

**Berichte aus der Statistik**

**Jürgen Buck**

**Datenfusion und Steuersimulation**

**Theorie und Empirie im Rahmen  
des Mikrosimulationsmodells GMOD**

D 100 (Diss. Universität Hohenheim)

**Shaker Verlag  
Aachen 2006**

# Inhaltsverzeichnis

<b>Abbildungsverzeichnis</b>	<b>xi</b>
<b>Tabellenverzeichnis</b>	<b>xiii</b>
<b>Abkürzungsverzeichnis</b>	<b>xv</b>
<b>Symbolverzeichnis</b>	<b>xvii</b>
<b>1 Motivation und Struktur</b>	<b>1</b>
1.1 Allgemeine Bedeutung von Mikrosimulationsmodellen . . . . .	1
1.2 Mikrosimulationsmodelle in der Steuerpolitik . . . . .	2
1.3 Bestehende Modelle und Grenzen der Datenbasis . . . . .	3
1.4 Datenfusion als mögliche Lösung . . . . .	4
1.5 Struktur . . . . .	4
<b>2 Grundlagen der Datenfusion</b>	<b>7</b>
2.1 Grundidee . . . . .	7
2.2 Abgrenzung . . . . .	8
2.2.1 Exakte Zusammenführung . . . . .	8
2.2.1.1 Unterschiede zur Datenfusion . . . . .	8
2.2.1.2 Einschränkung der Anwendbarkeit . . . . .	9
2.2.1.3 Vorgehen und Anwendungsbereiche . . . . .	9
2.2.2 Regressionsansätze . . . . .	10
2.2.2.1 Idee . . . . .	11
2.2.2.2 Kritik . . . . .	11
2.3 Anforderungen an Größe und Beschaffenheit der Files . . . . .	12
2.3.1 Grundsätzliche Ähnlichkeit der Merkmalsträger in Primär- und Sekundärfile . . . . .	12
2.3.2 Ausreichende Zahl von ähnlichen Beobachtungen im Sekundärfile . . . . .	12
2.4 Distanzfunktionen . . . . .	13
2.4.1 Sinn von Distanzfunktionen . . . . .	13
2.4.2 Mögliche Distanzfunktionen . . . . .	13
2.4.2.1 Absoluter Abstand . . . . .	14
2.4.2.2 Euklidischer Abstand . . . . .	14
2.4.2.3 Quadratischer Abstand . . . . .	15
2.4.2.4 MAHALANOBIS-Distanz . . . . .	15
2.4.3 Fusion unter Verwendung von Propensity Scores . . . . .	16

## Inhaltsverzeichnis

2.4.3.1	Herkunft des Ansatzes . . . . .	16
2.4.3.2	Idee . . . . .	16
2.4.3.3	Balancing Scores . . . . .	17
2.4.3.4	Propensity Scores . . . . .	17
2.5	Homogene Gruppen . . . . .	19
2.6	Monogame vs. polygame Algorithmen . . . . .	20
2.7	Restringierte vs. unrestringierte Datenfusion . . . . .	20
2.7.1	Beispiel zur restringierten bzw. unrestringierten Fusion . . . . .	21
2.7.2	Unrestringierte Datenfusion . . . . .	23
2.7.2.1	Vorgehensweise . . . . .	23
2.7.2.2	Ergebnis . . . . .	23
2.7.3	Restringierte Datenfusion . . . . .	23
2.7.3.1	Vorgehensweise . . . . .	23
2.7.3.2	Ergebnis . . . . .	25
2.7.4	Vergleichende Kritik . . . . .	25
2.8	Weitere Anwendungsmöglichkeiten der Datenfusion . . . . .	26
<b>3</b>	<b>Statistische Eigenschaften der fusionierten Daten</b>	<b>29</b>
3.1	Grundüberlegungen . . . . .	29
3.2	Dichtefunktionen und Momente . . . . .	30
3.2.1	Randverteilungen . . . . .	31
3.2.2	Gemeinsame Verteilungen . . . . .	31
3.2.3	Momente . . . . .	33
3.3	Statistische Eigenschaften bei Annahme einer multivariaten Normalverteilung . . . . .	34
3.3.1	Ausgangsverteilungen . . . . .	34
3.3.2	Verteilung im fusionierten File . . . . .	34
3.3.3	Vergleich mit Verteilung der Grundgesamtheit . . . . .	35
3.3.4	Aussagen über $\Sigma_{XY}$ . . . . .	36
3.3.4.1	Grundgedanken . . . . .	36
3.3.4.2	Univariater Fall . . . . .	36
3.3.4.3	Multivariate gemeinsame Variablen $Z$ , univariate Variablen $X$ und $Y$ . . . . .	39
3.3.4.4	Schlussfolgerungen . . . . .	39
3.4	Fusion nach Propensity-Score-Methode . . . . .	39
3.5	Ergebnis von Simulationsexperimenten . . . . .	40
3.5.1	Simulationsexperiment von GÄBLER . . . . .	40
3.5.2	Simulationsexperiment von RÄSSLER . . . . .	41
3.5.2.1	Design der Simulation . . . . .	42
3.5.2.2	Simulation ohne Vorliegen bedingter Unabhängigkeit . . . . .	43
3.5.2.3	Simulation bei Vorliegen bedingter Unabhängigkeit . . . . .	44
3.5.2.4	Schlussfolgerungen . . . . .	45
3.6	Bedeutung der Annahme bedingter Unabhängigkeit . . . . .	45
3.7	Erfolgskriterien der Datenfusion . . . . .	46
3.7.1	Erste Stufe . . . . .	46
3.7.2	Zweite Stufe . . . . .	47
3.7.3	Dritte Stufe . . . . .	47

