Rat für Sozial- und
WirtschaftsDaten (RatSWD)

www.ratswd.de

RatSWD
Working Paper Series

Working Paper Nr. 174

Enthüllungsrisiko beim Remote Access: Die Schwerpunkteigenschaft der Regressionsgerade

Alexander Vogel

März 2011

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 erschienen 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: Rat für Sozial- und Wirtschaftsdaten | Mohrenstraße 58 | 10117 Berlin | office@ratswd.de

Enthüllungsrisiko beim Remote Access: Die Schwerpunkteigenschaft der Regressionsgerade

Alexander Vogel*

März 2011

Zusammenfassung In der Literatur wird zunehmend untersucht, inwieweit Enthüllungsrisiken durch multivariate Analysemethoden beim indirekten Mikrodatenzugang über die kontrollierte Datenfernverarbeitung (Remote Access) bestehen. Daran anschließend zeigt der Beitrag, wie die Schwerpunkteigenschaft der Regressionsgerade von einem Datenangreifer für Enthüllungen genutzt werden könnte. Dabei wird sowohl auf Enthüllungsmöglichkeiten durch gezielt erstellte (strategische) Variablen als auch auf die Offenlegung geheim zu haltender Merkmalssummen natürlich vorkommender (also nicht gezielt erstellter) Variablen eingegangen. Hinsichtlich der Geheimhaltungsprüfung wird deutlich, dass es nicht ausreichend ist, ein einfaches Kriterium wie die in das Regressionsmodell eingehende Anzahl der Beobachtungen zu Grunde zu legen. Um Enthüllungen auf Grundlage der Schwerpunkteigenschaft zu verhindern ist vielmehr eine Überprüfung jeder einzelnen im Regressionsmodell verwendeten Variable notwendig.

Schlüsselwörter Enthüllungsrisiko, Schwerpunkteigenschaft der Regressionsgerade, Kontrollierte Datenfernverarbeitung, Forschungsdatenzentrum

JEL Klassifikationen C10, C13

---

* Alexander Vogel
Statistisches Amt für Hamburg und Schleswig-Holstein, Forschungsdatenzentrum,
Fröbelstraße 15 – 17, 24113 Kiel
e-Mail: Alexander.Vogel@statistik-nord.de

Dieser Beitrag spiegelt ausschließlich die persönliche Meinung des Autors und nicht notwendigerweise die des Statistikamtes Nord wider.

Der Autor dankt Julia Höninger, Gerd Ronning, Hendrik Tietje sowie den Teilnehmern des Forums „Future Data Access“ im Rahmen der 5. Konferenz für Sozial- und Wirtschaftsdaten für Ihre Anmerkungen und Kommentare.

1 Einleitung

Seit geraumer Zeit bieten Forschungsdatenzentren für die empirisch arbeitende Wissenschaft Zugangsmöglichkeiten zu Mikrodaten der amtlichen Statistik und weiteren öffentlichen Datenproduzenten. In der Diskussion um den Datenzugang der Zukunft rückt dabei die Mikrodatennutzung mittels kontrollierter Datenfernverarbeitung (Remote Access) immer stärker in den Fokus (vgl. z.B. Brandt/Zwick 2009). Beim Remote Access erhalten die Nutzer keinen direkten Zugang zu den Mikrodaten, vielmehr erhalten sie die Möglichkeit Syntax zu entwickeln, mit der die Mikrodaten durch die Mitarbeiterinnen und Mitarbeiter der Forschungsdatenzentren ausgewertet werden. Anschließend werden den Nutzern nach einer Geheimhaltungsprüfung die Ergebnisse ihrer Auswertung zur Verfügung gestellt. Während die Regeln zur Geheimhaltungsprüfung von Tabellen und aggregierten Kennzahlen hinreichend dokumentiert sind, befindet sich die systematische Untersuchung von Enthüllungsrisiken durch Schätzoutput (z.B. auf der Basis von Regressionsanalysen) noch in den Anfängen.

