20.02.2014

6. Konferenz für Sozial- und Wirtschaftsdaten
20./21. Februar 2014

Statistische Geheimhaltung der Zensusergebnisse: Wie wurden die Zensusdaten verändert, die in der Auswertungsdatenbank abgerufen werden können?

Dr. Jörg Höhne Abteilung 4 - Gesamtwirtschaft

Gliederung

- Besonderheiten der Ergebnisse zu VÖT2
- Veränderung der Daten
- Pretabulare Geheimhaltung und Mikroaggregation
- Das Verfahren SAFE
- Die Stichprobenanonymisierung
- Die Geheimhaltung des Zensus 2011 – VÖT2
- Interpretation / Besonderheiten

Besonderheiten der Ergebnisse zum Veröffentlichungstermin 2 (VÖT 2)

- erweiterte Informationen durch Einbeziehung weiterer Registerdaten und Stichprobeninformationen
  (z.B. Informationen zur Erwerbstätigkeit,
  Mischung verschiedener Informationsquellen über „Methodenbaukasten“)

- zusätzliche Informationen durch die Haushaltegenerierung
  (Angaben zu Familien, Haushalten)

- Angaben zu Verbindungen zwischen den verschiedenen Einheiten
  (z.B. von Personen und den von ihnen bewohnten Wohnungen/Gebäuden)

Der Veröffentlichungstermin 2 liefert mehr Informationspotential als zum Veröffentlichungstermin 1 verfügbar war.

Die Zensusdatenbank teilt sich weiterhin in ein öffentliches Auswertungssystem (ÖAWS) und ein internes Auswertungssystem (IAWS)

Veränderung der Daten im Rahmen der Aufbereitung für das ÖAWS

Im Rahmen des Zensus wurden verschiedenste Informationsquellen zusammengeführt und zusätzliche Erhebungen durchgeführt.

Datenverändernd wirken dabei:

- Plausibilisierung
- Korrektur der Registerfehler (Karteileichen, Fehlbestände)

Qualitäts-steigerung

- Haushaltegenerierung (Verbindung der vorhandenen Informationen zu Personen und Wohnungen / Gebäuden zur Bildung zusätzlicher Angaben über Familien, HH)

Informations-anreicherung

- Anonymisierung des Datenbestandes (SAFE für Registerinformationen, Stichprobenanonymisierung)

Datenschutz

Im IAWS können auch nicht anonymisierte Daten abgerufen werden.

Klassifikation Geheimhaltungsverfahren

	Informationsreduzierende Verfahren	Datenverändernde Verfahren
Pre-tabulare Verfahren	- Vergröberung (Zusammenfassen von Kategorien) - Entfernen von Merkmalen	- Mikroaggregation, z.B. SAFE - Swapping - Stochastische Überlagerung auf Mikrodatenebene
Post-tabulare Verfahren	- Zellsperrung - Zusammenfassung (Tabellen Re-Design)	- Deterministische (konventionelle) Rundung - Zufällige Rundung - Kontrollierte Rundung - Stochastische Überlagerung auf Tabellenfeldebene
	↓	↓
	Löschen oder unterdrücken Informationen (auch unkritische Felder bei Sekundärsperrungen)	Schutz entsteht durch Unsicherheit (auch bei unkritischen Feldern)

Pre-tabulare datenverändernde Geheimhaltungsverfahren

Vorteile:

- Die Lösung der Geheimhaltung ist eine einmalige Aufgabe.
- Alle Auswertungen
  - erfolgen aus den anonymen Daten
  - sind untereinander immer konsistent
  - sind stets additiv
- Eine Sperrung von Tabellenfeldern (Primär- und Sekundärsperrungen) ist nicht nötig.

Nachteile:

- Die Interpretation von mit anonymisierten Mikrodaten erstellten Tabellen muss durch den Nutzer neu "gelernt" werden.
- Die Sperrungen in Auswertungen werden durch Unsicherheiten ersetzt.
- Der einmalige Rechenaufwand kann relativ hoch sein.