3.7.4	Vierte Stufe . . . . .	48
<b>4</b>	<b>Spezielle Verfahren bei weiteren Annahmen oder Zusatzinformationen</b>	<b>49</b>
4.1	Anwendungsgebiete . . . . .	49
4.2	Erweiterter Regressionsansatz . . . . .	50
4.2.1	Idee . . . . .	50
4.2.2	Verfahren . . . . .	50
4.2.3	Erweiterung der Methodik . . . . .	53
4.3	Ansatz von KADANE . . . . .	54
4.3.1	Idee . . . . .	54
4.3.2	Verfahren und statistische Eigenschaften . . . . .	54
4.3.2.1	Regression . . . . .	54
4.3.2.2	Fusion . . . . .	55
4.3.3	Bewertung . . . . .	56
4.4	Multiple Imputation als Verfahren der Datenfusion . . . . .	56
4.4.1	Ursprung der Idee . . . . .	57
4.4.2	Vorgehensweise . . . . .	57
4.4.3	Grobeinteilung der statistischen Eigenschaften der fehlenden Daten . . . . .	58
4.4.4	Übertragung der Methodik auf das Problem der Datenfusion . . . . .	59
4.4.5	Grundlagen der BAYES-Statistik . . . . .	60
4.4.5.1	A-priori-Verteilung . . . . .	60
4.4.5.2	Likelihoodprinzip . . . . .	61
4.4.5.3	A-posteriori-Verteilung . . . . .	62
4.4.6	Nichtiteratives multivariates Imputationsverfahren . . . . .	63
4.4.6.1	Idee . . . . .	63
4.4.6.2	Univariater Fall . . . . .	63
4.4.6.3	Multivariater Fall . . . . .	66
4.4.6.4	Software-Implementierung . . . . .	68
4.4.7	Iterative univariate Imputations . . . . .	68
4.4.7.1	Idee . . . . .	68
4.4.7.2	Verfahren . . . . .	69
4.4.7.3	Software-Implementierung . . . . .	70
4.4.8	Data-Augmentation-Ansatz . . . . .	71
4.4.8.1	Idee . . . . .	71
4.4.8.2	Verfahren . . . . .	72
<b>5</b>	<b>Systematisierung und Anwendungsmöglichkeiten für Mikrosimulationsmodelle</b>	<b>75</b>
5.1	Allgemeine Systematisierung und Würdigung . . . . .	75
5.1.1	Systematisierung . . . . .	75
5.1.2	Anwendungsgebiete klassischer Verfahren . . . . .	76
5.1.2.1	Vorliegen bedingter Unabhängigkeit . . . . .	76
5.1.2.2	Keine bedingte Unabhängigkeit vorhanden . . . . .	76
5.1.3	Anwendungsgebiete erweiterter Verfahren . . . . .	76
5.1.3.1	Vorliegen bedingter Unabhängigkeit . . . . .	76
5.1.3.2	Keine bedingte Unabhängigkeit . . . . .	77
5.1.4	Anwendbarkeit der Verfahrenstypen . . . . .	77

