Die *RatSWD Working Papers* können nicht über den Buchhandel, sondern nur online über den RatSWD bezogen werden.

Um nicht deutsch sprechenden Nutzer/innen die Arbeit mit der neuen Reihe zu erleichtern, sind auf den englischen Internetseiten der *RatSWD Working Papers* nur die englischsprachigen Papers zu finden, auf den deutschen Seiten werden alle Nummern der Reihe chronologisch geordnet aufgelistet.

Einige ursprünglich in der *RatSWD Working Papers* Reihe erschienenen empirischen Forschungsarbeiten sind ab 2009 in der RatSWD Research Notes Reihe zu finden.

Die Inhalte der *RatSWD Working Papers* stellen ausdrücklich die Meinung der jeweiligen Autor/innen dar und nicht die des RatSWD.

Kontakt: S C I V E R O Verlag | Rat für Sozial- und Wirtschaftsdaten | Mohrenstr. 58 | 10117 Berlin | office@ratswd.de

Das Verfahren Morpheus – Auf dem Weg zu Remote Access

Dr. Jörg Höhne und Julia Höninger

Amt für Statistik Berlin-Brandenburg, Alt-Friedrichsfelde 60, 10315 Berlin

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

Abstract

Morpheus ist ein neuartiger Ansatz, einen echten Fernzugriff auf Mikrodaten der amtlichen Statistik zu gewähren. Wissenschaftler werten mit den üblichen Statistik-Softwarepaketen einen anonymen Datensatz aus und erhalten ihre Ergebnisse in Echtzeit zurück. Zusätzlich erhalten sie zu jedem einzelnen Ergebnis ein Gütemaß. Obwohl die Wissenschaftler mit den anonymen Ergebnissen arbeiten, können sie dadurch sicher sein, dass sie das gleiche oder ein sehr ähnliches Ergebnis mit den Originaldaten erhalten hätten. Alle statistischen Analysen sind erlaubt. Darüber hinaus können sich die Wissenschaftler d ie anonymen Daten anschauen. Dies ist sehr hilfreich, wenn man Programme entwickelt und es ist in den meisten anderen Remote-Access-Systemen bisher nicht möglich.
Forschungsdatenzentren würden von solch einem System ebenfalls in großem Maße profitieren, da die mühsame manuelle Geheimhaltungsprüfung entfallen würde. Alle Ergebnisse wären sicher und werden automatisch dem Wissenschaftler zugesandt. Dieses System würde die strengen gesetzlichen Datenschutzvorgaben für den Zugang zu Mikrodaten in Deutschland erfüllen.

1 Einleitung

In vielen Ländern haben die Statistischen Ämter und andere Datenproduzenten Forschungsdatenzentren (FDZ) eingerichtet, um der Wissenschaft einen Zugang zu Mikrodaten zu gewähren. Um den Datenschutz der Mikrodaten zu gewahren, werden die Daten je nach Zugangsweg unterschiedlich anonymisiert. Scientific-Use-Files (SUF) können von Wissenschaftlern an ihrem Arbeitsplatz analysiert werden. Wenn die Nutzung außerhalb der Räume der amtlichen Statistik (Off-Site) stattfindet, weisen die Daten ein höheres Anonymisierungsniveau und dadurch ein niedrigeres Informationspotenzial oder eine niedrigere Informationsqualität auf. Werden die Daten innerhalb der amtlichen Statistik ausgewertet und verlassen die Mikrodaten die Ämter somit nicht, kann ein höheres Informationsniveau bereitgestellt werden. Allerdings müssen die Wissenschaftler dann entweder zum Gastwissenschaftler-arbeitsplatz (GWAP) reisen oder ihre Analyseprogramme an die FDZ-Mitarbeiter senden, ohne dass sie selbst direkten Kontakt zu den Daten erhalten. Dieser zuletzt genannte Datenzugangsweg wird kontrollierte Datenfernverarbeitung (remote data processing, KDFV) genannt und ist noch kein Remote Access, da er weiterhin manueller Eingriffe der FDZ-Mitarbeiter bedarf und die Wissenschaftler auf die Freigabe der Analyseergebnisse warten müssen.

