Universität Trier

Warum reicht eine Stichprobe aus?

Möglichkeiten und Grenzen des Zensus 2011
DAGSTAT Symposium, DIW Berlin

Ralf Münnich

Universität Trier, Fachbereich IV, VWL
Wirtschafts- und Sozialstatistik

Berlin, 08. April 2011

Universität Trier

Warum reicht eine Stichprobe aus?

Universität Trier

Das Team des Zensus-Stichprobenforschungsprojektes

Universität Trier

- Ralf Münnich (CO)
- Jan Pablo Burgard, Jan-Philipp Kolb, Lucie Dostál, Martin Vogt

gesis Leibniz-Institut für Sozialwissenschaften

- Siegfried Gabler
- Matthias Ganninger

Externe Experten:

- Professor Partha Lahiri, University of Maryland, JPSM
- Professor Ulrich Rendtel, FU Berlin

Universität Trier

Aufgaben des Projektes

- Analyse von möglichen Stichprobendesigns
  - Design des Zensustests
  - Variationen des Stichprobendesigns
  - Optimales Stichprobendesign
- Ermittlung einer geeigneten Schätzmethodik

  Ziel 1 Korrektur der Registerdaten
  - Schätzung von Karteileichen und Fehlbeständen
  - Ermittlung der amtlichen Bevölkerungszahl

  Ziel 2 Schätzung von zusätzlichen Personenmerkmalen
- Analyse der Wirkung des Stichprobendesigns auf die Schätzungen
- Entwicklung adäquater Fehlerrechnungen
- Entwicklung geeigneter Qualitätskriterien

Universität Trier

Aufgaben des Projektes

- Analyse von möglichen Stichprobendesigns
  - Design des Zensustests
  - Variationen des Stichprobendesigns
  - Optimales Stichprobendesign
- Ermittlung einer geeigneten Schätzmethodik

  Ziel 1 Korrektur der Registerdaten
  - Schätzung von Karteileichen und Fehlbeständen
  - Ermittlung der amtlichen Bevölkerungszahl

  Ziel 2 Schätzung von zusätzlichen Personenmerkmalen
- Analyse der Wirkung des Stichprobendesigns auf die Schätzungen
- Entwicklung adäquater Fehlerrechnungen
- Entwicklung geeigneter Qualitätskriterien

Universität Trier

Wo passieren Fehler in Erhebungen

Nicht-Stichprobenfehler

- Nonresponse (Unit/Item)
- Messfehler
- Angabefehler
- Registerfehler (Karteileichen und Fehlbeständen)

Stichprobenfehler

- in Folge verschiedener möglicher Stichproben, die gezogen werden können (Zufallsfehler)
- auf Grund von Modellannahmen in bestimmten Verfahren
- auf Grund kleiner (Teil-) Stichprobenumfänge

Universität Trier

Wo passieren Fehler in Erhebungen

Nicht-Stichprobenfehler

- Nonresponse (Unit/Item)
- Messfehler
- Angabefehler
- Registerfehler (Karteileichen und Fehlbeständen)

Stichprobenfehler

- in Folge verschiedener möglicher Stichproben, die gezogen werden können (Zufallsfehler)
- auf Grund von Modellannahmen in bestimmten Verfahren
- auf Grund kleiner (Teil-) Stichprobenumfänge

Universität Trier

Ergebnisse des Zensustests 2001

- Karteileichen: 1,8 %
- Fehlbestände: 1,7 %
- Große Differenzen zwischen den Gemeinden
  Karteileichen:
  - 0,7% in Gemeinden bis 10 000 Einwohner
  - 1,4% in Gemeinden mit 10 000 - 50 000 Einwohner
  - 1,5% in Gemeinden mit 50 000 - 100 000 Einwohner
  - 3,4% in Gemeinden mit 100 000 und mehr Einwohner

  Fehlbestände:
  - 1,3% in Gemeinden bis 10 000 Einwohner
  - 1,3% in Gemeinden mit 10 000 - 50 000 Einwohner
  - 2,1% in Gemeinden mit 50 000 - 100 000 Einwohner
  - 2,4% in Gemeinden mit 100 000 und mehr Einwohner
- Weitere Abhängigkeiten: z.B. Adressgröße

Universität Trier

Präzisionsvorgaben für den Zensus 2011

Ziel 1 Für jede Gemeinde bzw. Stadtteil g:

$$\text{RRMSE}(\hat{\tau}_{Z<g>}) = \frac{\text{RMSE}(\hat{\tau}_{Z<g>})}{\tau_{Z<g>}} \leq 0,005$$

Ziel 2 Ist der Anteil interessierender Beobachtungen an der Zensus-Bevölkerung gleich p, dann muss im Falle von $p \geq 1/15$

$$\text{RRMSE}(\hat{\tau}_{Y<\text{area}>}) \leq \frac{1}{p \cdot 100}$$

gelten.

