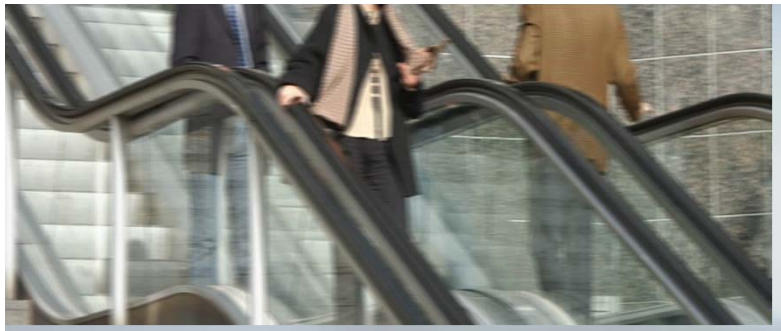

Wir bewegen

Informationen

STATISTIK AUSTRIA

Die Informationsmanager

KWSD

(6) Zensus 2011: wie kann der Datenzugang für die Wissenschaft sichergestellt werden?

Im Spannungsfeld von Wissenschaft und Datenschutz

Manuela Lenk

Bereichsleiterin Registerzählung

14. Jänner 2011

www.statistik.at

Zugang zu Mikrodaten

Bundesstatistikgesetz 2000

- Mikrodaten der amtlichen Statistik sind Daten, die im Zusammenhang mit statistischen Erhebungen der Statistik Austria anfallen
- Zugang nur zu nicht-personenbezogenen statistischen Daten → Anonymisierung notwendig
- Zugang nur für fachlich geeignete Personen und wissenschaftliche Einrichtungen (z.B. Universitäten, Dissertanten)
- Zugang nur für wissenschaftliche Projekte ohne kommerziellem Nutzen
- Ergebnisse sollen der Öffentlichkeit/Allgemeinheit zugänglich gemacht werden → Grundlagenforschung
- Durch Datensicherheitsmaßnahmen ist Vorsorge zu treffen, dass eine Ermittlung von personenbezogenen Daten mit Mitteln, die vernünftigerweise angewendet werden können, und eine Abspeicherung von personenbezogenen statistischen Daten auf externe Datenträger nicht möglich ist

Datenbereitstellung

- Volkszählung 2001
  - vordefinierte Abfragen via ISIS (Integriertes Statistisches Informationssystem)
    - kostenfreie Überblicksdaten, Detaildaten kostenpflichtig
  - ohne spezielle Datenschutzvorkehrungen
- Registerzählung 2011
  - Datenbank Superstar → interaktive Datenbank
  - Test mit Daten der Abgestimmten Erwerbsstatistik 2008 (AEST 2008) → Erwerbsstatistik im Rahmen der Registerzählung, auf Bundesland und Gemeindeebene
    - Anonymisierung noch nicht befriedigend gelöst
  - Webseite der Abgestimmten Erwerbsstatistik: http://www.statistik.at/web_de/frageboegen/registerzaehlung/abgestimmte_erwerbsstatistik/index.html

Anforderungen (1)

Publikationspflichten

- Interaktive Datenbank mit niedrigem Aggregationslevel (Gemeinde)
- Eurostat → Hypercubes
  - Bevölkerungsdaten und Wohnungssituation auf nationaler regionaler und örtlicher Ebene
  - 58 Kreuztabellen mit bis zu acht Merkmalen
  - anonymisiert
- Mikrodatenbereitstellung
  - Wissenschaft und Forschung
  - Landesstatistische Ämter

Anforderungen (2)

Datennutzbarkeit

- Additivität
  - Randwerte = Summe der Zellwerte
- Konsistenz
  - Zellwerte in unterschiedlichen Tabellen für gleiche Merkmalskombination ident
- universelle Methode
- Akzeptanz der Nutzer
  - Datensicherheit mit verständlichem Verfahren
- Sonderauswertungen
  - Auswertungen für detaillierte Fragestellungen
  - Zusätzliches Service für Datennutzer

Case Study (1)

Beurteilung der Effekte unterschiedlicher Anonymisierungsstrategien

- Record Swapping
  - **Idee:** Tausch (Umsetzen) von Personen bzw. Haushalten innerhalb eines bestimmten Gebiets, Mikrodatenverschmutzung
  - **Vorgangsweise**:
    - Ziehen einer bestimmten Anzahl (Swapping Rate) von Personen/Haushalten aus der Grundmasse
    - Suchen von passenden (z.B.: Übereinstimmungen bei BDL, Geschl.) Personen/Haushalten (Swapping Partner)
    - Abtausch der Merkmale dieser Einheiten
    - Für alle Einheiten durchführen bis die gewünschte Swapping Rate erreicht ist.

Vor- und Nachteile Record Swapping

Stärken	Schwächen
Anteil der geswappten Datensätze ist frei wählbar	Um höheren Datenschutz zu gewährleisten, sollte die *swap rate* nicht veröffentlicht werden.
Randverteilungen bleiben im Vergleich zu den Originaldaten auf hohem bis mittlerem Aggregationsniveau unverändert.	Spezielle Merkmalskombinationen bleiben mit großer Wahrscheinlichkeit unverändert.
Bei Verschmutzung auf Mikroebene wird immer der verschmutzte Datensatz verwendet → Konsistenz	Genauigkeitsabschätzung (wie zB.: Stichprobenfehler) nicht möglich
Wahl der Identifikationsvariablen kann einfach den Anforderungen angepasst werden.	
Erfahrung des ONS und US Census Bureau	

Case Study (2)