Untersuchungen, ob und in welchem Ausmaß Enthüllungsrisiken durch multivariate Analysemethoden im Rahmen des indirekten Mikrodatenzugangs über die kontrollierte Datenfernverarbeitung bestehen, finden sich zum Beispiel bei Reznek (2003), Reznek/Riggs (2005) und Ronning et al. (2010). Hinsichtlich der Regressionsanalyse wird unter anderem auf die Enthüllungsmöglichkeiten durch gezielt gesetzte Dummy-Variablen sowie durch die Erzeugung künstlicher Ausreißer eingegangen. Ronning et al. (2010) schließen ihren Artikel jedoch mit dem Hinweis, dass die „präsentierten Beispiele sicher nur ein kleiner Ausschnitt dessen sind, was möglich ist“ (Ronning et al. 2010, S. 46). Motiviert durch diese Aussage, wird im Folgenden eine weitere Möglichkeit gezeigt, wie ein Datenangreifer die Ergebnisse von Regressionsanalysen dazu nutzen könnte, geheim zu haltende Merkmalssummen aufzudecken sowie mit Hilfe gezielt erstellter metrischer Variablen Einzelwerte zu enthüllen.

Genutzt wird hierfür die Schwerpunkteigenschaft der Regressionsgerade, welche besagt, dass die Regressionsgerade durch den Schnittpunkt der Mittelwerte[1] verläuft.

[1] In diesem Beitrag wird unter dem Begriff Mittelwert immer das arithmetische Mittel verstanden.

Formal lässt sich dies für eine abhängige Variable *y* und *k* unabhängige Variablen x_1 *bis* x_k wie folgt darstellen:

(1) $\bar{y} = b_0 + b_1\bar{x}_1 + b_2\bar{x}_2 ... + b_k\bar{x}_k$

wobei b_0 bis b_k für die geschätzten Regressionskoeffizienten stehen.[2]

Im Weiteren gliedert sich der Beitrag wie folgt: Abschnitt 2 stellt die Grundidee vor, wie die Schwerpunkteigenschaft der Regressionsgerade für Enthüllungen genutzt werden kann. Abschnitt 3 behandelt die Enthüllung von geheim zu haltenden Merkmalssummen natürlich vorkommender (also nicht gezielt erstellter) Variablen, während sich Abschnitt 4 mit den Enthüllungsmöglichkeiten durch gezielt erstellte (strategische) Variablen beschäftigt. Abschließend zieht Abschnitt 5 ein Fazit und benennt die Konsequenzen für die heutige manuelle bzw. für eine zukünftig automatisierte Outputprüfung in den Forschungsdatenzentren.

2 Grundidee

Bevor statistische Ergebnisse veröffentlicht werden dürfen, müssen sie auf Geheimhaltung geprüft werden. Die Geheimhaltungsprüfung der Statistischen Ämter des Bundes und der Länder ist dabei im Wesentlichen die Geheimhaltungsprüfung von Merkmalssummen. Zur primären Sperrung von Merkmalssummen kommt es aufgrund von zu geringen Fallzahlen (eine Mindestanzahl von Einheiten (zumeist drei) muss zu einem Wert beitragen) und wenn Dominanz- bzw. Konzentrationsfälle vorliegen (ein Überblick über die verschiedenen Regeln der primären Geheimhaltung findet sich zum Beispiel bei Giessing/Dittrich 2006, S. 806). Zusätzlich kommt es durch Sekundärsperrungen zur Geheimhaltung von weiteren Merkmalssummen, um zu verhindern, dass primär geheim gehaltene Werte durch Summen- oder Differenzenbildung zurückgerechnet werden können.[3]

---

[2] Reznek (2003) stellt in seinem Artikel zu Enthüllungsrisiken des Regressionsmodels ebenfalls kurz den in Gleichung 1 gezeigten Zusammenhang dar, geht allerdings nicht ausführlich darauf ein.

[3] Werden in einer Tabelle z.B. der Wert des Gesamtumsatzes, des Auslandsumsatzes sowie des Inlandsumsatzes eines Wirtschaftszweiges angegeben und muss z.B. der Wert des Auslandsumsatzes aufgrund einer Geheimhaltungsregel gesperrt werden (primäre Geheimhaltung), so muss auch der Wert des Inlandsumsatzes gesperrt werden (sekundäre Geheimhaltung) um eine Berechnung des Auslandsumsatzes durch Differenzenbildung zu verhindern.