Das Verfahren SAFE

Grundidee des SAFE Verfahrens (1)

- Gruppen von mindestens 3 (möglichst ähnlichen) Merkmalsträgern bilden
- Ersetzen der Einzelangaben in der Gruppe von Merkmalsträgern mit einheitlichen Ausprägungen

Original

Person ID	Region	Alter
01	A	34
02	B	28
03	A	17
04	C	28
05	A	29
06	C	15
07	A	15
08	A	29
09	B	24
10	A	29
11	B	15
12	A	29

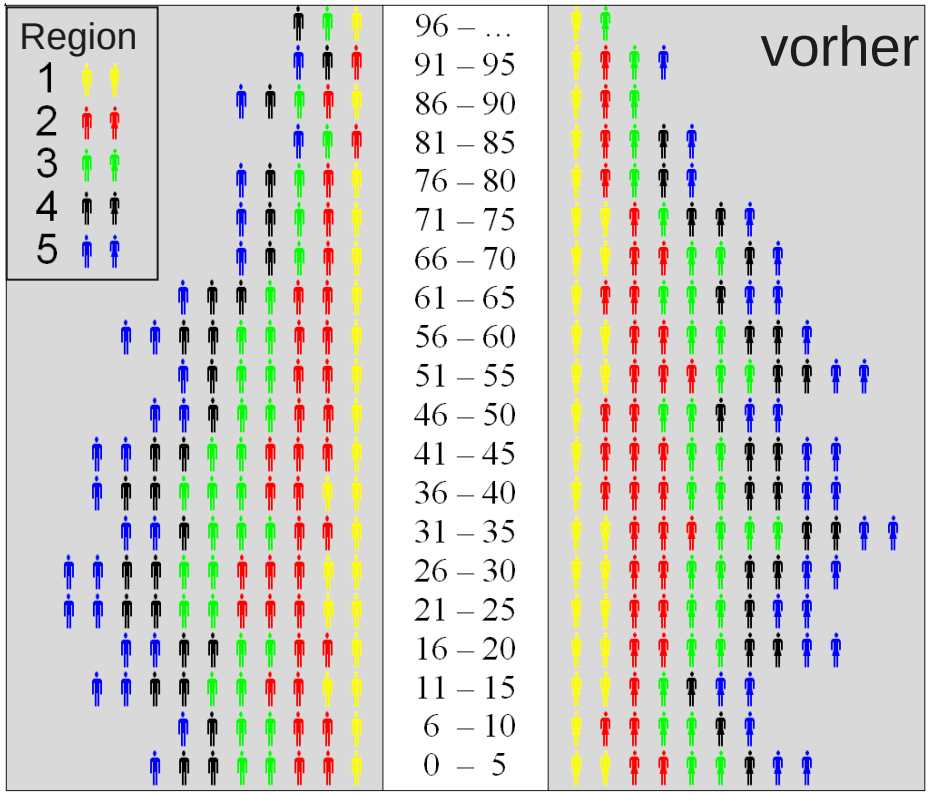
Saldo der Schlüssel	
Alter	
34	++
17	-
29	-
24	-
28	+
Region	
A	-
C	+