Wissenschaftler in ihrem Wunschszenario wollen Analysen gerne von ihrem eigenen Computer anstoßen und die Ergebnisse in Echtzeit zurückbekommen, wobei sie auf Daten mit dem höchsten Informationsgehalt zurückgreifen. Bei den traditionellen Zugangswegen (SUF, GWAP, KDFV) muss man bei mindestens einem dieser Aspekte Kompromisse eingehen. Daher wird in diesem Beitrag ein neues System namens Morpheus vorgeschlagen, das alle drei Eigenschaften in einem Zugangsweg kombiniert. Diese Arbeit entstand in dem Projekt “Eine informationelle Infrastruktur für das E-Science Age (infinitE)”, das vom Bundesministerium für Bildung und Forschung gefördert wird (Brandt und Zwick 2009). Projektpartner sind das Statistische Bundesamt, das Institut für Arbeitsmarkt- und Berufsforschung, das Institut für Angewandte Wirtschaftsforschung, das Hessische Statistische Landesamt und das Amt für Statistik Berlin-Brandenburg.

Die generelle Idee von Morpheus ist folgende: Der Wissenschaftler analysiert einen anonymen Datensatz, der auf einem Server innerhalb eines Statistischen Amtes oder bei einem anderen Datenproduzenten[1] gespeichert ist. Alle Analysen sind erlaubt und die Ergebnisse werden in Echtzeit zurückgegeben. Zusätzlich zu dem Ergebnis mit anonymen Daten wird ein Gütemaß ausgegeben, das es dem Wissenschaftler erlaubt zu beurteilen, ob die

[1] Der Begriff “Statistisches Amt” wird von nun an stellvertretend für alle Datenproduzenten verwendet.

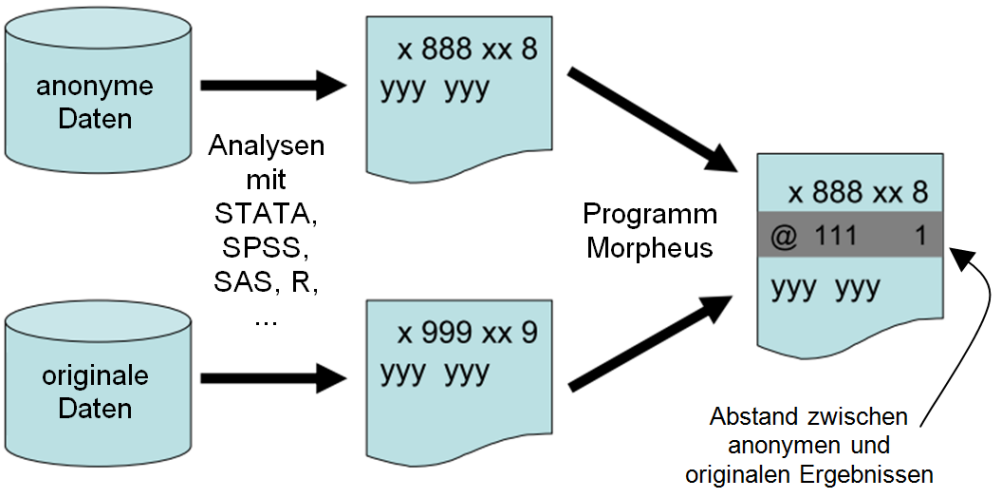
Interpretation, die von den anonymen Daten abgeleitet wird, die gleiche ist wie mit Originaldaten oder nicht. So lange das Gütemaß eine gute Qualität des anonymen Datensatzes anzeigt, können Wissenschaftler mit dem Morpheus-System arbeiten.

2 Morpheus – Vorteile und Herausforderungen

Das Morpheus-System gewährt als neuer Ansatz Zugang zu Mikrodaten und die Ergebnisse werden in Echtzeit angezeigt. Dabei wird der ganze Prozess des Datenzugangs gestaltet. Dieser Beitrag zeigt, dass das System große Vorteile sowohl für Datennutzer als auch für die Datenproduzenten mit sich bringen würde. Es soll der Nutzen für beide Gruppen hervorgehoben werden und die neuen Herausforderungen durch das neue System benannt werden. Zunächst folgt dazu ein Überblick über die Funktionsweise des Systems.