Anteil Merkmalsträger (in %)	6,7	10	20	30	50	80
Relativer Standardfehler (in %)	15	10	5	3,33	2	1,5

Universität Trier

Ausgangspunkt: Schäfer-Design

Gemeinde ab 10.000 EW Ziehung von 550 Anschriften

Gemeinde unter 10.000 EW Ziehung eines *Anteils* von 550 Anschriften (proportional zu EW einer Gemeinde am Kreis)

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

Universität Trier

Ausgangspunkt: Schäfer-Design

Gemeinde ab 10.000 EW Ziehung von 550 Anschriften

Gemeinde unter 10.000 EW Ziehung eines *Anteils* von 550 Anschriften (proportional zu EW einer Gemeinde am Kreis)

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

Universität Trier

Welche Besonderheiten müssen berücksichtigt werden?

- Welche gesetzlichen Bestimmungen sind zu beachten?
  - Ziehung von Anschriften
  - Etwa 10% 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?

Universität Trier

Welche Besonderheiten müssen berücksichtigt werden?

- Welche gesetzlichen Bestimmungen sind zu beachten?
  - Ziehung von Anschriften
  - Etwa 10% 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?

Universität Trier

Welche Besonderheiten müssen berücksichtigt werden?

- Welche gesetzlichen Bestimmungen sind zu beachten?
  - Ziehung von Anschriften
  - Etwa 10% 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?

Universität Trier

Wie erfolgt die Einteilung in Stichprobenbasiseinheiten

Typ 0 (SDT): Stadtteile ab 200.000 Einwohner (EW) aus Gemeinden mit mindestens 400.000 EW

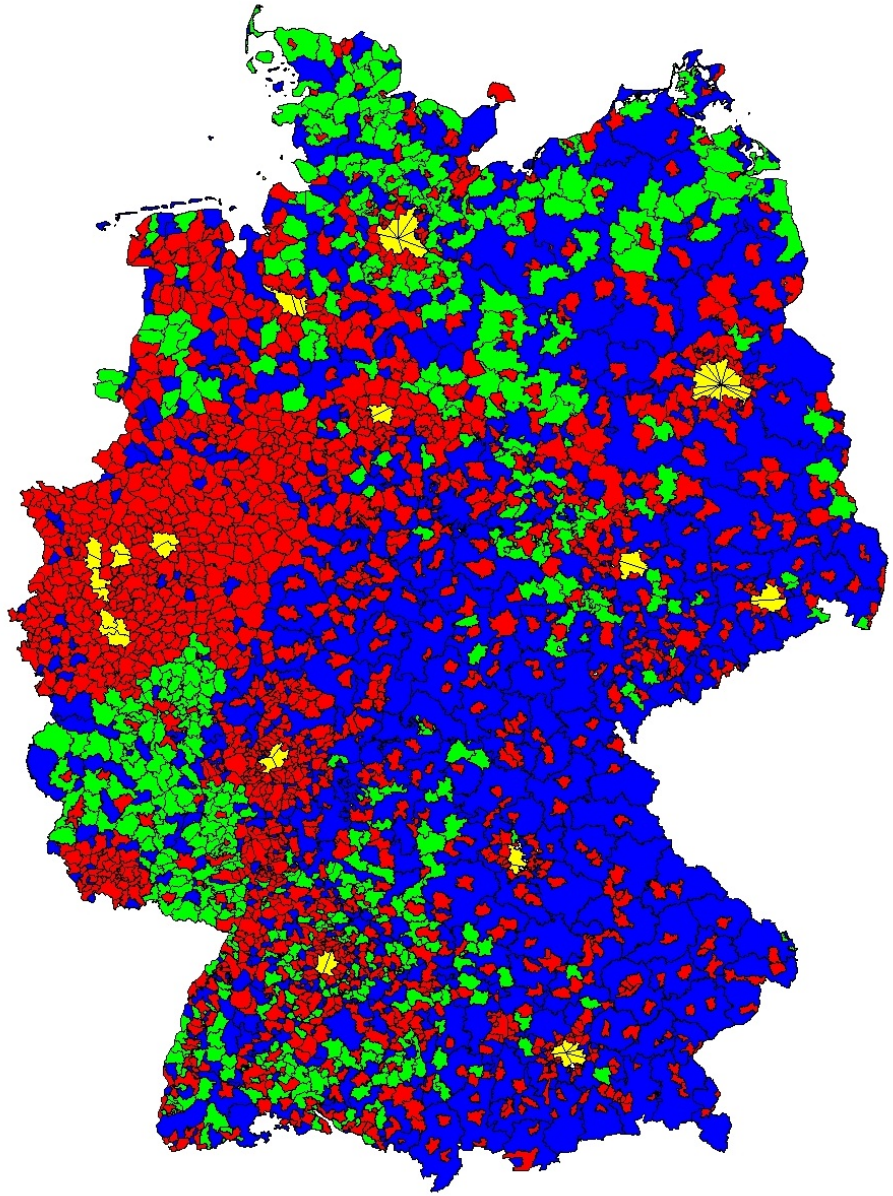
Typ 1 (GEM): Gemeinden mit mindestens 10.000 EW, sofern sie nicht zum Typ 0 gehören