## Inhaltsverzeichnis

5.1.5	Bewertung der klassischen Verfahren . . . . .	77
5.1.5.1	Einfacher Regressionsansatz . . . . .	77
5.1.5.2	Nearest-Neighbour-Verfahren . . . . .	78
5.1.5.3	Propensity-Score-Ansätze . . . . .	79
5.1.6	Bewertung der erweiterten Verfahren . . . . .	80
5.1.6.1	Erweiterter Regressionsansatz . . . . .	80
5.1.6.2	Ansätze von KADANE und MORIARITY/SCHEUREN . . . . .	80
5.1.6.3	Multiple Imputation . . . . .	80
5.2	Anwendbarkeit für Mikrosimulationsmodelle . . . . .	81
5.2.1	Möglichkeiten . . . . .	81
5.2.2	Grenzen . . . . .	81
5.2.3	Mögliche Algorithmen und Erfolgsvoraussetzungen . . . . .	82
5.2.3.1	Geeignete Algorithmen . . . . .	82
5.2.3.2	Wahl der gemeinsamen Variablen . . . . .	82
<b>6</b>	<b>Mikrosimulationsmodell GMOD</b>	<b>83</b>
6.1	Mögliche Datenbasis für Mikrosimulationsmodelle . . . . .	83
6.1.1	Sozio-ökonomisches Panel (SOEP) . . . . .	83
6.1.2	Faktisch anonymisierte Steuerdaten (FAST 98) . . . . .	84
6.1.3	Einkommens- und Verbrauchsstichprobe . . . . .	84
6.2	Mikrosimulationsmodell GMOD . . . . .	85
6.2.1	Evolution des Modells . . . . .	85
6.2.2	Datenbasis und Funktionsweise . . . . .	85
<b>7</b>	<b>Ergänzung der Datenbasis des GMOD</b>	<b>87</b>
7.1	Einschränkungen der gegenwärtigen Datenbasis . . . . .	87
7.2	Überblick über relevante Aspekte der FAST-Daten . . . . .	87
7.2.1	Stichprobe und Grundgesamtheit . . . . .	87
7.2.2	Fälle ohne Steuerklasse . . . . .	88
7.2.3	Steuerklasse und Splitting . . . . .	89
7.3	Werbungskosten bei nichtselbständiger Tätigkeit . . . . .	90
7.3.1	Modellierung im GMOD . . . . .	90
7.3.2	Situation im FAST . . . . .	90
7.3.2.1	Werbungskostenpauschale . . . . .	90
7.3.2.2	Bedeutung Fahrtkosten . . . . .	94
7.3.3	Verbesserung der GMOD-Datenbasis . . . . .	97
7.3.3.1	Funktionale Zusammenhänge . . . . .	97
7.3.3.2	Konsequenzen für die Wahl der Methodik . . . . .	102
7.3.3.3	Ergebnisse . . . . .	105
7.4	Sonderausgaben für Steuerberatung . . . . .	112
7.4.1	Modellierung im GMOD . . . . .	112
7.4.2	Situation im FAST-Datenbestand . . . . .	112
7.4.3	Verbesserung der GMOD-Datenbasis . . . . .	113
7.4.3.1	Funktionale Zusammenhänge . . . . .	113
7.4.3.2	Konsequenzen für die Wahl der Methodik . . . . .	113
7.4.3.3	Ergebnisse . . . . .	114

7.5	Sonderausgaben für Beiträge und Spenden . . . . .	118
7.5.1	Modellierung im GMOD . . . . .	118
7.5.2	Situation im FAST-Datenbestand . . . . .	119
7.5.3	Verbesserung der GMOD-Datenbasis . . . . .	120
7.5.3.1	Funktionale Zusammenhänge . . . . .	120
7.5.3.2	Konsequenzen für die Wahl der Methodik . . . . .	121
7.5.3.3	Ergebnisse . . . . .	121
7.6	Zugehörigkeit zu einer Konfession (Kirchensteuerpflicht) . . . . .	122
7.6.1	Modellierung im GMOD . . . . .	122
7.6.2	Situation im FAST-Datenbestand . . . . .	130
7.6.3	Verbesserung der GMOD-Datenbasis . . . . .	131
7.6.3.1	Methodik . . . . .	131
7.6.3.2	Ergebnisse . . . . .	131
<b>8</b>	<b>Fazit und Ausblick</b>	<b>135</b>
	<b>Literaturverzeichnis</b>	<b>137</b>