Morpheus besteht, wie in Abbildung 1 dargestellt, aus drei Komponenten. In einem ersten Schritt arbeitet ein Nutzer mit anonymen Mikrodaten. Diese Daten sind auf einem Server gespeichert und der Wissenschaftler kann sie mit einem der üblichen Statistik-Softwarepakete (SPSS, SAS, Stata oder R) auswerten. Um die Beschreibung nachfolgend zu vereinfachen, wird der Fokus im Folgenden auf das Statistik-Programm Stata gelegt. Auch der Prototyp wurde für Stata programmiert. Idealerweise kann der Wissenschaftler normal mit dem Programm und all seinen Funktionalitäten arbeiten. Insbesondere besteht die Möglichkeit die Daten anzusehen. Da die Mikrodatensätze anonymisiert sind, bestehen bei der Anzeige der Daten keine Bedenken bezüglich des Datenschutzes. Darüber hinaus müssen für Wissenschaftler keine Regeln über die am Gastwissenschaftlerarbeitsplatz üblicherweise geltenden Regeln hinaus (ein guter Leitfaden ist in Office for National Statistics 2008 veröffentlicht) aufgestellt werden. Programmsyntaxen sollten stets gut dokumentiert sein und bei jeder Analyse müssen die Fallzahlen ausgewiesen werden. Die Arbeit mit den anonymisierten Daten sollte für die Wissenschaft unkompliziert und gewohnt sein. Alle Analysen und Befehle sind erlaubt, da die Daten anonym sind und somit auch Off-Site analysiert werden können. Keine Analyse stellt ein Enthüllungsrisiko dar.

Die zweite Komponente in Morpheus sind die korrespondierenden Originalmikrodaten, die ebenfalls auf dem Server gespeichert sind. Der Nutzer sieht die Originalergebnisse jedoch nicht. Sie werden nur als Input für die dritte Komponente verwendet. Der Nutzer bekommt die Ergebnisse mit den anonymen Daten angezeigt und erhält dazu jeweils ein Gütemaß, das als Qualitätsmaß für die anonymen Ergebnisse fungiert. Dieser Indikator wird als Differenz zwischen dem Ergebnis mit anonymen Daten und Originaldaten berechnet. Um eine direkte Rückrechenbarkeit zu verhindern wird nur der absolute Abstand angezeigt. In Abschnitt 3 werden mögliche Enthüllungsrisiken, die durch die Ergänzung der anonymen und sicheren Ergebnisse um den Qualitätsindikator entstehen könnten, diskutiert.



.
Abb. 1 Überblick über die Funktionsweise von Morpheus, eigene Darstellung

Der Qualitätsindikator wird für jedes statistische Ergebnis, das der Nutzer mit der Statistiksoftware anfordert, berechnet. Er wird in eine neu eingefügte Zeile jeweils direkt unter das numerische Ergebnis geschrieben. Die neue Zeile wird mit einem @-Zeichen an der ersten Stelle gekennzeichnet. Dieses wurde zur Kennzeichnung ausgewählt, da es üblicherweise in Programmiersprachen nicht verwendet wird. Zusätzlich wird die neue Zeile grau hinterlegt.

Morpheus ist insbesondere für die Erstellung von Zwischenergebnissen vorgesehen. Für Veröffentlichungen sollen die Wissenschaftler weiterhin Ergebnisse auf Basis der Originaldaten erhalten, die vor der Freigabe manuell überprüft werden. Dieses Vorgehen soll die Akzeptanz von Morpheus erhöhen und gleichzeitig die Arbeitsbelastung in den FDZ erheblich reduzieren, da in einem typischen Forschungsprojekt der Großteil aller Ergebnisse Zwischenergebnisse sind, die nicht in die Veröffentlichung eingehen.