Typ 2 (VBG): Kleine Gemeinden (unter 10.000 EW) innerhalb eines Gemeindeverbands beziehungsweise einer Verbandsgemeinde werden zusammengefasst, sofern sie in der Summe mindestens 10.000 EW betragen

Typ 3 (KRS): Zusammenfassung aller Gemeinden eines Kreises, die bis dahin noch keinem Typ zugeordnet wurden

Universität Trier

SMPs in Deutschland



SMP 0
SMP 1
SMP 2
SMP 3

Universität Trier

Der Optimierungsalgorithmus

- Ausgangspunkt: Neyman-Optimierung
- Unter- und Obergrenzen der Entnahmeanteile der Anschriften in jeder Schicht und in jedem Sampling Point
  - Berücksichtigung von Anschriftengröße als Schichtmerkmal
  - Variierende Grenzen nach Gemeindegrößenklasse
  - Einschränkung der Variabilität der Gewichte (Gelman)
- Obergrenze des Entnahmeanteils an Personen in Deutschland
- Ziel: Simultane Minimierung des RRMSE-Vektors:

$$||\mathbf{RRMSE}_{<\cdot>}(\widehat{\tau})||_2 = \sqrt[2]{\sum_{g=1}^{G} \mathrm{RRMSE}(\tau_{<g>})^2}$$

- Lösung durch iterativen Algorithmus:
  Gabler, Ganninger und Münnich (2010), Metrika (on-line first)
  Münnich, Sachs und Wagner (2011)

Referenz-Schätzer $\widehat{\tau}_{Y,\text{GREG}}$ (je SMP)

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

Notation:

N_h	Anzahl der Anschriften in der h-ten Schicht
$\overline{y}_h$	Stichprobenmittel von y in der h-ten Schicht
$\overline{\mathbf{X}}_h$	Vektor der Mittelwerte von x in der h-ten Schicht
$\overline{\mathbf{x}}_h$	Vektor der Stichprobenmittel von x in der h-ten Schicht
$\widehat{\beta}$	geschätzter Regressionsparameter

$$\text{mit} \qquad \mathrm{V}\left(\widehat{\tau}_y\right) = \sum_{h=1}^{H} N_h^2 \cdot \frac{S_{h,Y}^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot \left(1 - \varrho^2\right)$$

Universität Trier

Referenz-Schätzer $\widehat{\tau}_{Y,\text{GREG}}$ (je SMP)

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

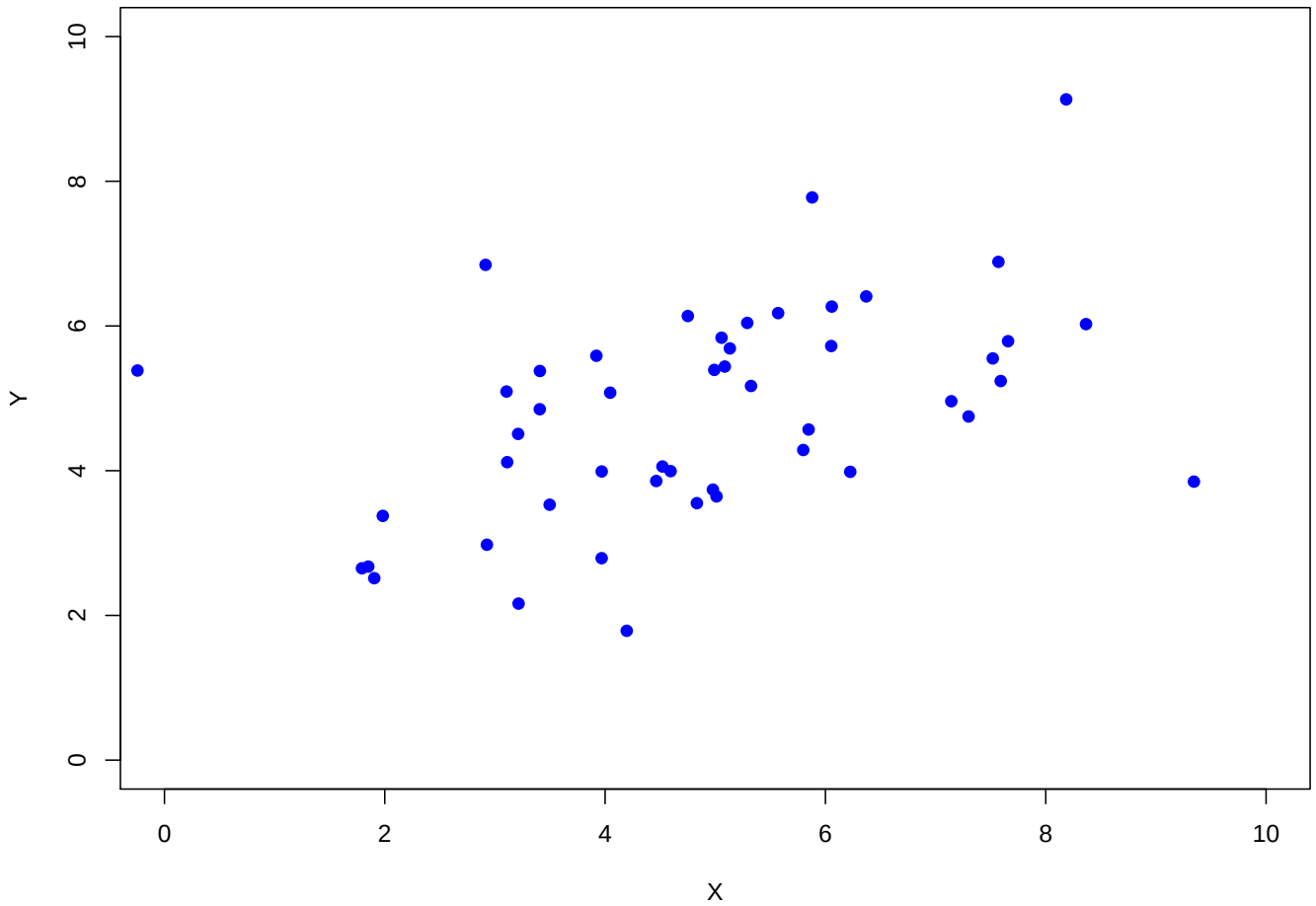
Notation:

N_h	Anzahl der Anschriften in der h-ten Schicht
$\overline{y}_h$	Stichprobenmittel von y in der h-ten Schicht
$\overline{\mathbf{X}}_h$	Vektor der Mittelwerte von x in der h-ten Schicht
$\overline{\mathbf{x}}_h$	Vektor der Stichprobenmittel von x in der h-ten Schicht
$\widehat{\boldsymbol{\beta}}$	geschätzter Regressionsparameter

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

Universität Trier

Wie funktioniert der Referenz-Schätzer

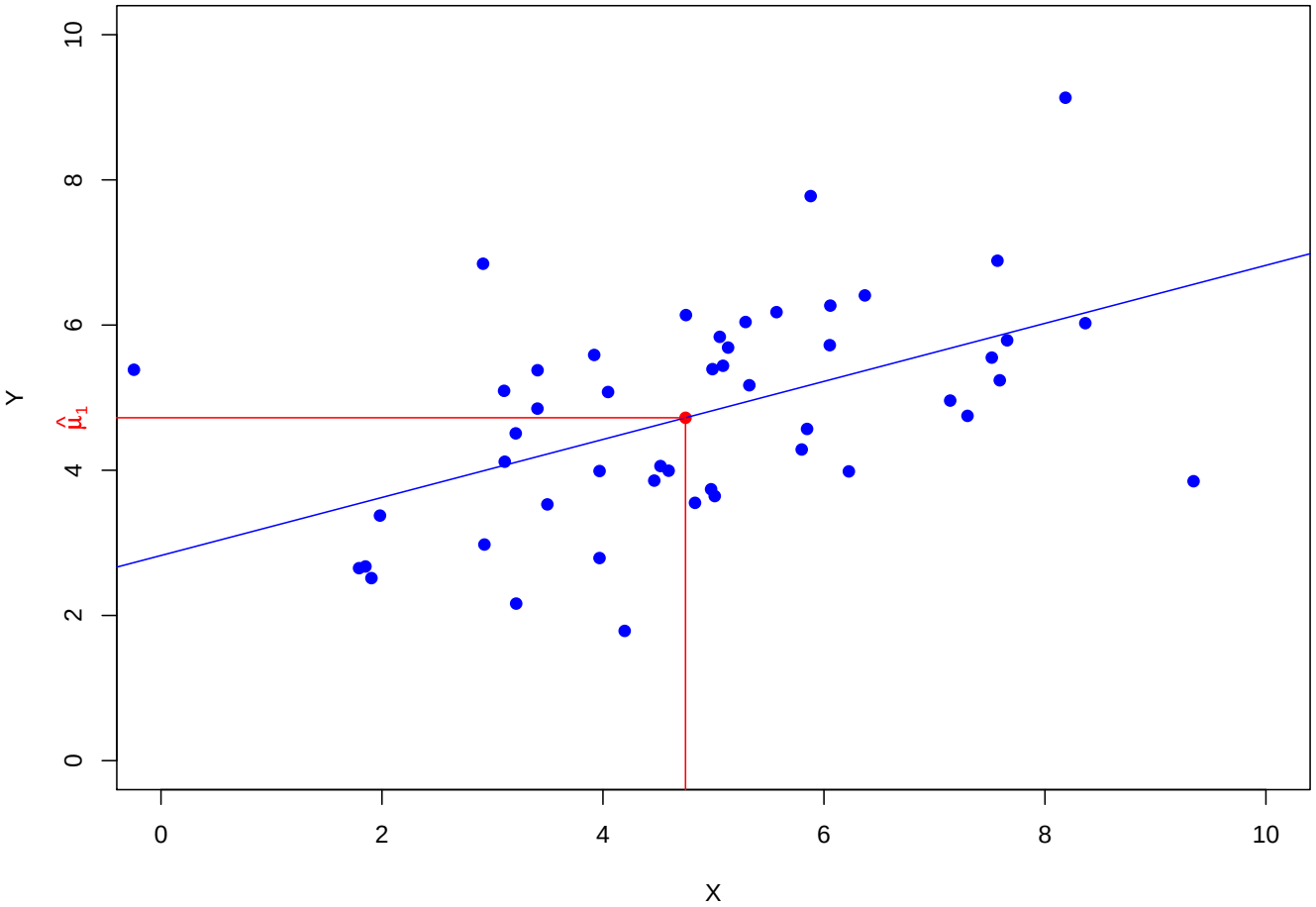


Y: Anzahl der (Zensus-) Personen in einer Anschrift

X: Anzahl der registrierten Personen in einer Anschrift

Universität Trier

Wie funktioniert der Referenz-Schätzer

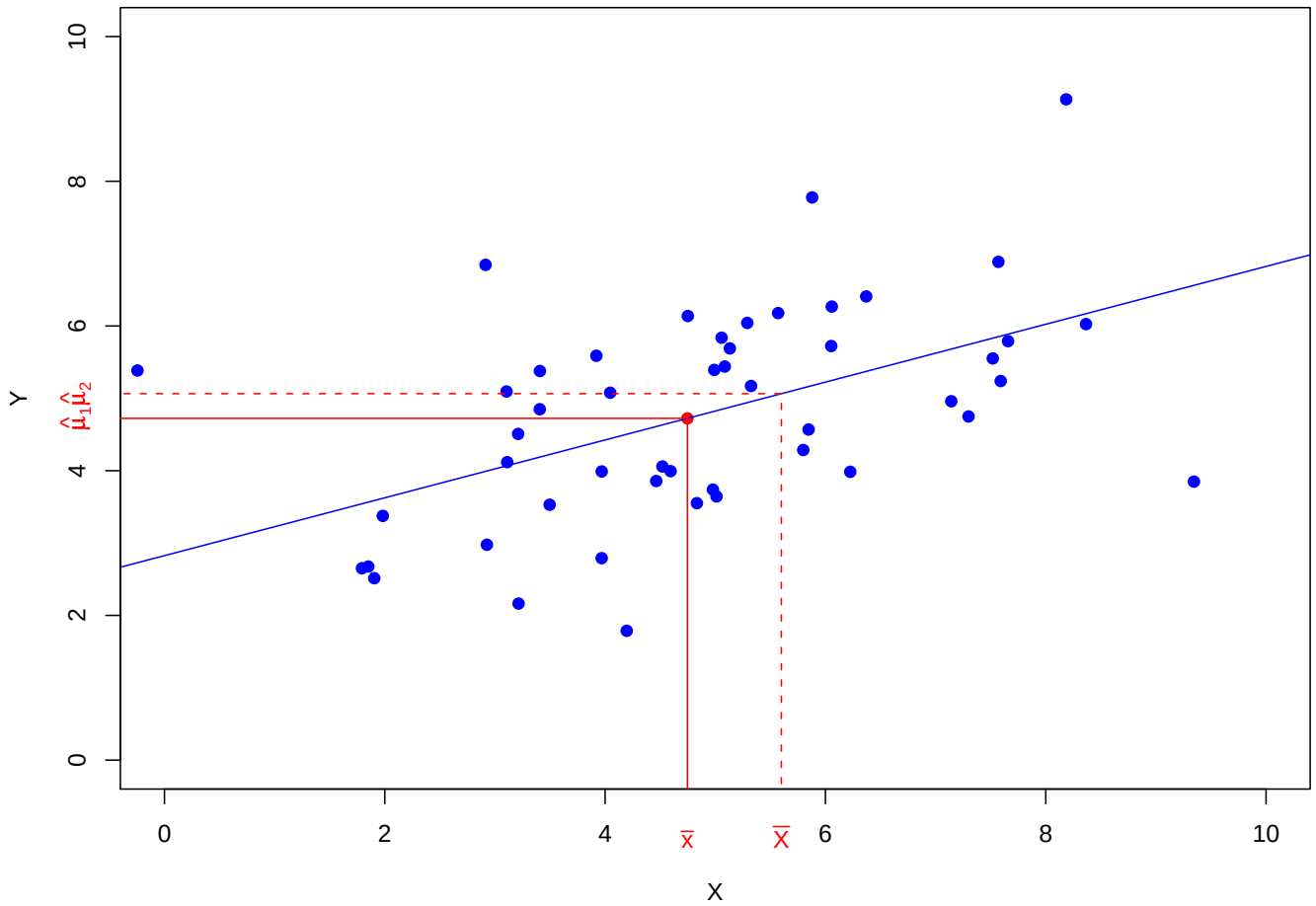


Y: Anzahl der (Zensus-) Personen in einer Anschrift