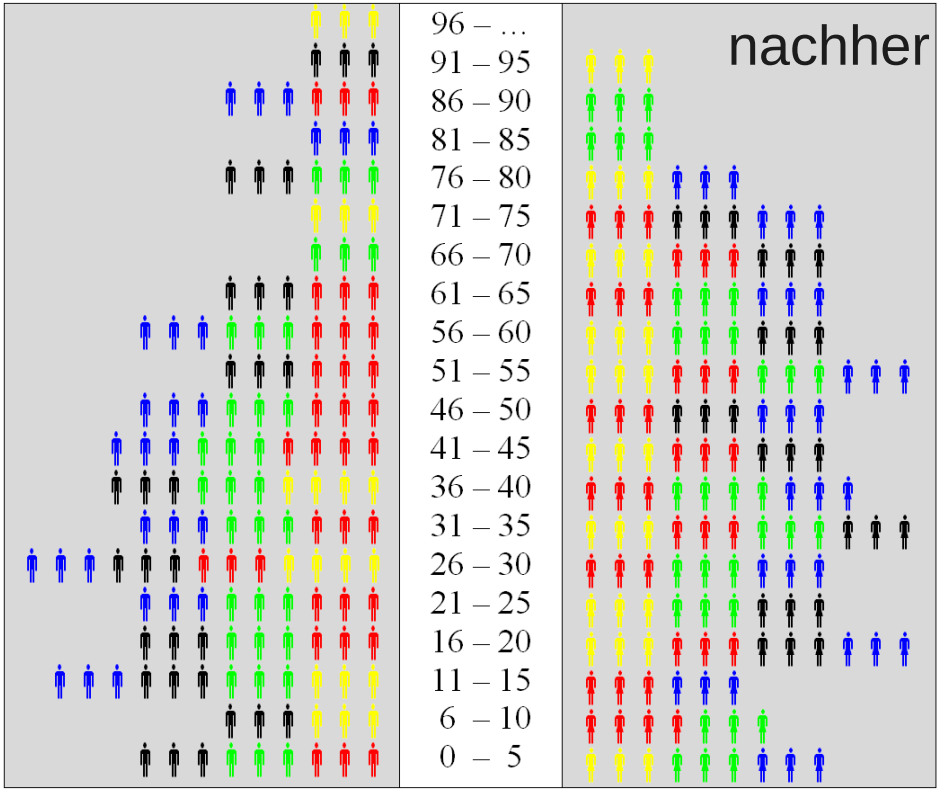
Anonym

Region	Alter
A	29
A	29
A	29
A	34
A	34
A	34
B	28
B	28
B	28
C	15
C	15
C	15

Saldo der Schlüssel: Merkmalswert ist im anonymen Bestand einmal häufiger (+) oder weniger (-) als im Original.

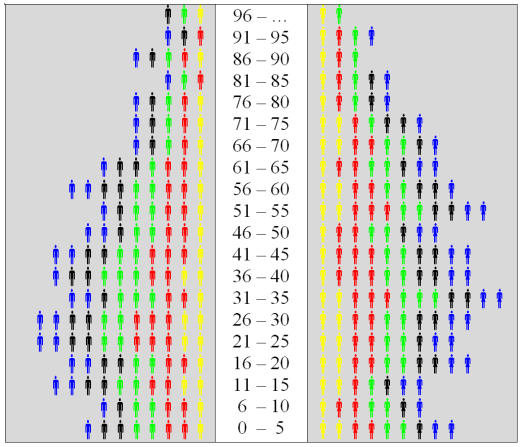
Grundidee des SAFE Verfahrens (2)



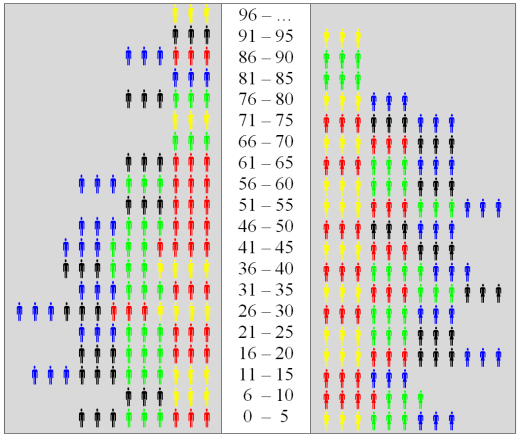


- Innerhalb des anonymisierten Datenbestandes sind alle statistischen Objekte mindestens dreimal mit identischen Merkmalsausprägungen vorhanden.
- Es gibt keine Geheimhaltungsprobleme mehr, da die Merkmalsausprägungen nicht mehr eindeutig einzelnen Objekten zugeordnet werden können.

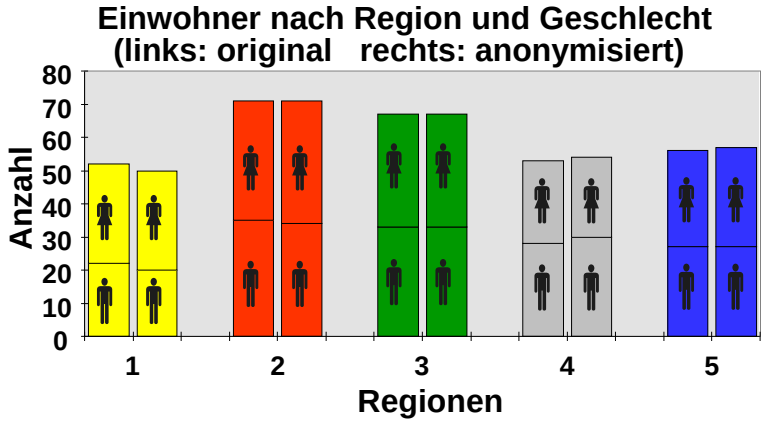
Grundidee des SAFE Verfahrens (3)