2.1 Vorteile für die Wissenschaftler

Das Morpheus-Verfahren ist nutzerfreundlich, da die Nutzer die Ergebnisse in Echtzeit erhalten. Eine der bemerkenswertesten Eigenschaften ist, dass bei Morpheus alle statistischen Analysen und Programmierbefehle erlaubt sind. Dies ist bei Remote–Access-Systemen nicht selbstverständlich, da die meisten Systeme entweder nur eine Liste an Befehlen zulassen oder eine Liste an Befehlen verbieten (O'Keefe et al. 2009, Lucero et al. 2009, National Center for Health Statistics 2010). Fast alle Wissenschaftler kombinieren in ihrer Forschung deskriptive und inferenzstatistische Befehle. Oft berechnen und benötigen sie dabei keine Konfidenzintervalle für deskriptive Analysen, weder für Stichprobenerhebungen noch für Vollerhebungen. Daher unterscheidet sich das hier vorgeschlagene System deutlich von den von Oganian et al. (2009) vorgeschlagenen "verification servers" (auf deutsch etwa Überprüfungsserver). Bei Morpheus kann wie gewohnt mit den Statistik-Softwarepaketen gearbeitet werden und es sind keine Programmänderungen nötig.

Darüber hinaus bietet Morpheus viele Vorteile gegenüber den bisher von den meisten FDZ angebotenen Zugangswegen. Bei der kontrollierten Datenfernverarbeitung (KDFV) können Wissenschaftler die Daten nicht in Augenschein nehmen und müssen ggf. mehrere Tage warten, bis ihr Programm von einem FDZ-Mitarbeiter gestartet und manuell auf Enthüllungsrisiken untersucht wurde. Diese Wartezeit entfällt bei Morpheus. Wenn Wissenschaftler einen Gastwissenschaftlerarbeitsplatz (GWAP) nutzen, entstehen Reise- und Aufenthaltskosten, und dennoch sind die Daten, wenn auch leicht, doch faktisch anonymisiert. Bei Scientific-Use-Files (SUF) müssen die Nutzer den Datenproduzenten vertrauen, dass sie die Mikrodaten „gut“ anonymisiert haben und die Anonymisierungsmethoden korrekt angewendet wurden (Alexander et al. 2010). Bei Morpheus entfallen die Reisekosten. Datennutzer können einen Blick auf Mikrodaten werfen und so die Datenstruktur sehen. Und sie erhalten die Ergebnisse sowohl in Echtzeit als auch zusätzlich ein Qualitätsmaß, dass es ihnen erlaubt zu beurteilen, ob sie den anonymen Ergebnissen vertrauen wollen. Morpheus kann somit auch als Instrument interpretiert werden, um das Vertrauen in Anonymisierungsverfahren zu stärken (Oganian et al. 2009).

Solange die Indikatoren eine ausreichend gute Qualität aufweisen, können Nutzer mit den Morpheus-Ergebnissen arbeiten. Die Entscheidung liegt beim Nutzer selbst: Er kann entscheiden, ob die Indikatoren für ihn ausreichend gut sind oder ob er lieber auf manuell auf Enthüllungsrisiken geprüfte Ergebnisse mit Originaldaten warten möchte. Berechnet ein Wissenschaftler beispielsweise Veränderungsraten, z. B. in der Variable Einkommen, könnte er folgende Ergebnisse erhalten: In Szenario 1 sei das Einkommen um 10 % gestiegen, wobei das Qualitätsmaß eine Abweichung von ±2 Prozentpunkten anzeigt. Dann kann der Wissenschaftler mit Sicherheit formulieren, dass das Einkommen gestiegen ist. Wenn er in einem Szenario 2 als Ergebnis erhält, dass das Einkommen um 10 % gestiegen ist und das Qualitätsmaß eine Abweichung von ±20 Prozentpunkten anzeigt, dann kann der Wissenschaftler aus der Analyse keine definitiven Schlussfolgerungen ziehen und würde für seine Studie Morpheus-Ergebnisse wahrscheinlich nicht akzeptieren.

Ein Nachteil für die Nutzer ist die ungewohnte Ergebnispräsentation. In den durch Stata produzierten LogFiles wird die neu eingefügte graue Zeile mit dem @-Zeichen zu Beginn ungewöhnlich sein. Unter Berücksichtigung der großen Vorteile gegenüber den heutigen Zugangswegen werden die Nutzer sich an diese Darstellung jedoch rasch gewöhnen.

