

Design der Haushaltsstichprobe für den Zensus 2011

Ausschuss für Regionalstatistik
DStatG und VDSt

Ralf Münnich

Universität Trier, Fachbereich IV, Wirtschafts- und Sozialstatistik
in Zusammenarbeit mit Siegfried Gabler und Matthias Ganninger, GESIS

Düsseldorf, 18. Januar 2010

Ausgangspunkt: Schäfer-Design

Gemeinde ab 10.000 EW Ziehung von 550 Anschriften

Gemeinde unter 10.000 EW Ziehung von *Anteil der EW der Gemeinde am Kreis* an 550 Anschriften

Schäfer, J. (2004): Ergänzende Verfahren für einen künftigen registergestützten Zensus. In: Statistische Analysen und Studien Nordrhein-Westfalen, Band 17, S. 20-27, Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen.

Variationen

- ▶ Berücksichtigung von Stadtteilen
- ▶ Berücksichtigung von Verbandsgemeinden

Ausgangspunkt: Schäfer-Design

Gemeinde ab 10.000 EW Ziehung von 550 Anschriften

Gemeinde unter 10.000 EW Ziehung von *Anteil der EW der Gemeinde am Kreis* an 550 Anschriften

Schäfer, J. (2004): Ergänzende Verfahren für einen künftigen registergestützten Zensus. In: Statistische Analysen und Studien Nordrhein-Westfalen, Band 17, S. 20-27, Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen.

Variationen

- ▶ Berücksichtigung von Stadtteilen
- ▶ Berücksichtigung von Verbandsgemeinden

Welche Besonderheiten müssen berücksichtigt werden?

- ▶ Welche gesetzlichen Bestimmungen sind zu beachten?
 - ▶ Ziehung von Anschriften
 - ▶ 9,1% der Bevölkerung
- ▶ Weitere zu beachtende Besonderheiten
 - ▶ Effizienz versus Machbarkeit
 - ▶ Berücksichtigung der hierarchischen Struktur (Ziel 2)
 - ▶ Beurteilungskriterien: RRMSE
 - ▶ Multikriterielle Betrachtung
- ▶ Wahl eines Designs, das die Präzisionsanforderungen einhält
 - ▶ Welches Schätzverfahren wird zu Grunde gelegt?
 - ▶ Wie werden die Zielvorgaben präzisiert?
 - ▶ Wo werden *tatsächlich* Stichproben gezogen?

Welche Besonderheiten müssen berücksichtigt werden?

- ▶ Welche gesetzlichen Bestimmungen sind zu beachten?
 - ▶ Ziehung von Anschriften
 - ▶ 9,1% der Bevölkerung
- ▶ Weitere zu beachtende Besonderheiten
 - ▶ Effizienz versus Machbarkeit
 - ▶ Berücksichtigung der hierarchischen Struktur (Ziel 2)
 - ▶ Beurteilungskriterien: RRMSE
 - ▶ Multikriterielle Betrachtung
- ▶ Wahl eines Designs, das die Präzisionsanforderungen einhält
 - ▶ Welches Schätzverfahren wird zu Grunde gelegt?
 - ▶ Wie werden die Zielvorgaben präzisiert?
 - ▶ Wo werden *tatsächlich* Stichproben gezogen?

Welche Besonderheiten müssen berücksichtigt werden?

- ▶ Welche gesetzlichen Bestimmungen sind zu beachten?
 - ▶ Ziehung von Anschriften
 - ▶ 9,1% der Bevölkerung
- ▶ Weitere zu beachtende Besonderheiten
 - ▶ Effizienz versus Machbarkeit
 - ▶ Berücksichtigung der hierarchischen Struktur (Ziel 2)
 - ▶ Beurteilungskriterien: RRMSE
 - ▶ Multikriterielle Betrachtung
- ▶ Wahl eines Designs, das die Präzisionsanforderungen einhält
 - ▶ Welches Schätzverfahren wird zu Grunde gelegt?
 - ▶ Wie werden die Zielvorgaben präzisiert?
 - ▶ Wo werden *tatsächlich* Stichproben gezogen?