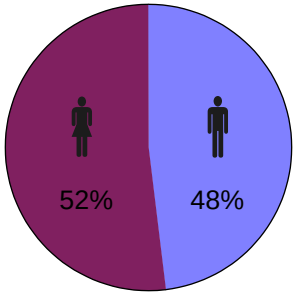


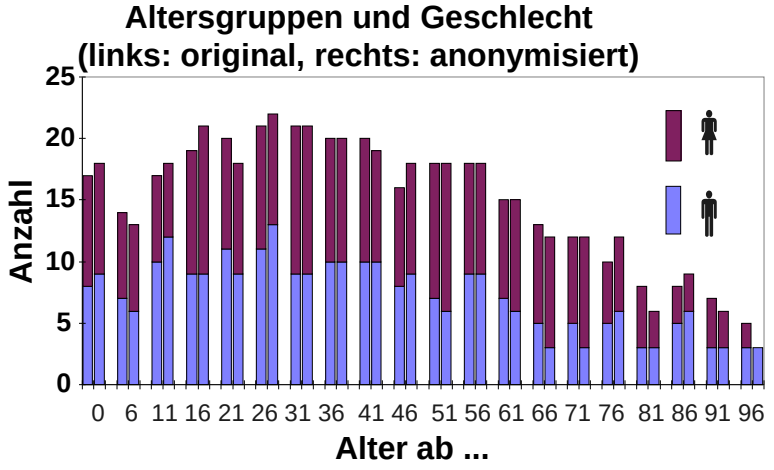
vorher



nachher







Besonderheiten von mit SAFE anonymisierten Daten

1. Veröffentlichte Tabellenfelder sind mit einem gewissen Fehler überlagert.

2. Der Fehler wird relativ um so größer, je kleiner die Zellwerte sind.

3. Die Größe des Fehlers ist bekannt und kann berücksichtigt werden.

4. In Tabellen fehlende Merkmalskombinationen können trotzdem existieren.

   In einer Reihe (Zeile/Spalte) einmalige Kombinationen (Randwertprobleme) sind nicht unbedingt in der originalen Gesamtheit einmalig.

5. Strukturen in den Daten werden gut abgebildet, und nicht wegen ggf. erforderlicher „Komplementärsperrungen“ verschleiert.

Mathematisches Modell SAFE

$$ZF: \quad \min_i \left(\max_i \left(|e_i| - g_i \right) \right)$$

$$Ax = t + e$$

$$\text{mit}: \quad x_j = 0,3,4,...$$

mit:

x - Vektor der Häufigkeiten möglicher Sätze im Datenbestand
- einmalige Originalsätze ($x_j^o=1$)
- völlig identische Originalsätze zusammengefasst ($x_j^o>1$)
- künstliche Sätze ($x_j^o=0$)

t - Vektor der Häufigkeiten aller Tabellenfelder in den zu kontrollierenden Tabellen bei Auswertung des Originaldatenbestandes

A - Zuordnungsmatrix
($a_{ij}=1$, wenn Objekt im Tabellenfeld j gezählt wird, sonst $a_{ij}=0$)

e - Vektor aller Fehler bei der Tabellierung; e_i ist Fehler im Tabellenfeld t_i

g - Vektor eventueller Gewichtungen der Tabellenfelder

Zulässige Abweichungen in Kontrolltabellen

- Durch Steuerungsparameter den Vektor g beeinflussen
- Bei eindimensionalen Auswertungen kleinere Abweichungen als bei mehrdimensionalen Auswertungen
- Bei größeren Feldern zusätzliche Abweichungen zulässig
  - Früher: 9 Größenklassen nach dem dekadischen Logarithmus
  - Neu für Zensus:
    <10, <20, <50, <100, <200, <1000, <10.000, <100.000, >100.000
    (zusätzliche Größenklassen im Bereich 10 und 1000 Einheiten)
  - Erzeugt „Stufenform“ bei Maximalabweichungen

SAFE–Tests mit der Volkszählung 1987

Mikrodatenfile für das Personenregister

Datensätze (Personen):	63 202 834		
kontrollierte Tabellen:	416		
Tabellenfelder:	11 485 249		
Davon mit Geheimhaltungsfällen:	2 134 034		
Maximale Abweichung:	10	insgesamt	(3 eindim.)