Aus der Prüfung von Merkmalssummen leitet sich direkt die Prüfung von Mittelwerten ab, da sich aus dem Mittelwert und der Fallzahl die Merkmalssumme berechnen lässt.

In Gleichung 2 wird Gleichung 1 um einen Regressor *z* erweitert, dessen Merkmalssumme für die in der Regression betrachtete Gruppe von Merkmalsträgern geheim gehalten werden muss.

(2) $\bar{y} = b_0 + b_1\bar{x}_1 + b_2\bar{x}_2 ... + b_k\bar{x}_k + b_z\bar{z}$

Sind die geschätzten Regressionskoeffizienten sowie die Mittelwerte der Variablen *y* und x_1 bis x_k bekannt, ergibt sich der Mittelwert von *z* durch einfaches Umstellen der Gleichung 2 zu:

(3) $\bar{z} = \frac{b_0 + b_1\bar{x}_1 + b_2\bar{x}_2 ... + b_k\bar{x}_k - \bar{y}}{-b_z}$

Die Merkmalssumme der Variable *z* kann nun durch Multiplikation des in Gleichung 3 berechneten Mittelwertes mit der Anzahl der Beobachtungen der Regression (*n*) berechnet werden (siehe Gleichung 4).

(4) $\sum_{i=1}^{n} z_i = \bar{z} \cdot n$

Um die Aufdeckung zu verhindern, ist es daher notwendig, den geschätzten Koeffizienten b_z geheim zu halten.

In den folgenden beiden Abschnitten wird gezeigt, dass ein einfaches Kriterium wie die Anzahl der Fälle, die in die Regression einfließen, nicht genügt, um zu erkennen, ob ein Koeffizient geheim gehalten werden muss. Abschnitt 3 beschäftigt sich mit Fällen von natürlich vorkommenden – nicht gezielt erstellten – Variablen in denen trotz großer Fallzahl eine geheim zu haltende Merkmalssumme errechnet werden kann. Abschnitt 4 behandelt Enthüllungsmöglichkeiten durch strategische – gezielt erstellte – Variablen.

3 Enthüllung von Merkmalssummen nicht gezielt erstellter Variablen

3.1 Konstellationen von Geheimhaltungsfällen trotz hoher Fallzahl

Grundsätzlich sind drei Konstellationen denkbar, in denen trotz einer großen Anzahl von Beobachtungen, die in die Regression einfließen, die geheim zu haltende Merkmalssumme einer Variable *z* aufgrund der Schwerpunkteigenschaft errechnet werden kann:

a) Die Variable *z* nimmt für sehr viele Beobachtungen den Wert Null an, nur für ein oder zwei Merkmalsträger liegt ein Wert ungleich Null vor. Die Merkmalssumme muss daher aufgrund der gängigen Mindestfallzahlregel (mindestens drei Einheiten müssen zur Merkmalssumme beitragen) gesperrt werden, obwohl die Regression auf über zwei Beobachtungen basiert. Als Beispiele für Variablen welche für viele Beobachtungen den Wert Null aufweisen, und für die es auch auf höheren Gliederungsebenen zu Geheimhaltungsfällen kommt, sind unter anderem der Auslandsumsatz oder auch Angaben zu Subventionen zu nennen.[4]

b) Die Variable *z* weist für viele Merkmalsträger nur einen sehr kleinen Wert und für einen oder zwei Merkmalsträger einen sehr großen Wert auf. Hier muss die Merkmalssumme aufgrund von Dominanz- bzw. Konzentrationsregeln gesperrt werden.

c) Die Merkmalssumme der Variable *z* fungiert in einer Veröffentlichung der amtlichen Statistik als Sekundärsperrpartner und ist daher geheim zu halten.

Während die Fälle a) und b) relativ schnell durch primäre Geheimhaltungsprüfungen aller im Regressionsmodell enthaltenden Variablen aufgedeckt werden können (z.B. durch das Auszählen der Anzahl der Merkmalsträger mit Werten ungleich Null, bzw. durch den Test auf Dominanz bzw. Konzentration), erfordert der dritte Fall c) ein recht aufwendiges Abgleichen jeder Regressionsvariable mit den Veröffentlichungen der amtlichen Statistik (sofern für die Untergruppe, für welche die Regression geschätzt wird, Veröffentlichungen der amtlichen Statistik vorliegen).