X: Anzahl der registrierten Personen in einer Anschrift

Universität Trier

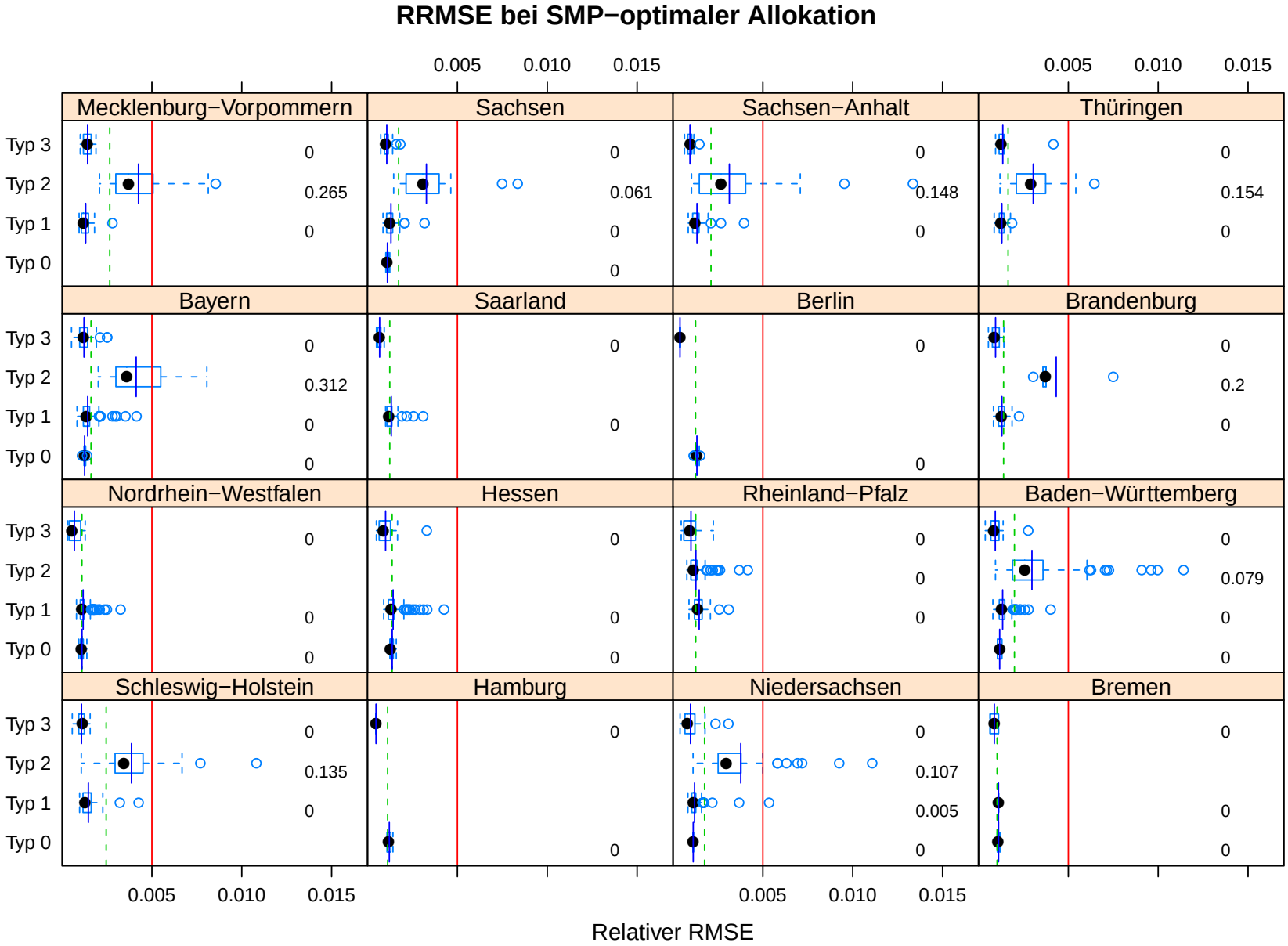
Wie funktioniert der Referenz-Schätzer



Y: Anzahl der (Zensus-) Personen in einer Anschrift

X: Anzahl der registrierten Personen in einer Anschrift

Universität Trier



Universität Trier

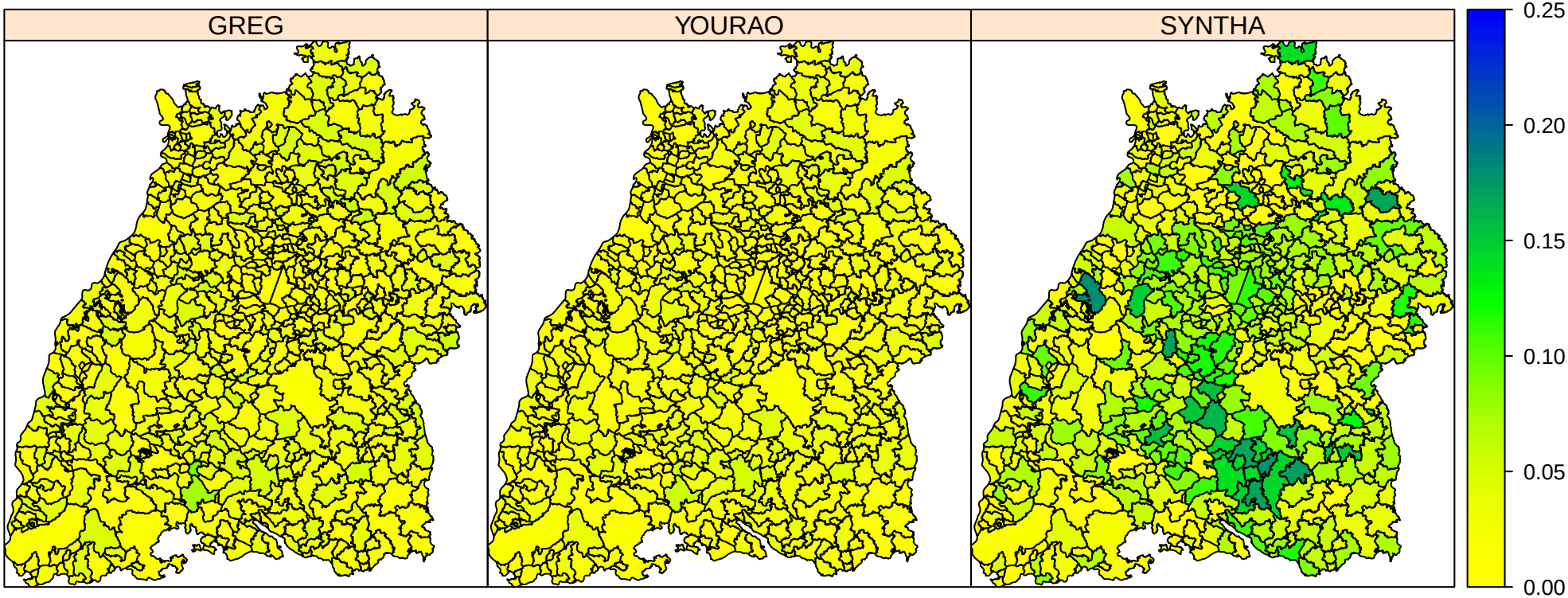
Verwendung von BA-Daten

- Schätzung *überwiegender Lebensunterhalt* (UBS)
- Hilfsvariablen:
  - Anzahl der Personen unter der Anschrift (ADN)
  - Altersklassen (15-24, 25-39, 40-64)
  - Geschlecht
  - Registervariable: Erwerbstätigkeit (EWT; Daten der BA)
- Schätzmethoden:

GREG:	Referenzschätzverfahren
YOURAO:	Modernes Small Area-Verfahren
SYNTHA:	Synthetisches Verfahren (Modell)