Referenzschätzverfahren $\hat{\tau}_{Y,GREG}$

$$\hat{\tau}_{Y,GREG} = \sum_{h=1}^H N_h \cdot \left(\bar{y}_h + (\bar{\mathbf{X}}_h - \bar{\mathbf{x}}_h)' \cdot \hat{\beta} \right)$$

Dabei ist

- N_h Anzahl Anschriften in h -ter Schicht
- \bar{y}_h Stichprobenmittel der y -Werte in h -ter Schicht
- $\bar{\mathbf{X}}_h$ Vektor der Mittelwerte der x -Werte in der Gesamtheit in h -ter Schicht
- $\bar{\mathbf{x}}_h$ Vektor der Mittelwerte der x -Werte in der Stichprobe in h -ter Schicht
- $\hat{\beta}$ Schätzung des Regressionsparameters.

mit
$$V(\hat{\tau}_y) = \sum_{h=1}^H N_h^2 \cdot \frac{S_{h,Y}^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h} \right) \cdot (1 - \rho^2)$$

Referenzschätzverfahren $\hat{\tau}_{Y,GREG}$

$$\hat{\tau}_{Y,GREG} = \sum_{h=1}^H N_h \cdot \left(\bar{y}_h + (\bar{\mathbf{X}}_h - \bar{\mathbf{x}}_h)' \cdot \hat{\beta} \right)$$

Dabei ist

- N_h Anzahl Anschriften in h -ter Schicht
- \bar{y}_h Stichprobenmittel der y -Werte in h -ter Schicht
- $\bar{\mathbf{X}}_h$ Vektor der Mittelwerte der x -Werte in der Gesamtheit in h -ter Schicht
- $\bar{\mathbf{x}}_h$ Vektor der Mittelwerte der x -Werte in der Stichprobe in h -ter Schicht
- $\hat{\beta}$ Schätzung des Regressionsparameters.

mit
$$V(\hat{\tau}_y) = \sum_{h=1}^H N_h^2 \cdot \frac{S_{h,Y}^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h} \right) \cdot (1 - \rho^2)$$

Präzisionsanforderungen im Zensus 2011

Ziel 1: Schätzungen nur bei Gemeinden ab 10.000 EW

- ▶ Gemeinden ab 10.000 EW: $RRMSE(\hat{\tau}_{Z, <area>}) \leq 0,5\%$
- ▶ Stadtteile von Großstädten: $RRMSE(\hat{\tau}_{Z, <area>}) \leq 0,5\%$

Ziel 2: Betrachtet wird bei $\frac{\tau_{Y, <area>}}{\tau_{Z, <area>}} \approx \rho$ mit $\rho \geq \frac{1}{15}$:

- ▶ Gemeinden ab 10.000 EW: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{\rho}$
- ▶ Stadtteile von Großstädten: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{\rho}$
- ▶ Kreise: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{\rho}$
- ▶ VBG in RLP: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{\rho}$

Ziel:	1	2	2	2	2	2	2
ρ (in %):	100	80	50	30	20	10	6,7
$RRMSE_{\max}$ (in %):	0,5	1,25	2	3,33	5	10	15

Präzisionsanforderungen im Zensus 2011

Ziel 1: Schätzungen nur bei Gemeinden ab 10.000 EW

- ▶ Gemeinden ab 10.000 EW: $RRMSE(\hat{\tau}_{Z, <area>}) \leq 0,5\%$
- ▶ Stadtteile von Großstädten: $RRMSE(\hat{\tau}_{Z, <area>}) \leq 0,5\%$

Ziel 2: Betrachtet wird bei $\frac{\tau_{Y, <area>}}{\tau_{Z, <area>}} \approx p$ mit $p \geq \frac{1}{15}$:

- ▶ Gemeinden ab 10.000 EW: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{p}$
- ▶ Stadtteile von Großstädten: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{p}$
- ▶ Kreise: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{p}$
- ▶ VBG in RLP: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{p}$

Ziel:	1	2	2	2	2	2	2
p (in %):	100	80	50	30	20	10	6,7
$RRMSE_{max}$ (in %):	0,5	1,25	2	3,33	5	10	15

Präzisionsanforderungen im Zensus 2011

Ziel 1: Schätzungen nur bei Gemeinden ab 10.000 EW

- ▶ Gemeinden ab 10.000 EW: $RRMSE(\hat{\tau}_{Z, <area>}) \leq 0,5\%$
- ▶ Stadtteile von Großstädten: $RRMSE(\hat{\tau}_{Z, <area>}) \leq 0,5\%$

Ziel 2: Betrachtet wird bei $\frac{\tau_{Y, <area>}}{\tau_{Z, <area>}} \approx p$ mit $p \geq \frac{1}{15}$:

- ▶ Gemeinden ab 10.000 EW: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{p}$
- ▶ Stadtteile von Großstädten: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{p}$
- ▶ Kreise: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{p}$
- ▶ VBG in RLP: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{p}$

Ziel:	1	2	2	2	2	2	2
p (in %):	100	80	50	30	20	10	6,7
$RRMSE_{max}$ (in %):	0,5	1,25	2	3,33	5	10	15

Präzisionsanforderungen im Zensus 2011

Ziel 1: Schätzungen nur bei Gemeinden ab 10.000 EW

- ▶ Gemeinden ab 10.000 EW: $RRMSE(\hat{\tau}_{Z, <area>}) \leq 0,5\%$
- ▶ Stadtteile von Großstädten: $RRMSE(\hat{\tau}_{Z, <area>}) \leq 0,5\%$

Ziel 2: Betrachtet wird bei $\frac{\tau_{Y, <area>}}{\tau_{Z, <area>}} \approx p$ mit $p \geq \frac{1}{15}$:

- ▶ Gemeinden ab 10.000 EW: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{p}$
- ▶ Stadtteile von Großstädten: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{p}$
- ▶ Kreise: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{p}$
- ▶ VBG in RLP: $RRMSE(\hat{\tau}_{Y, <area>}) \leq \frac{1}{p}$

Ziel:	1	2	2	2	2	2	2
p (in %):	100	80	50	30	20	10	6,7
$RRMSE_{\max}$ (in %):	0,5	1,25	2	3,33	5	10	15

Einteilung in Sampling Points (SMP)

- Typ 0 (SDT):** Stadtteile ab 200.000 Einwohnern aus Städten mit mindestens 400.000 Einwohnern;
- Typ 1 (GEM):** Gemeinden und Städte mit mindestens 10.000 Einwohnern, sofern sie nicht zum Typ 0 gehören;
- Typ 2 (VBG):** Kleine Gemeinden (unter 10.000 Einwohnern) innerhalb eines Gemeindeverbands beziehungsweise einer Verbandsgemeinde werden zusammengefasst, sofern sie in der Summe mindestens 10.000 Einwohner betragen;
- Typ 3 (KRS):** Zusammenfassung aller Gemeinden eines Kreises, die bis dahin noch keinem Typ zugeordnet wurden.

- ▶ Berücksichtigung der Präzisionsanforderungen zu Ziel 1 und 2
- ▶ Berücksichtigung von Sonderproblemen: Rheinland-Pfalz