[4] Neben natürlich existierenden Variablen dieser Art, kann eine Variable (analog zu einem strategischen Dummy) auch durch einen Datenangreifer gezielt erzeugt werden (siehe Abschnitt 4).

Im Folgenden wird an einem Beispiel aus dem Agrarbereich eine Möglichkeit zur Offenlegung der Merkmalssumme einer „natürlichen“ Variablen demonstriert, auf die die unter Fall a) genannten Bedingungen zutreffen.

3.2 Beispiel: Zuckerrübenfläche in Hamburg 2003

In der über die Regionaldatenbank erhältlichen Tabelle „Landwirtschaftliche Betriebe mit Ackerland und deren Ackerfläche nach Fruchtarten“ wurde für das Land Hamburg im Berichtsjahr 2003 der Wert der Zuckerrübenfläche gesperrt, da nur ein landwirtschaftlicher Betrieb in Hamburg Zuckerrüben anbaut (siehe Abbildung 1, letzte Spalte, dritte Zeile von unten).

ABBILDUNG 1: LANDWIRTSCHAFTLICHE BETRIEBE MIT ACKERLAND UND DEREN ACKERFLÄCHE NACH FRUCHTARTEN

Landwirtschaftliche Betriebe mit Ackerland und deren Ackerfläche nach Fruchtarten - Erhebungsjahr - regionale Tiefe: Bundesländer

Allgemeine Agrarstrukturerhebung

Bundesländer Landwirtschaftliche Betriebe mit Ackerland Landwirtschaftlich genutzte Fläche			Einheit		Hackfrüchte	
				Arnen	Kartoffeln	Zuckerrüben
2003						
DG	Deutschland	Landwirtschaftliche Betriebe mit Ackerland	Anzahl	5 449	72 695	43 353
		Landwirtschaftlich genutzte Fläche	ha	9 433	287 264	445 630
01	Schleswig-Holstein	Landwirtschaftliche Betriebe mit Ackerland	Anzahl	8 969	610	1 008
		Landwirtschaftlich genutzte Fläche	ha	7 833	5 809	12 557
02	Hamburg	Landwirtschaftliche Betriebe mit Ackerland	Anzahl	114	40	1
		Landwirtschaftlich genutzte Fläche	ha	2 667	27	.
03	Niedersachsen	Landwirtschaftliche Betriebe mit Ackerland	Anzahl	7 524	8 612	8 562
		Landwirtschaftlich genutzte Fläche	ha	8 435	125 903	113 530

Quelle: Auszug aus der Tabelle „Landwirtschaftliche Betriebe mit Ackerland und deren Ackerfläche nach Fruchtarten“. Erhebungsjahr: 2003. regionale Tiefe: Bundesländer (siehe Statistische Ämter des Bundes und der Länder, 2011).

Dieser gesperrte Wert könnte zum Beispiel mit folgendem Regressionsmodell offen gelegt werden:

Auf der linken Seite der Regressionsgleichung steht als abhängige Variable der Standarddeckungsbeitrag der landwirtschaftlichen Betriebe in Euro (EF70). Auf der

rechten Seite werden die landwirtschaftlich genutzte Fläche (EF258), die Anzahl der Rinder (EF119), die Zuckerrübenfläche (EF220) sowie ein Dummy, der angibt ob es sich um ein Einzelunternehmen handelt oder nicht (EF13_einzel), aufgenommen.[5]

Die Ergebnisse der Regressionsschätzung sind in Tabelle 1 dargestellt. Die Schätzung wurde in Form einer normalen OLS-Regression mit Stata 10 durchgeführt. Als Datensatz dient die über die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder verfügbare Agrarstrukturerhebung 2003 für das Land Hamburg.[6]

TABELLE 1: REGRESSIONSERGEBNISSE DER SCHÄTZUNG DES STANDARDDECKUNGSBEITRAGES DER LANDWIRTSCHAFTLICHEN BETRIEBE IN HAMBURG 2003