Morpheus könnte grundsätzlich von allen Datenproduzenten eingesetzt werden. In Deutschland, wo die Datenschutzgesetze besonders streng sind, würde das Verfahren beim Mikrodatenzugang einen deutlichen Schritt nach vorn bedeuten. Hierzulande wurde argumentiert, dass die Anzeige von Mikrodaten am Bildschirm bei einem Fernzugriff bereits eine Datenübermittlung darstellt (laut juristischen Gutachten, die von der Arbeitsgruppe „Future Data Access“ des Rates für Sozial- und Wirtschaftsdaten 2010 in

Auftrag gegeben wurden). Daher dürfen nur Daten und statistische Ergebnisse, die mindestens faktisch anonym sind, am Bildschirm angezeigt werden. Der Gesamtprozess muss garantieren, dass dies stets sichergestellt ist.

2.2 Vorteile und Herausforderungen für die Datenproduzenten

Der große Vorteil für Datenproduzenten ist, dass Wissenschaftler für den überwiegenden Teil ihrer Analysen selbstständig mit Morpheus arbeiten können. Bei diesen Analysen entfällt die manuelle Prüfung. Derzeit ist geplant, dass für die Endergebnisse, die in Veröffentlichungen verwendet werden, die Analyseergebnisse auf Basis der Originaldaten manuell geprüft und den Wissenschaftlern zugesandt werden. Die Unterscheidung in Zwischenergebnisse aus dem Morpheus-System und finale Ergebnisse, die manuell geprüft werden, hätte den Vorteil, dass das Morpheus-System als faktisch anonym akkreditiert werden könnte. Veröffentlicht werden dürfen jedoch nur absolut anonyme Ergebnisse. Diese Unterscheidung ist eine Besonderheit des deutschen Rechtssystems. In anderen Ländern und für andere Datenbestände gilt eine andere Rechtslage und eine Unterscheidung wäre eventuell nicht zwingend notwendig.

Die größte Herausforderung besteht darin, dass für jede Statistik ein anonymer Datenbestand erstellt werden muss. Es wäre eine Investition jeden Mikrodatenbestand, zu dem die Statistikämter oder andere Datenproduzenten den Zugang gewähren, ein erstes Mal zu anonymisieren. Wenn neue Wellen einer Erhebung in Morpheus eingestellt werden, können die Anonymisierungsverfahren und die verwendeten Parameter meist erneut angewandt werden. Die meisten Datenproduzenten sind bereits aktiv dabei, die besten Anonymisierungsverfahren für ihre Daten zu identifizieren, da viele bereits faktisch anonyme Scientific-Use-Files oder absolut anonyme Datenstrukturfiles für das Erarbeiten von Programmcode bei der KDFV anbieten. Morpheus ist dabei unabhängig von spezifischen Anonymisierungsverfahren. Die einzige Anforderung an das Anonymisierungsverfahren ist, die Dimensionen des Datensatzes zu erhalten, d. h. die Anzahl der Beobachtungen und die Ausprägungen in den kategorialen Variablen müssen bestehen bleiben. Die Vergröberung scheidet daher als Anonymisierungsverfahren aus, aber z.B. Swapping, Stochastische Überlagerung oder Imputation (für eine Übersicht siehe Hundepool et al. 2010) können angewendet werden.

3 Stochastische Veränderung des Abstandes

Im Morpheus-System sind alle Analysen erlaubt, da diese in erster Instanz an einem anonymen Datensatz ausgeführt werden. Allerdings könnte die zusätzliche Angabe des Gütemaßes zu neuen Enthüllungsrisiken führen. Daher wurde untersucht, ob die Zusatzinformation ein neues Enthüllungsrisiko darstellt.

Da Originalergebnisse geheim zu haltende Informationen enthüllen könnten, muss verhindert werden, dass Nutzer den exakten Wert des Originalergebnisses zurückrechnen können. Um die Möglichkeit auszuschließen, dass anhand des anonymen Ergebnisses und dem Abstand das Originalergebnis ausgerechnet wird, erfolgt keine Veröffentlichung darüber, ob das Originalergebnis höher oder niedriger als das mit den anonymen Daten errechnete Ergebnis ist. Wenn man den Abstand einmal vom Ergebnis mit anonymen Daten subtrahiert und einmal darauf addiert, erhält man zwei potenzielle Originalergebnisse. Um die Mehrdeutigkeit und Unsicherheit zu erhöhen, wird eine stochastische Veränderung des absoluten Abstandes vorgeschlagen.