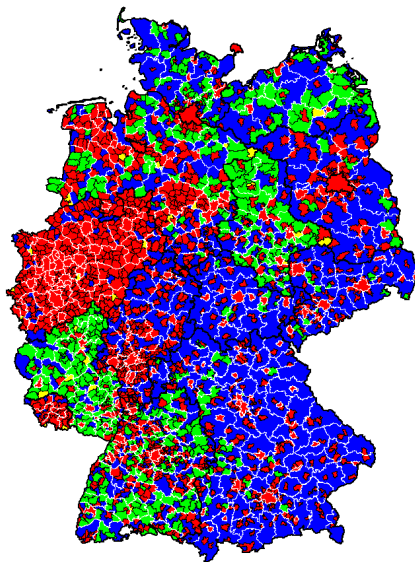
SMPs in Deutschland

SMP 1

SMP 2

SMP 3

Stadtteile nicht in der Karte; gelbe Bezirke sind wegen Gemeindereform nicht zuordenbar



Auswahlsätze von Anschriften

Schäfer-Design	Schäfer (in %)		SMP-Typ (abs. Hfk)				Summe
	GEM	SMP	SDT	GEM	VBG	KRS	
Baden-Württemberg	6.15	9.54	2	244	126	35	407
Bayern	5.16	6.19	8	216	30	71	325
Berlin	2.19	2.19	12	0	0	0	12
Brandenburg	6.71	7.84	0	71	5	14	90
Bremen	1.61	1.61	3	1	0	0	4
Hamburg	1.56	1.56	7	0	0	0	7
Hessen	7.26	7.78	3	168	0	21	192
Mecklenburg-Vorpommern	4.87	9.56	0	24	30	12	66
Niedersachsen	5.67	7.83	2	205	68	34	309
Nordrhein-Westfalen	5.12	5.26	12	339	0	17	368
Rheinland-Pfalz	3.22	8.94	0	46	122	20	188
Saarland	7.50	7.94	0	40	0	5	45
Sachsen	5.86	7.42	4	69	13	22	108
Sachsen-Anhalt	6.12	9.63	0	59	30	11	100
Schleswig-Holstein	4.21	8.21	0	53	52	11	116
Thüringen	4.74	5.99	0	33	6	17	56
Deutschland	5.34	7.13	53	1568	482	290	2393

Stichprobendesign als multikriterielles Optimierungsproblem

- ▶ Es sollten alle Genauigkeitsanforderungen eingehalten werden
- ▶ Zensus soll flächendeckend besser als Mikrozensus sein
- ▶ Das Design sollte MSE-minimale Schätzungen unterstützen
- ▶ Das Design sollte *robust* sein
- ▶ Es sollten statistische Modellierungen möglich sein:
Sozio-ökonomische und Small Area-Modelle
- ▶ Vermeidung allzu *ungleicher* Behandlung der Bevölkerung

MGG-Ansatz (naiv): Anwendung von

$$\min_{n_{\langle 1 \rangle}, \dots, n_{\langle G \rangle}} \max_{g=1, \dots, G} \left(\frac{\sqrt{V_{\langle g \rangle}(\hat{\tau}_{\text{GREG}})}}{\tau_{\langle g \rangle}} \right) \leq \zeta$$

ist direkt lösbar, aber im Allgemeinen problematisch

MGG-Design mit fester Allokation

Für die proportionale Allokation erhält man

$$n \geq \frac{N \cdot \Xi_{\text{prop}} \cdot (1 - \varrho^2)}{(N_P \cdot p \cdot \zeta)^2 + \Xi_{\text{prop}} \cdot (1 - \varrho^2)}$$

und für die optimale Allokation

$$n \geq \frac{\Xi_{\text{opt}} \cdot (1 - \varrho^2)}{(N_P \cdot p \cdot \zeta)^2 + \Xi_{\text{prop}} \cdot (1 - \varrho^2)}$$

$$\text{mit } \Xi_{\text{prop}} := \sum_{h=1}^H N_h \cdot S_{h,Y}^2 \text{ und } \Xi_{\text{opt}} := \left(\sum_{h=1}^H N_h \cdot S_{h,Y} \right)^2.$$

$$d_{\text{max}}/d_{\text{min}} : 329,2 \text{ (prop. Allok.) und } 6337,1 \text{ (opt. Allok.)}$$