	Schätzung des Standarddeckungsbeitrages der landwirtschaftlichen Betriebe in Euro (EF70)					
Unabhängige Variablen	Geschätzte Koeffizienten	Standard-fehler	t-Wert	p-Wert	95% Konfidenzintervall	
landwirtschaftlich genutzte Fläche (EF258)	443.6301	320.7652	1.38	0.167	-185.743	1073.003
Anzahl der Rinder (EF119)	-517.899	343.9492	-1.51	0.132	-1192.762	156.9637
Einzelunternehmen ja/nein (EF13_einzel)	51988.56	21057.29	-2.47	0.014	-93305.05	-10672.07
Zuckerrübenfläche (EF220)	x	x	-0.28	0.780	x	x
Konstante	143833.1	20594.82	6.98	0.000	103424	184242.2
Fallzahl	1.117					

Quelle: Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, Agrarstrukturerhebung für Hamburg, 2003, eigene Berechnungen.
Anmerkung: Gesperrte Werte sind durch „x“ gekennzeichnet.

[5] Das aufgestellte Regressionsmodell hat nicht den Anspruch den Standarddeckungsbeitrag sinnvoll zu erklären. Es dient lediglich als Beispiel.

[6] Nähere Informationen zu den Nutzungsmöglichkeiten von Mikrodaten der amtlichen Statistik finden sich bei Zühlke et al. (2004). Nähere Informationen zu den Agrarstatistiken finden sich bei Heinze/Vogel (2010).

Mit Blick auf die recht hohe Anzahl von 1.117 Beobachtungen wirken die Regressionsergebnisse unter Rückgriff auf das Kriterium einer hohen Fallzahl unverdächtig. Lässt sich der Nutzer in einem zweiten Schritt die ebenfalls unverdächtigen Mittelwerte der Variablen EF70, EF258, EF124, und EF13_einzel für alle landwirtschaftlichen Betriebe in Hamburg ausgeben (siehe Tabelle 2), ist die Merkmalssumme der Zuckerrübenfläche (EF220) jedoch mit den in Gleichung 3 und 4 gezeigten Zusammenhängen zu berechnen. Zwar kommt es zu einer leichten Abweichung durch die gerundeten Mittelwerte. In diesem Beispiel beträgt die Abweichung zwischen der berechneten und der tatsächlichen Zuckerrübenfläche in Hamburg jedoch nur 0.019 Prozent.

TABELLE 2: DESKRIPTIVER ÜBERBLICK ÜBER DIE IN DER REGRESSION VERWENDETEN VARIABLEN

Variablen	Fallzahl	Mittelwert	Standard-abweichung	Minimum	Maximum
Standarddeckungs-beitrag in Euro (EF70)	1117	98043.08	188397.8	x	x
landwirtschaftlich genutzte Fläche (EF258)	1117	12.29731	26.82767	0	x
Anzahl der Rinder (EF119)	1117	6.382274	25.02801	0	x
Einzelunternehmen ja/nein (EF13_einzel)	1117	0.9212175	0.2695196	0	1

Quelle: Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, Agrarstrukturerhebung für Hamburg, 2003, eigene Berechnungen.
Anmerkung: Gesperrte Werte sind durch „x“ gekennzeichnet.

Die zusätzliche Darstellung deskriptiver Statistiken über die in der Regression verwendeten Variablen ist ein gängiges Vorgehen. Im vorliegenden Fall genügt jedoch sogar ein Blick in den Statistischen Bericht zur Agrarstruktur in Hamburg 2003, um die benötigten Mittelwerte zu erhalten. Für den Standarddeckungsbeitrag findet sich hier bereits der gesuchte Mittelwert und hinsichtlich der landwirtschaftlich genutzten Fläche sowie der Rinderanzahl sind jeweils die Merkmalssummen

angeben. Zusätzlich ist die Anzahl der Einzelunternehmen ablesbar, welche die Berechnung des Mittelwertes der Dummy-Variable EF13_einzel ermöglicht (siehe Statistisches Amt für Hamburg und Schleswig-Holstein 2006, S. 12 – 13 sowie S. 16). Mit Hilfe dieser Werte ist ebenfalls eine Ermittlung der Hamburger Zuckerrübenfläche möglich. Zwar ist über diesen Weg die Abweichung zwischen berechneter und tatsächlicher Zuckerrübenfläche aufgrund von stärkeren Rundungen höher als bei der direkten Berechnung der Mittelwerte über die kontrollierte Datenfernverarbeitung, sie beträgt hier jedoch ebenfalls nur relativ niedrige 0.103 Prozent.