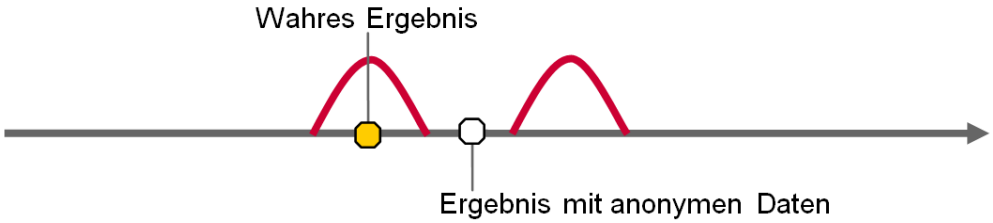


Abb. 2a Variante 1 den Abstand stochastisch zu verändern: unverzerrter Punktschätzer, eigene Darstellung.

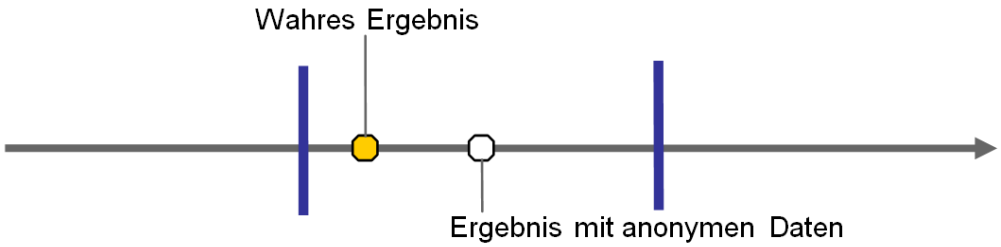


Abb. 2b Variante 2 den Abstand stochastisch zu verändern: maximale Entfernung, eigene Darstellung.

Zuerst war geplant, den absoluten Abstand mit einem zufälligen Faktor zu multiplizieren. Den Zufallsfaktor u_1 würde man aus einer Normalverteilung mit Erwartungswert 1 und Varianz σ wie in Gleichung 1 dargestellt ziehen:

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

So kann ein punktgenauer Wert für die Entfernung angeben werden. Die Entfernung wäre unverzerrt, da in der Hälfte aller Fälle die Abweichung vergrößert würde und in der anderen Hälfte verkleinert. Im Erwartungswert würde die Entfernung, die zwischen dem originalen und dem anonymen Ergebnis besteht, erhalten bleiben. Diese Variante der stochastischen Veränderung ist in Abbildung 2a dargestellt. Allerdings hat diese Variante einen entscheidenden Nachteil: Jedes Mal, wenn die stochastische

Veränderung den Abstand reduziert, erscheint das anonyme Ergebnis von besserer Qualität, als es tatsächlich ist. Da die Wissenschaftler dann nicht mehr korrekt selbst entscheiden können, ob das Arbeiten mit dem anonymen Datensatz äquivalent zum Arbeiten mit den Originaldaten ist, war dies für die Nutzer nicht akzeptabel.

Daher wird eine andere zufällige Veränderung des absoluten Abstandes vorgeschlagen. Der Abstand wird mit einem Zufallsfaktor u_2 multipliziert, der aus einer Gleichverteilung auf dem Intervall [1; x] wobei x>1 gezogen wird:

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

Der genaue Wert für x sollte nicht veröffentlicht werden, da er von Datenangreifern verwendet werden könnte. Der Abstand zwischen dem originalen und dem anonymen Ergebnis wird dann nicht verkleinert, nur vergrößert. Das neue Qualitätsmaß kann als Maximalabweichung des Originalergebnisses vom Ergebnis mit anonymen Daten interpretiert werden. Diese Variante ist in Abbildung 2b illustriert. Dem Nutzer wird das Ergebnis mit anonymen Daten und die maximale Entfernung mitgeteilt. Die beiden blauen Linien spannen ein Intervall um das anonyme Ergebnis mit der Intervallbreite zweimal Maximalabstand, da der Nutzer nicht weiß, ob das Originalergebnis nach oben oder unten vom anonymen Ergebnis abweicht. Das Originalergebnis kann jeder Punkt innerhalb im Intervall oder eine der beiden Intervallgrenzen sein.