Tabellenfelder nach Größe von ... - bis ...	Anzahl an Tabellenfeldern	maximale Abweichung insgesamt	maximale Abweichung eindim.	mittlere Abweichung
1 - 9	4 471 429	6	2	1.65
10 - 49	2 822 413	7	2	2.39
50 - 99	1 008 565	8	2	2.67
100 - 149	505 857	9	2	2.95
150 - 199	318 245	9	2	3.19
200 - 999	1 254 910	10	2	3.41
1 000 - 9 999	854 690	10	2	3.41
10 000 - 99 999	215 132	10	3	3.28
100 000 - 999 999	31 177	10	3	3.13
1 000 000 und mehr	2 831	9	3	3.08

SAFE–Tests mit der Volkszählung 1987

Mikrodatenfile Personenregister

Abweichung im Tabellenfeld	Anzahl an Tabellenfeldern nach der Größe von … bis ...							
	1 - 9	10 - 49	50 - 99	100-149	150-199	200-999	1 000 - 9 999	10 000 und mehr
0	521 383	357 877	113 036	51 228	30 265	110 616	77 217	23 772
1	1 842 827	637 243	219 070	98 896	57 794	214 492	148 823	45 021
2	1 224 414	550 070	176 839	91 929	54 089	200 664	136 501	41 396
3	539 738	526 629	150 607	72 071	47 656	179 562	119 681	35 595
4	248 322	424 206	159 125	61 408	36 894	153 140	102 549	30 200
5	89 584	235 239	119 751	62 601	31 166	123 561	85 691	25 240
6	5 161	88 668	55 700	45 068	30 230	109 787	72 499	21 143
7	-	2 481	14 358	18 576	20 577	91 814	57 427	14 858
8	-	-	79	4 068	8 310	51 827	35 858	8 069
9	-	-	-	12	1 264	19 343	17 482	3 778
10	-	-	-	-	-	104	962	68
>10	-	-	-	-	-	-	-	-
Σ	4 471 429	2 822 413	1 008 565	505 857	318 245	1 254 910	854 690	249 140

Verschiedene Datenquellen – verschiedene Risiken?

- Register / Vollerhebungen
  - Umfang der enthaltenen Informationen ist begrenzt
  - Informationen sind aber für alle vorhanden (Teilnahmekenntnis)

  → Anonymisierung auf Ebene der Mikrodaten (Verfahren SAFE)

- Stichprobeninformationen
  - Umfang der erhobenen Informationen ist höher
  - (teilweise) fehlende Teilnahmekenntnis
  - höhere Qualität, da aktuelle Selbstauskunft

  → Anonymisierung bei der Auswertung, da Hochrechnungen erforderlich

- Amtliche Einwohnerzahl

  → ohne Geheimhaltung

Die Stichprobenanonymisierung

Grundidee der Stichprobenanonymisierung (1)

» Grundsätzlich:

Hochrechnungsfaktor = 1 / Auswahlsatz

» Aber: Hochrechnungsfaktoren werden skaliert

Hochrechnungsfaktor ≠ 1 / Auswahlsatz

(verschiedene Einheiten → verschiedene Hochrechnungsfaktoren)

» Bei fein untergliederten Tabellen und genauen Hochrechnungen lässt sich die originale Zusammensetzung aus den einzelnen Einheiten unter bestimmten Konstellationen reproduzieren.

Grundidee der Stichprobenanonymisierung (2)

Beispiel: Kleine Gemeinde ca. 200 EW, Auswahlsatz 10%
(20 Personen in der Stichprobe)

Tabelle 1:

	männlich	weiblich	ins-gesamt
Deutsche	83	94	177
Ausländer	9	11	20
Insgesamt	92	105	197

Tabelle 2:

	erwerbs-tätig	erwerbs-los	ins-gesamt
Deutsche	155	22	177
Ausländer	9	11	20
Insgesamt	164	33	197

Wer von den beiden Ausländern in der Stichprobe ist erwerbslos?
Stichprobenergebnisse sind nicht automatisch sicher!

Grundidee der Stichprobenanonymisierung (3)

- Hochrechnung allein ist nicht sicher
- Rundung auf Vielfache von 10
  - ist auch wegen der statistischen Unsicherheit sinnvoll
  - auch mit Rundung noch nicht ausreichend
- Kombination mit Mindestfallzahlregel

- Schritte bei Auswertung der Stichprobe
  - Hochrechnung
  - Sperrung bei Unterschreitung der Mindestpersonenzahl (s. nächste Folie)
  - Rundung der Ergebnisse (auf Vielfache von 10)

- Keine klassische Unterscheidung in den Tabellen zwischen nicht existent „-“ und geheim „/“ zu halten. Diese Unterscheidung ist als sichere Aussage für die Gesamtheit aus Stichproben nicht möglich.