Um die Aufdeckung zu verhindern, ist es notwendig den geschätzten Koeffizienten der Variable EF220 geheim zu halten. Um die Möglichkeit der Rückrechnung des Koeffizienten über den Standardfehler bzw. das Konfidenzintervall zu unterbinden, müssen zusätzlich auch diese Werte gesperrt werden.

4 Enthüllung mit gezielt erstellten (strategischen) Variablen

4.1 Enthüllung einer geheim zu haltenden Merkmalssumme

Im vorangegangenen Abschnitt wurde ein Beispiel der Enthüllung einer geheim zu haltenden Summe auf Bundeslandebene dargestellt. Ein Blick in die Statistischen Berichte der amtlichen Statistik zeigt jedoch geheim zu haltende Merkmalssummen auf den unterschiedlichsten fachlichen und regionalen Gliederungsebenen. Theoretisch wären auch gesperrte Merkmalssummen auf tiefer regionaler Gliederung und/oder auf tief aufgeschlüsselter Wirtschaftszweigebene nach dem oben beschriebenen Vorgehen möglich. Ein Datenangreifer würde die Regressionsanalyse in diesem Fall über die gewünschte Untergruppe (z.B. eine bestimmte Wirtschaftszweig/Kreis Kombination) laufen lassen. Allerdings führt dies bei kleinen Untergruppen auch zu einer kleinen Anzahl an Beobachtungen welche in die Regression einfließen (was grundsätzlich bei der Outputprüfung auffälliger ist).

Um diese kleine Anzahl von Beobachtungen zu umgehen, bietet sich dem Datenangreifer die Möglichkeit die ihn interessierende Variable *z* so zu transformieren, dass sie nur dann Werte enthält, wenn die Einheit zur gewünschten Untergruppe gehört. Für einen bestimmten Kreis zeigt dies formal Gleichung 5.

$$(5)\quad z\_transformiert = \begin{cases} z & falls\ kreis = gewünschter\ kreis \\ 0 & sonst \end{cases}$$

Da es in Deutschland im Jahr 2003 über 400.000 landwirtschaftliche Betriebe gab, lässt sich nach diesem Vorgehen die (in Abschnitt 3.2 gezeigte) Aufdeckung der Hamburger Zuckerrübenfläche in einem Regressionsmodell verbergen, dem eine sehr hohe Anzahl an Beobachtungen zu Grunde liegt. Dazu könnte eine „strategische“ Variable erzeugt werden, die für alle landwirtschaftlichen Betriebe außerhalb Hamburgs den Wert Null enthält. Für die Hamburger Betriebe werden die Angaben zur Zuckerrübenfläche in die strategische Variable übertragen. Diese neue Variable wird in ein Regressionsmodell für alle deutschen landwirtschaftlichen Betriebe aufgenommen. Die Merkmalssumme der Hamburger Zuckerrübenflächen ließe sich nun nach der in Abschnitt 2 beschriebenen Prozedur ermitteln.[7]

Im Nicht-Agrarbereich wäre zum Beispiel die Enthüllung der Merkmalssumme des Auslandsumsatzes eines bestimmten Wirtschaftszweiges in einem bestimmten Kreis möglich, wenn die strategische Variable nur die Werte für diese bestimmte Wirtschaftszweig-Kreis-Kombination enthalten würde.