Aktuelle Festlegungen

Für $p_h := \frac{m_h}{N_h} \leq \frac{n_h}{N_h} \leq \frac{M_h}{N_h} =: P_h$ gelten folgende Festlegungen:

GemGK	SMP-Typ									
	0		1		2 (RLP)		2 ($\bar{R}LP$)		3	
	p_h	P_h	p_h	P_h	p_h	P_h	p_h	P_h	p_h	P_h
I	—	—	—	—	—	—	—	—	0,05	0,05
II	—	—	0,05	0,50	0,05	0,50	0,05	0,05	0,05	0,05
III	—	—	0,04	0,40	0,04	0,40	0,05	0,05	0,05	0,05
IV	0,02	0,40	0,02	0,40	0,02	0,40	0,05	0,05	0,05	0,05

- ▶ Festlegung der GemGK-Grenzen
 I: 0 bis unter 10.000 EW / II: 10.000 bis unter 30.000 EW
 III: 30.000 bis unter 100.000 EW / IV: ab 100.000 EW
- ▶ Alle SMPs werden nach Anschriftengrößenklassen in 8 gleich große Schichten eingeteilt.
- ▶ max / min der Design-Gewichte: 25

Optimierung mit Box-Constraints

Ziel: Minimierung der 2-Norm des RRMSE-Vektors:

$$\|\mathbf{RRMSE}_{\langle \cdot \rangle}(\hat{\tau})\|_2 = \sqrt{\sum_{g=1}^G \text{RRMSE}(\tau_{\langle g \rangle})^2}$$

unter den Nebenbedingungen:

- ▶ Unter- und Obergrenzen der Entnahmeanteile in jedem Sampling Point und jeder Schicht
- ▶ Obergrenze des Entnahmeanteils an Personen in Deutschland

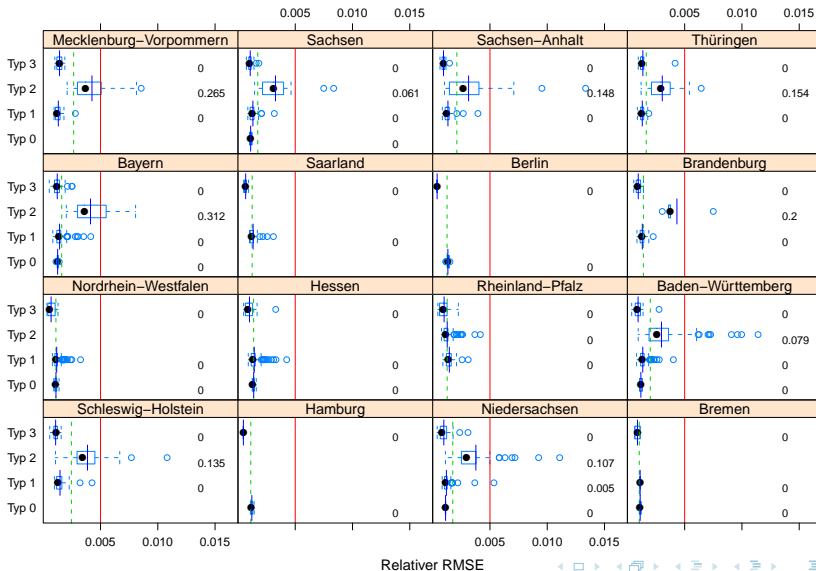
Lösung mittels besonderem *box-constraints optimization*-Algorithmus