- ABS Perturbation
  - **Idee:** Perturbation der Output Tabellen durch ein nachvollziehbares Verfahren
  - **Vorgangsweise**:
    - Jeder Einheit im Mikrodatensatz wird ein Record-Key zugewiesen (flexible) Erstellung einer Look-Up Tabelle, die die Verschmutzungsverteilung enthält
    - Bei der Erzeugung einer Tabelle werden alle Record-Keys der Einheiten, die zu einer beliebigen Tabellenzelle beitragen, aggregiert und ein Cell-Key wird berechnet
    - In Abhängigkeit von Cell-Key und dem originalen Zellwert wird das Ausmaß der Perturbation in der Look-Up Tabelle nachgeschlagen
    - Dieser Wert (positiv, negativ oder 0) wird mit dem original Zellwert addiert und ausgegeben
    - Z.B. bei geringer Zellbesetzung, nur Ausgabe von 0 und >4

Vor- und Nachteile ABS Perturbationsmethode

Stärken	Schwächen
Methode ist sehr flexibel, da das Design der Look-Up Tabelle flexibel angepasst werden kann.	Alle Outputtabellen sind entweder additiv oder konsistent.
Das Verfahren bietet guten Schutz gegen Identifikation einzelner Einheiten	Perturbation ist für den Nutzer sichtbar (Nicht-Additivität bzw. Nicht-Konsistenz der Output-Tabellen).
Die Mikrodaten selbst werden nicht verschmutzt.	Problematische Mikrodatenweitergabe, da die Perturbationsmethode in die interaktive Datenbank implementiert ist.
Erfahrungen bei Publikation von Census Daten beim Australien Bureau of Statistics	Datenlieferung an Eurostat → inkonsistente Tabellen.

Case Study (3)

- Sampling
  - **Idee:** Zufallsauswahl von x% der Personen (bzw. Gebäuden/Haushalten) aus den Einzeldaten mit entsprechenden Hochrechnungsgewichten
  - **Vorgangsweise**:
    - Zufallsauswahl von Einheiten (Gebäuden, Haushalten, Personen) nach definiertem Stichprobenplan
    - ausgewählte Einheiten werden mit dazugehörigen Gewichten in die interaktive Datenbank eingelagert
  - **Bemerkungen:**
    - Einschränkungen der Methodik etwa hinsichtlich ganzzahliger Hochrechnungsgewichte oder dem Beharren auf gewissen Eckzahlen
    - komplexes Stichprobendesign möglich
    - Gewichtung einer Zufallsstichprobe führt zu nicht-Additivität der Outputtabellen durch Rundungsfehler

Vor- und Nachteile Sampling

Stärken	Schwächen
verbleibende Datensätze werden nicht verändert	Umgekehrter Zusammenhang zwischen Stichprobenfehler und dem Re-Identifikationsrisiko einzelner Einheiten
Haushaltsintere Konsistenz bleibt erhalten	(großer) Teil der Daten wird nicht verwendet
einfache Implementierung	Rundungsfehler bei nicht-ganzzahligen Gewichten
Einfach zu kommunizierendes Verfahren	sehr starke Veränderung der Auswertungen
Drittelstichprobe würde alle Zellen mit Häufigkeiten 1 oder 2 eliminieren	Keine Erfahrungen mit dem Verfahren in anderen NSIs
klassische Fehlerrechnung ist möglich	

Case Study (4)

- Calib 80
  - **Idee:** (Quasi)zufällige Auswahl von etwa 90% der Personen (bzw. Gebäuden) aus den Mikrodaten mit Hochrechnungsgewichten auf Personenebene
  - **Vorgangsweise**:
    - Zufälliges ziehen von etwa 80% der Personen (Gebäuden). Alle Personen (bzw. Personen in gezogenen Gebäuden) bekommen Gewicht=1 zugewiesen
    - Aus der restlichen Maße werden iterativ Personen (Gebäude) zufällig ausgewählt.
    - Allen Personen in den ausgewählten Einheiten (Gebäuden) wird das Gewicht 2 zugewiesen
    - Dies wird solange durchgeführt, bis ein Abbruchkriterium erreicht ist. (z.B. hochgerechnete Anzahl von In/Ausländern auf Gemeindeebene entspricht den Echtzahlen)
    - Falls der Algorithmus nicht konvergiert, wird mit einer neuen, zufälligen Auswahl in Schritt 1 begonnen.

Vor- und Nachteile Calib 80

Stärken	Schwächen
Verfahren liefert konsistente und additive Tabellen	Umgekehrter Zusammenhang zwischen Stichprobenfehler und Re-Identifikationsrisiko einzelner Einheiten
Haushaltsinterne Konsistenz bleibt erhalten	ein Teil der Daten wird nicht verwendet
Methode ist für Anwender einfach zu verstehen	möglicher (systematischer) Bias (etwa Überschätzung von 1-Personen Haushalten)
Eckzahlen (Anzahl der In-/Ausländer auf Gemeindeebene) bleiben unverändert	keine Erfahrungen mit dem Verfahren in anderen NSIs
Näherungsweise Fehlerrechnung möglich	

Ausblick / Zusammenfassung

- **alle Methoden** weisen spezifische Stärken und Schwächen auf
- **kein Verfahren liefert eine vollständige Lösung** für das divergierende Anforderungsprofil
- **Swapping** ist flexibel anpassbar, bietet allerdings oftmals wenig Schutz und ist Datennutzern möglicherweise schwer kommunizierbar
  - zusätzliche **Kalibration** auf Randsummen mit einem iterativen Algorithmus **grundsätzlich möglich**, wenn ganzzahle Gewichte gewünscht sind
- die **ABS-Methode** ist flexibel anpassbar und bietet guten Schutz. Tabellen sind jedoch entweder konsistent (Zellen haben in allen Tabellen den gleichen Wert) oder additiv → bei Mikrodatenweitergabe keine Konsistenz

Vielen Dank für die Aufmerksamkeit !