4.2 Enthüllung von Werten einer bestimmten Einheit

Ist ein Datenangreifer an dem ihm unbekannten Wert der Variable *z* einer bestimmten Einheit *m* interessiert, ist es grundsätzlich möglich, eine transformierte Variante der Variable *z* zu erstellen, welche nur für die Einheit *m* den Wert z_m enthält und für alle anderen Einheiten den Wert Null annimmt. In Anlehnung an die durch Ronning et al. (2010) gezeigte Erstellung eines gezielt gesetzten Dummies sind (unter anderem) die folgenden zwei Möglichkeiten gegeben, eine auf diese Weise transformierte Variable *z* zu erstellen:

a) Der Datenangreifer kennt ein Merkmal *w*, welches im Datensatz enthalten ist und von dem er weiß, dass die Ausprägung w_m nur einmal im Datensatz vorkommt (die Einheit *m* also eindeutig beschreibt). Erzeugt wird nun eine Variable *z_transformiert* die nur dann die Werte der Variable *z* enthält, wenn das Merkmal *w* die Ausprägung

[7] In diesem Zusammenhang sei ferner darauf hingewiesen, dass analog zur Geheimhaltungsprüfung von Merkmalssummen und Mittelwerten auch bei Regressionsmodellen ausgeschlossene Untergruppen ein Enthüllungsrisiko darstellen. Sind zum Beispiel die Mittelwerte für alle deutschen landwirtschaftlichen Betriebe bekannt, genügt die Schätzung des in Abschnitt 3.2 beschriebenen Regressionsmodels für alle landwirtschaftlichen Betriebe außerhalb Hamburgs und die Berechnung der Mittelwerte für alle landwirtschaftlichen Betriebe außerhalb Hamburgs, um wiederum die Hamburger Zuckerrübenfläche durch Differenzbildung zu enthüllen.

w_m annimmt. Für alle anderen Merkmalsträger erhält die Variable *z_transformiert* den Wert Null (siehe Gleichung 6).

$$(6)\quad z\_transformiert = \begin{cases} z & falls\ w = w_m \\ 0 & sonst \end{cases}$$

Wird die Variable *z_transformiert* in ein Regressionsmodell eingebunden, ist eine Enthüllung des Wertes z_m durch das in Abschnitt 2 beschriebene Vorgehen möglich.

Im Falle des Zuckerrübenanbaus in Hamburg bedeutet dies Folgendes: Aus der Regionaldatenbank weiß ein Datenangreifer, dass es nur einen Zuckerrübenanbauer in Hamburg gibt (siehe Statistische Ämter des Bundes und der Länder, 2011). Mit dieser Information ist es ihm möglich, jede gewünschte im Datensatz enthaltene Information über den Zuckerrübenanbauer zu enthüllen. Ist er zum Beispiel an der Größe der landwirtschaftlichen Fläche insgesamt des Zuckerrübenanbauers interessiert, genügt die Generierung einer Variable, die nur dann den Wert der landwirtschaftlichen Fläche enthält, wenn die Zuckerrübenanbaufläche größer null ist. Diese Variable kann dann als erklärende Variable in ein Regressionsmodell aufgenommen werden.

b) Der Datenangreifer kennt die Intervalle, in denen sich die Merkmalsausprägungen der Variablen w_1 bis w_k einer anzugreifenden Einheit *m* befinden und weiß, dass die Kombination genügt, um die anzugreifende Einheit eindeutig zu identifizieren. Ist die Einheit *m* damit eindeutig identifiziert, lässt sich für jede dem Datenangreifer unbekannte Variable *z* eine transformierte Variable *z_transformiert* erstellen, die nur im Falle der Einheit *m* den Wert z_m und ansonsten Nullen enthält.

5 Fazit und Konsequenzen für die Geheimhaltungsprüfung

Durch die Schwerpunkteigenschaft der Regressionsgerade ist es einem Datenangreifer grundsätzlich möglich mit Hilfe der geschätzten Regressionskoeffizienten und ihm bekannten Mittelwerten einen geheim zu haltenden Mittelwert und damit eine geheim zu haltende Merkmalssumme bzw. einen geheim zu haltenden Einzelwert offen zu legen. Im Gegensatz zu den von Ronning et al. (2010) aufgezeigten Enthüllungsmöglichkeiten, ist mit der hier gezeigten Methode nicht nur ein Angriff auf die abhängige Variable sondern auch auf die unabhängigen Variablen eines Regressionsmodells möglich. Das Enthüllungsrisiko besteht dabei nicht nur

durch gezielt durch den Datenangreifer erzeugte strategische Variablen sondern auch durch „natürlich“ vorkommende Variablen. Der letztere Fall ist dabei besonders kritisch, da keine – in der Syntax dokumentierten – Änderungen oder Neugenerierungen von Variablen notwendig sind.