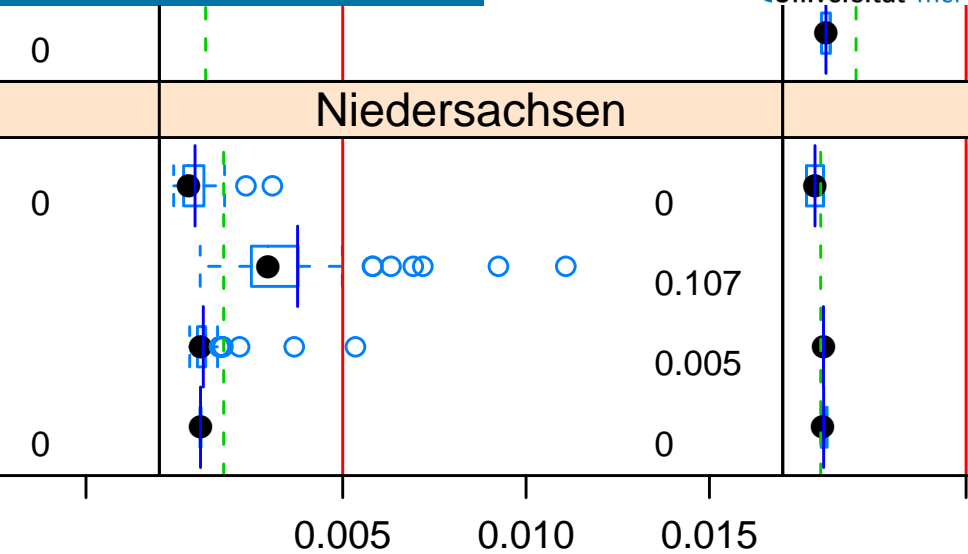
Exakt: Gabler, Ganninger und Münnich (2010), Metrika, sowie

Numerisch: Münnich, Sachs und Wagner (2010), in Fertigstellung

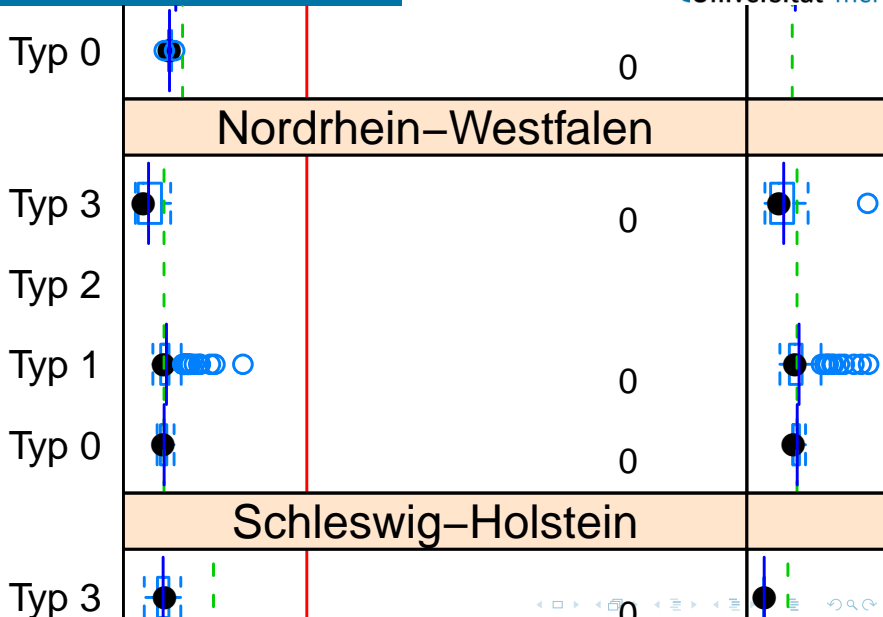
Anmerkung: Supremums-Norm problematisch (siehe zuvor)