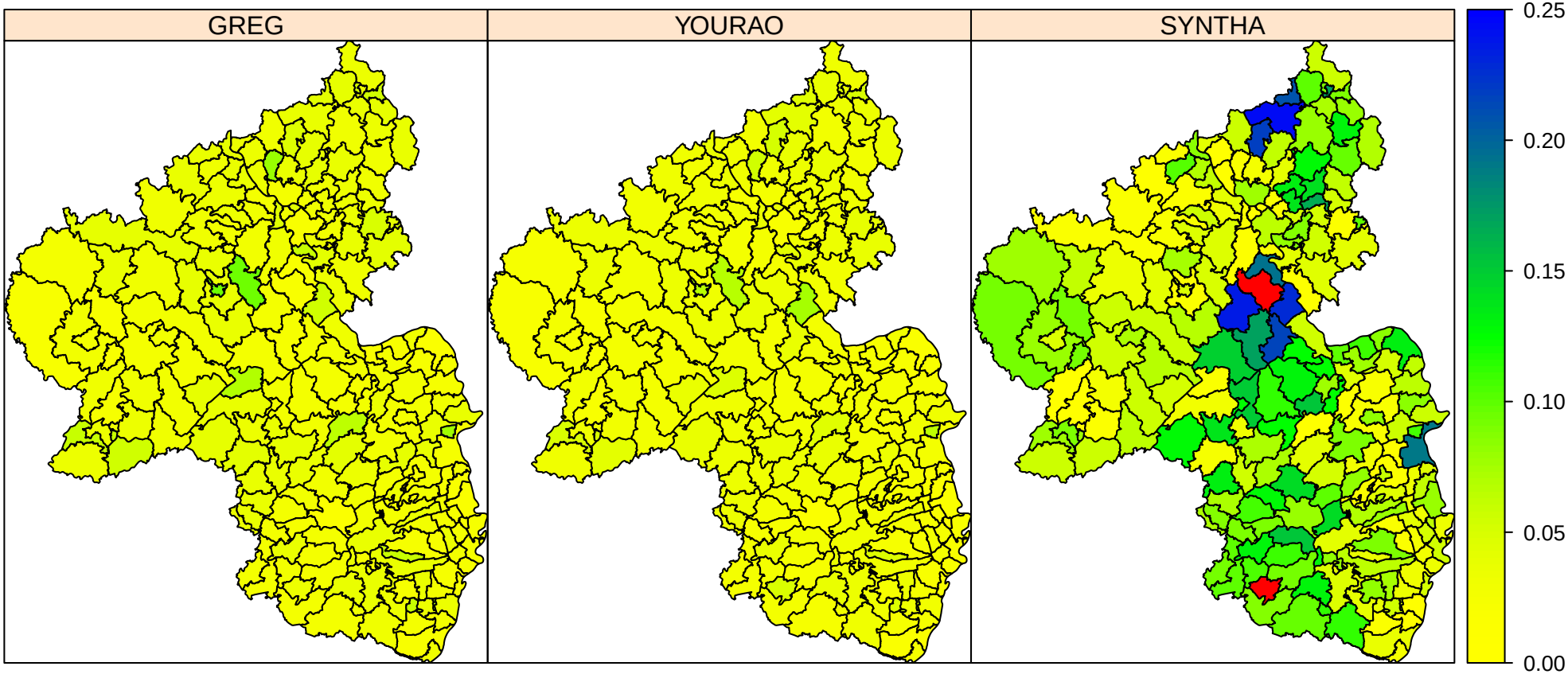
Universität Trier

Schätzung UBS in Baden-Württemberg



Universität Trier

Schätzung UBS in Rheinland-Pfalz



Universität Trier

Zusammenfassung

- Die Stichprobe reicht aus, um die gegebenen (a priori) Qualitätskriterien zu erfüllen
- Optimierung der Stichprobenumfänge als multikriterielles Ziel (über alle SMPs) ist gelungen
- Ziel 1 kann *klassisch* geschätzt werden
- Ziel 2: in tieferen Tabellen werden erweiterte Methoden benötigt

- Veröffentlichungen
  - Statistisches Bundesamt: Haushaltebefragung beim Zensus 2011, `http://www.zensus2011.de`
  - MGGBK: Stichprobendesign und Schätzmethodik im registergestützten Zensus (provisorischer Titel)

- Wir sind schon ganz gespannt auf den Zensus 2011!

Universität Trier

Zusammenfassung

- Die Stichprobe reicht aus, um die gegebenen (a priori) Qualitätskriterien zu erfüllen
- Optimierung der Stichprobenumfänge als multikriterielles Ziel (über alle SMPs) ist gelungen
- Ziel 1 kann *klassisch* geschätzt werden
- Ziel 2: in tieferen Tabellen werden erweiterte Methoden benötigt

- Veröffentlichungen
  - Statistisches Bundesamt: Haushaltebefragung beim Zensus 2011, `http://www.zensus2011.de`
  - MGGBK: Stichprobendesign und Schätzmethodik im registergestützten Zensus (provisorischer Titel)

- Wir sind schon ganz gespannt auf den Zensus 2011!

Universität Trier

Zusammenfassung

- Die Stichprobe reicht aus, um die gegebenen (a priori) Qualitätskriterien zu erfüllen
- Optimierung der Stichprobenumfänge als multikriterielles Ziel (über alle SMPs) ist gelungen
- Ziel 1 kann *klassisch* geschätzt werden
- Ziel 2: in tieferen Tabellen werden erweiterte Methoden benötigt

- Veröffentlichungen
  - Statistisches Bundesamt: Haushaltebefragung beim Zensus 2011, `http://www.zensus2011.de`
  - MGGBK: Stichprobendesign und Schätzmethodik im registergestützten Zensus (provisorischer Titel)

- Wir sind schon ganz gespannt auf den Zensus 2011!