Ein wichtiger Aspekt beim Generieren der Zufallsfaktoren ist, dass bei jeder Wiederholung der Analyse zu einem anderen Zeitpunkt oder unter Nutzung einer anderen Befehlssyntax wieder der gleiche Faktor gezogen werden muss. Jedes Mal, wenn die gleiche Analyse über die gleiche Gruppe an Merkmalsträgern berechnet wird, muss die Zufallsveränderung gewährleisten, dass der gleiche Zufallsfaktor verwendet wird, sodass es durch wiederholte Berechnungen nicht möglich ist, die dahinterliegende Verteilung der Zufallsfaktoren zu enthüllen.

Die Zufallsveränderung soll auch verhindern, dass bei logischen Einschränkungen die Richtung der Abweichung eindeutig wird. Solch ein Fall könnte u. a. bei Variablen eintreten, die keine negativen Werte annehmen können, wie beispielsweise die Anzahl der Arbeitnehmer. Sobald der Abstand betragsmäßig größer ist als der Wert des anonymen Ergebnisses, kann vermutet werden, dass das Originalergebnis größer als das anonyme Ergebnis sein muss. Allerdings könnte dieser Effekt nur ein Ergebnis der stochastischen Veränderung des Abstandes sein, und das wahre Ergebnis ist in Wirklichkeit trotzdem kleiner als das anonyme. Dieser Fall kann eintreten, wenn durch die Multiplikation der Abstand deutlich vergrößert wurde. Dadurch schützt die stochastische Veränderung der absoluten Abweichung zwischen anonymem und originalem Ergebnis die Anonymität der Ergebnisse und der Mikrodaten, die bei der Analyse verwendet wurden.

4 Technische Umsetzung

Ein Prototyp von Morpheus, der Stata-Programme verarbeiten kann, wurde bereits entwickelt. Diese erste Version von Morpheus erzeugt eine Outputdatei mit den Ergebnissen, die aufgrund der anonymen Mikrodatendatei erzeugt wurden, und den jeweils korrespondierenden Qualitätsmaßen in einer separaten Zeile direkt darunter. Die neu eingefügte Zeile ist markiert durch ein @-Zeichen an erster Position und erhält einen grauen Hintergrund. Die Morpheus-Ergebnisdatei kann wie normaler Stata-Output und das Qualitätsmaß kann wie folgt interpretiert werden: Das Ergebnis auf Basis der Originaldaten weicht maximal um den Betrag des Qualitätsmaßes vom Ergebnis auf Basis der anonymen Daten ab.

```
. 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
----------------------------------------------------------------------------
```

Abb. 3a Beispiel einer Morpheus-Ergebnisdatei – deskriptive Analysen, eigene Darstellung.

Die Abbildungen 3a und 3b enthalten Ausschnitte eines Beispieloutputs. Die Abbildung 3a zeigt einige deskriptive Analysen der Variable Exportanteil im Jahr 2000 („expshare2000“). Abbildung 3b weist Ergebnisse einer Regression mit fixen Effekten aus. Hier wird der Logarithmus der Arbeitsproduktivität durch einen Exportdummy, die Firmengröße gemessen an der Anzahl der Mitarbeiter und der quadrierten Mitarbeiteranzahl und einem Dummy für hohe Humankapitalintensität erklärt. Die Regression reproduziert Analysen im Stil von Dr. Fryges und Prof. Wagner, die die Bestimmungsgründe für Exportaktivität in der Industrie in Deutschland untersuchten (Fryges and Wagner 2008).

```
.   xtreg 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
-------------+----------------------------------------------------------------
```

Abb. 3b Beispiel einer Morpheus-Ergebnisdatei – Regression, eigene Darstellung.