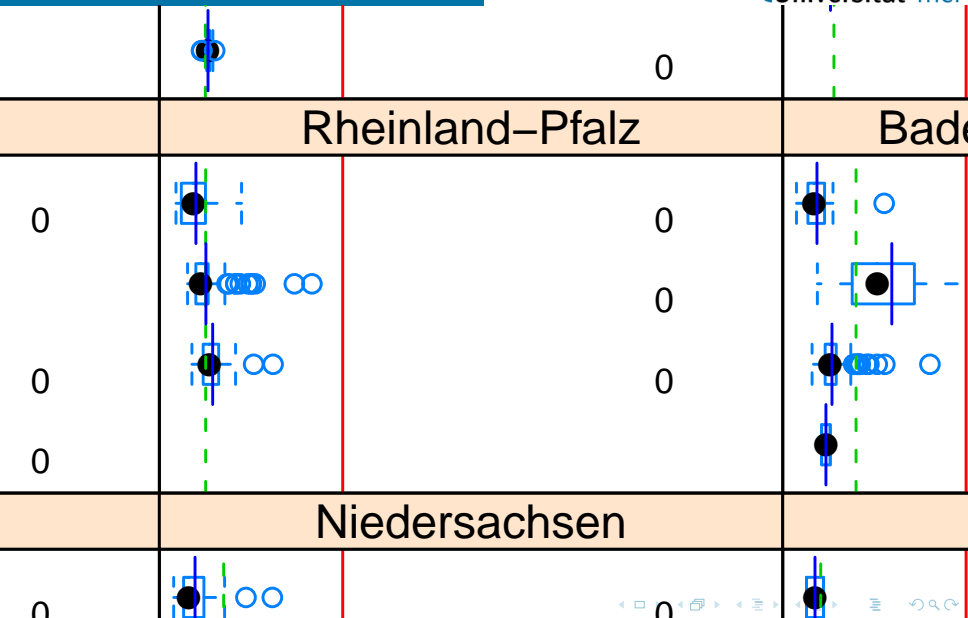
RRMSE bei SMP-optimaler Allokation



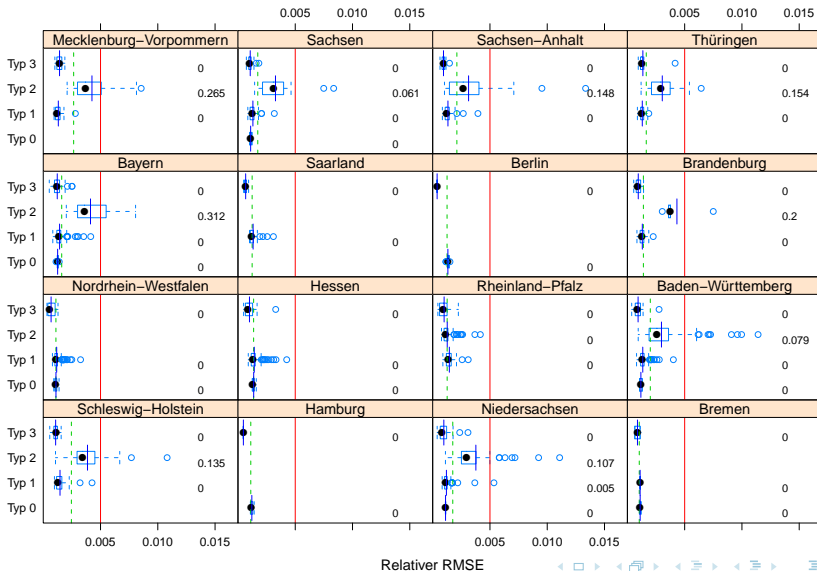


Relativer RMSE





RRMSE bei SMP-optimaler Allokation



Zusammenfassung und Ausblick

- ▶ Die Verwendung der SMPs ermöglicht:
 - ▶ geeignete Berücksichtigung der hierarchischen Struktur
 - ▶ Berücksichtigung der Präzisionsvorgaben
 - ▶ flächendeckende Allokation
 - ▶ Schichtung nach Anschriften-Größenklassen
- ▶ Allokationsproblem via Boxconstraints-Optimierung gelöst
- ▶ Exakte sowie schnelle numerische Algorithmen können herangezogen werden; Lösung allgemeiner verwendbar
- ▶ Ausblick:
 - ▶ Small Area-Verfahren
 - ▶ MSE-Schätzung
 - ▶ Benchmarking für Small Domains
 - ▶ Phase 2 sowie Ziel 2