- Stichprobentabellen enthalten nur statistisch belastbare Tabellenwerte.

Grundidee der Stichprobenanonymisierung (4)

» Vorgabe in Zensusgesetz: relativer Standardfehler maximal 15%
» Mindestpersonenzahl in der Stichprobe

$$n_g = \frac{(1-f)}{0{,}15^2}$$

wobei:
» n_g: Zahl der Personen in der Untergruppe g in der Stichprobe
» f: Auswahlwahrscheinlichkeit

» Für Tabellenfelder mit f=0,10 gilt dann

$$n_g = \frac{(1-0{,}1)}{0{,}15^2} \approx 40$$

» Mindestpersonenzahl ändert sich mit Auswahlsatz!

Vorgehen Geheimhaltung bei VÖT2

- Kleine Gemeinden (< 10.000 Einwohner)
  - Nur SAFE für Registermerkmale, HHG
  - Keine Hochrechnungsergebnisse

- Große Gemeinden (> 10.000 Einwohner)
  1. SAFE-Anonymisierung der Registermerkmale, HHG
  2. Hochrechnung der Stichprobe ggf. Addition der beiden Auswertungen
  3. Evtl. Sperrung bei Unterschreitung der Mindestfallzahl,

$$n_g = \frac{(1-f)}{0{,}15^{\,2}} a_g^{\,2}$$

modifizierte Formel
wobei a_g: Anteil, den der (hochgerechnete) Stichprobenteil zum Ergebnis des Tabellenfelds g beiträgt.

  4. Rundung

- Kreise, Regierungsbezirke, Bundesländer
  - Methodik wie große Gemeinden

Vorgehen Geheimhaltung

- vorgegebene Kontrolltabellen
  - Gemeindeblätter, Datenquader im IAWS und ÖAWS

- Je 1 „Gemeindeblatt“/ „Ergebnisblatt“ für Bevölkerung und für GWZ
  - Auswertungskanon für kleine Gemeinden <10.000 EW
  - Größerer Auswertungskanon für große Gemeinden >10.000 EW, Verbandsgemeinden, Kreise, Regierungsbezirk, Land, Bund
  - Bei GWZ gleicher Auswertungskanon für alle regionalen Einheiten

- GWZ → Vollerhebung: SAFE

- Bevölkerung (Register, Stichprobe oder Kombination)
  - Register: SAFE
  - Stichprobe: Hochrechnung, Sperrung und Rundung
  - Kombination: Baukasten (SAFE und Hochrechnung mit Sperrung und Rundung kombinieren)
  - Für jede Auswertung qualitativ hochwertigste Quelle verwenden

» Interpretation / Besonderheiten

Qualitätsangaben

» Qualitätsangaben für Tabellenfelder im ÖAWS verfügbar

Beispiel: SAFE–Tests mit der Volkszählung 1987
Gekennzeichnete Tabellenfelder im Mikrodatenfile Personenregister

Abweichung im Tabellenfeld	Anzahl an Tabellenfeldern nach der Größe von … bis ...							
	1 - 9	10 - 49	50 - 99	100 - 149	150 - 199	200 - 999	1 000 - 9 999	10 000 und mehr
0	521 383	357 877	113 036	51 228	30 265	110 616	77 217	23 772
1	1 842 827	637 243	219 070	98 896	57 794	214 492	148 823	45 021
2	1 224 414	550 070	176 839	91 929	54 089	200 664	136 501	41 396
3	539 738	526 629	150 607	72 071	47 656	179 562	119 681	35 595
4	248 322	424 206	159 125	61 408	36 894	153 140	102 549	30 200
5	89 584	235 239	119 751	62 601	31 166	123 561	85 691	25 240
6	5 161	88 668	55 700	45 068	30 230	109 787	72 499	21 143
7	-	2 481	14 358	18 576	20 577	91 814	57 427	14 858
8	-	-	79	4 068	8 310	51 827	35 858	8 069
9	-	-	-	12	1 264	19 343	17 482	3 778
10	-	-	-	-	-	104	962	68
>10	-	-	-	-	-	-	-	-
Σ	4 471 429	2 822 413	1 008 565	505 857	318 245	1 254 910	854 690	249 140

Vielen Dank für Ihre Aufmerksamkeit!

Haben Sie noch Fragen?

Kontakt:
Joerg.Hoehne@statistik-bbb.de