Das Qualitätsmaß ist für alle statistischen Ergebnisse relativ gut: Die anonymen deskriptiven Ergebnisse weichen nur minimal von den korrespondierenden Originalergebnissen ab. Der wahre Mittelwert und die wahren Perzentile weichen um weniger als 1 % von den anonymen Ergebnissen ab. In der Regression werden die gleichen Signifikanzniveaus und die gleiche Größenordnung bei den Koeffizienten für fast alle Regressoren ausgewiesen. Alle exogenen Variablen sind bei einem 1 %-Niveau signifikant, dies gilt sowohl in dem anonymen als auch dem originalen Datensatz. Die einzige Ausnahme ist die Variable quadrierte Mitarbeiteranzahl, die nur auf einem Niveau von 5 % signifikant ist. Die Größenordnung der Koeffizienten ist gut erhalten, Vorzeichenänderungen treten nicht auf. Ein Wissenschaftler, der diese Ergebnisse erhält, kann seine Analyse und die Spezifikationen seiner Regressionen mit Morpheus entwickeln und sogar einen ersten Entwurf seines Artikels schreiben, da die Zwischenergebnisse aus Morpheus eindeutig zu interpretieren sind.

5 Übertragbarkeit auf andere Statistikprogramme

Die am häufigsten verwendeten Statistik-Softwarepakete in FDZ der meisten Länder sind SPSS, SAS, Stata und R. Während dieser Beitrag sich schwerpunktmäßig auf das Programm Stata konzentrierte, kann Output der Statistikpakete SPSS, SAS und R generell ebenfalls mit dem Prototyp bearbeitet werden; SPSS-Outputdateien müssen nur zuvor in eine ASCII-Datei umgewandelt werden.

6 Zusammenfassung

Das Morpheus-Verfahren bietet einen innovativen Zugangsweg zu Mikrodaten. Die enormen Vorteile für die Wissenschaftler sind die Anzeige von statistischen Ergebnissen in Echtzeit, die Möglichkeit einen Blick auf die Mikrodaten zu werfen und die Tatsache, dass alle Befehle und Analysemethoden erlaubt sind. Um diese Vorteile zu erhalten, werden die Befehle zwar an anonymen Mikrodaten ausgeführt, aber der Wissenschaftler erhält zusätzlich ein Qualitätsmaß zu jedem statistischen Ergebnis.

Datenproduzenten müssen jedoch Arbeitszeit investieren, um eine anonyme Version für jeden Datensatz zu erstellen. Allerdings besitzen die meisten Datenproduzenten bereits gute Kenntnisse in Anonymisierungsverfahren, da sie oft einige Datensätze bereits als Scientific-Use-Files anbieten oder Datenstrukturfiles erstellen, um das Programmieren von Syntax bei der KDFV zu erleichtern.

In einem nächsten Schritt muss ein Zugangssystem entwickelt werden. Es bedarf einer Serverinfrastruktur mit einer Oberfläche, die registrierte Nutzer identifizieren kann und die Aufträge automatisch abarbeitet. Diese technische Lösung kann auf bereits existierenden Systemen wie LISSY (Coder and Cigrang 2003) aufbauen. Morpheus könnte auf Servern installiert werden, auf

die via Remote Desktop übers Internet zugegriffen werden kann. Zusätzliche Sicherungsmaßnahmen durch Zertifikate, Passwörter oder Smart Cards wären denkbar. Wenn leistungsstarke Maschinen verwendet werden, die zwei Programme parallel abarbeiten können, sollte es auch möglich sein, dass sich die Rechenzeit nicht verdoppelt.

Auch wenn bereits ein erster Prototyp von Morpheus entwickelt und programmiert wurde, so konnten noch nicht alle Details erschöpfend untersucht werden. Morpheus befindet sich noch in der Entwicklungsphase. Ein Aspekt, der noch weiter erforscht werden muss, ist beispielsweise die Frage, wann ein Abstand von Null ein Risiko darstellt und wie er stochastisch verändert werden könnte.

Literatur

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

Brandt, M. und 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. und 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. und 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 und P.-P. de Wolf (2010): Handbook on Statistical Disclosure Control. Version 1.2, verfügbar unter http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf

Lucero, J., Singh, L. und 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. Webseiten zuletzt besucht am 27. August 2012, http://www.cdc.gov/rdc/Data/B4/DisclosureManual.pdf und http://www.cdc.gov/rdc/Data/B2/SASSUDAANRestrictions.pdf

O'Keefe, C.M. und 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). Webseite zuletzt besucht am 27. August 2012, http://www.ons.gov.uk/ons/about-ons/who-we-are/services/vml/accessing-the-vml/how-to-access-the-vml/index.html

Oganian, A., Reiter, J. P. und 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.