Um Enthüllungen aufgrund der Schwerpunkteigenschaft zu verhindern, ergeben sich für die manuelle bzw. für eine eventuell automatisierte Geheimhaltungsprüfung in den Forschungsdatenzentren folgende Konsequenzen:

a) Die Erzeugung von strategischen Variablen muss (auch im Hinblick auf die von Ronning et al. 2010 gezeigten Enthüllungsrisiken durch strategische Dummys bzw. künstlich generierte Ausreißer) im Rahmen der Syntax-Kontrolle unterbunden werden.

b) Metrische Variablen, die in ein Regressionsmodell einfließen, müssen für die untersuchte Untergruppe an Einheiten auf Mindestfallzahl der Nicht-Null-Werte, Dominanz bzw. Konzentration und eventuelle Geheimhaltungen durch Sekundärsperrungen in bisherigen Veröffentlichungen geprüft werden. Ist für eine Variable des Regressionsmodells die Merkmalssumme geheim zu halten, ist der geschätzte Koeffizient dieser Variable zu sperren. Um Rückrechnungen des Koeffizienten zu verhindern, ist zusätzlich die Sperrung des ggf. angegebenen Standardfehlers sowie des Konfidenzintervalls notwendig.

Als Ausblick bleibt die Aussage von Ronning et al. (2010) auch weiterhin bestehen, dass die bereits bekannten Enthüllungsmöglichkeiten durch multivariate Analyseverfahren nur ein kleiner Ausschnitt dessen sind, was möglich ist, „vor allem wenn mathematisch-statistisch Begabte sich mit diesem "Problem" beschäftigen“ (Ronning et al. 2010, S. 46). Die Aufdeckung und Verhinderung von Enthüllungsmöglichkeiten durch multivariate Analyseverfahren ist vielmehr ein laufender Prozess, welcher in den Forschungsdatenzentren bei der Geheimhaltungsprüfung berücksichtig werden muss.

Literatur

Brandt M, Zwick M (2009) infinitE – Eine informationelle Infrastruktur für das E-Science Age. Verbesserung des Mikrodatenzugangs durch „Remote Access“. Wirtschaft und Statistik 7/2009: 670 – 675

Giessing S, Dittrich S (2006) Tabellengeheimhaltung im Statistischen Verbund - ein Verfahrensvergleich am Beispiel der Umsatzsteuerstatistik. Wirtschaft und Statistik 8/2006: 805 - 814

Heinze S, Vogel A (2010) The AFiD-Panel Agriculture: New Potential for Agricultural Research. Schmollers Jahrbuch/ Journal of Applied Social Science Studies 130, im Erscheinen

Reznek AP (2003) Disclosure Risks in Cross-Section Regression Models. American Statistical Association 2003, Proceedings of the Section on Government Statistics and Section on Social Statistics: 3444 - 3451

Reznek AP, Riggs TL (2005) Disclosure Risks in Releasing Output Based on Regression Residuals. American Statistical Association 2005, Proceedings of the Section on Government Statistics and Section on Social Statistics: 1397 - 1404

Ronning G, Bleninger P, Drechsler J, Gürke C (2010) Remote Access - Eine Welt ohne Mikrodaten?. IAW Discussion Papers, Tübingen, Nr. 66

Statistisches Amt für Hamburg und Schleswig-Holstein (2006) Agrarstruktur in Hamburg 2003. Ausgewählte Strukturdaten, Statistischer Bericht, C IV 9 / S - 2003 H

Statistische Ämter des Bundes und der Länder (2011) Landwirtschaftliche Betriebe mit Ackerland und deren Ackerfläche nach Fruchtarten. Erhebungsjahr: 2003. regionale Tiefe: Bundesländer, www.regionalstatistik.de, Stand: 24.01.2011

Zühlke S, Zwick M, Scharnhorst S, Wende T (2004) The research data centres of the Federal Statistical Office and the statistical offices of the Länder. Schmollers Jahrbuch/ Journal of Applied Social Science Studies 124: 567 - 578