Johannes Gussenbauer
Qualitätsmanagement und
Methodik (QM)

Wien
February 2020

Geheimhaltung

Statistische Geheimhaltung in der Statistik Austria

Johannes Gussenbauer
Qualitätsmanagement und
Methodik (QM)

Wien
February 2020

Geheimhaltung

Gründe für statistische Geheimhaltung

Statistische Geheimhaltung bedeutet, Daten so abzusichern, dass nach der Veröffentlichung nicht mehr auf vertrauliche Information von Einzelangaben rückgeschlossen werden kann.

Statistische Geheimhaltung bedeutet, Daten so abzusichern, dass nach der Veröffentlichung nicht mehr auf vertrauliche Information von Einzelangaben rückgeschlossen werden kann.

- statistische Geheimhaltung ist relevant für:
  - Tabellen
  - (interaktive) Datenbanken
  - Mikrodaten

Definition und Grundlagen

Statistische Geheimhaltung bedeutet, Daten so abzusichern, dass nach der Veröffentlichung nicht mehr auf vertrauliche Information von Einzelangaben rückgeschlossen werden kann.

- statistische Geheimhaltung ist relevant für:
  - Tabellen
  - (interaktive) Datenbanken
  - Mikrodaten

- in Ö: zentrale Grundlage ist das Bundesstatistikgesetz 2000:

  Statistiken sind grundsätzlich in solcher Weise zu veröffentlichen, dass ein Rückschluss auf Angaben über bestimmte oder bestimmbare Betroffene ausgeschlossen werden kann

Fragestellungen

Fragestellungen

- steigende Nachfrage nach Mikrodaten

Fragestellungen

- steigende Nachfrage nach Mikrodaten

- ist Zugriff auf Mikrodaten durch Forscher ‘von außen’ möglich

Fragestellungen

- steigende Nachfrage nach Mikrodaten
- ist Zugriff auf Mikrodaten durch Forscher ‘von außen’ möglich
- Erstellung anonymisierter Public-Use Files für Nutzer, die unterschiedliche Interessen haben

Fragestellungen

- steigende Nachfrage nach Mikrodaten
- ist Zugriff auf Mikrodaten durch Forscher ‘von außen’ möglich
- Erstellung anonymisierter Public-Use Files für Nutzer, die unterschiedliche Interessen haben
- Geheimhaltung von ‘dynamischen’ Tabellen (Superstar)

Fragestellungen

- steigende Nachfrage nach Mikrodaten
- ist Zugriff auf Mikrodaten durch Forscher ‘von außen’ möglich
- Erstellung anonymisierter Public-Use Files für Nutzer, die unterschiedliche Interessen haben
- Geheimhaltung von ‘dynamischen’ Tabellen (Superstar)
- wie kann die Anwendung von Geheimhaltungsmaßnahmen für Datennutzer gut dokumentiert werden?

Arten von Disclosure

als Disclosure versteht man, wenn aus veröffentlichten Daten Information über eine einzelne, spezifische statistische Einheit abgeleitet ('gelernt') werden kann.

Arten von Disclosure

- Identity-Disclosure: gegeben seien Daten x:

Wohnort	**Geschlecht**	**Beruf**
Vorau	männlich	Univ.-Prof.

Arten von Disclosure

- Identity-Disclosure: gegeben seien Daten x:

Wohnort	Geschlecht	Beruf
Vorau	männlich	Univ.-Prof.

- Attribut-Disclosure: Tabelle 1: (Fußballinteresse nach Geschlecht)

	männlich	weiblich	Gesamt
Fußballfan	12	0	22
kein Fußballfan	93	85	168
Gesamt	105	85	190

Einteilung von Methoden

Einteilung von Methoden

- Datenreduzierende Verfahren
  - Unterdrückung
  - Umkodieren

Einteilung von Methoden

- Datenreduzierende Verfahren
  - Unterdrückung
  - Umkodieren
- Datenmodifizierende Verfahren
  - Runden
  - Mikroaggregation
  - Postrandomisierung
  - Überlagern mit ‘Noise’

Einteilung von Methoden

- Datenreduzierende Verfahren
  - Unterdrückung
  - Umkodieren
- Datenmodifizierende Verfahren
  - Runden
  - Mikroaggregation
  - Postrandomisierung
  - Überlagern mit ‘Noise’
- (synthetische) Datengenerierungsverfahren
  - anstelle der echten Daten, werden synthetische Daten mit möglichst gleichen statistischen Eigenschaften generiert und publiziert

Spannungsfeld Geheimhaltung

Spannungsfeld Geheimhaltung

- ständiger Tradeoff zwischen Datenschutz und Datenqualität

Spannungsfeld Geheimhaltung

- ständiger Tradeoff zwischen Datenschutz und Datenqualität
- Respondentenschutz
  - Vertrauen der Respondenten in die (amtliche) Statistik erhalten

Spannungsfeld Geheimhaltung

- ständiger Tradeoff zwischen Datenschutz und Datenqualität
- Respondentenschutz
  - Vertrauen der Respondenten in die (amtliche) Statistik erhalten
- Wünsche von Forschern und Datennutzern

Spannungsfeld Geheimhaltung

- ständiger Tradeoff zwischen Datenschutz und Datenqualität
- Respondentenschutz
  - Vertrauen der Respondenten in die (amtliche) Statistik erhalten
- Wünsche von Forschern und Datennutzern
- Statistik Austria agiert nach eigenen, veröffentlichten Richtlinien

Johannes Gussenbauer
Qualitätsmanagement und
Methodik (QM)

Wien
February 2020

Geheimhaltung

Geheimhaltung von Mikrodaten

- **direkte Identifizierungsvariablen**: z.B. SVNr, Name, usw.

- **direkte Identifizierungsvariablen**: z.B. SVNr, Name, usw.
- **indirekte Identifizierungsvariablen**: jegliche kategorielle Variable ausser direkte Identifizierungsvariablen.

- **direkte Identifizierungsvariablen**: z.B. SVNr, Name, usw.
- **indirekte Identifizierungsvariablen**: jegliche kategorielle Variable ausser direkte Identifizierungsvariablen.
- **Schlüsselvariablen**: jene indirekte Identifizierungsvariablen für welche anzunehmen ist, dass Datenangreifer Informationen besitzen.

- **direkte Identifizierungsvariablen**: z.B. SVNr, Name, usw.
- **indirekte Identifizierungsvariablen**: jegliche kategorielle Variable ausser direkte Identifizierungsvariablen.
- **Schlüsselvariablen**: jene indirekte Identifizierungsvariablen für welche anzunehmen ist, dass Datenangreifer Informationen besitzen.
- **typische Schlüsselvariablen**: Nationalität, Beruf, NACE, Forschungsausgaben von Unternehmen, ...

- **direkte Identifizierungsvariablen**: z.B. SVNr, Name, usw.
- **indirekte Identifizierungsvariablen**: jegliche kategorielle Variable ausser direkte Identifizierungsvariablen.
- **Schlüsselvariablen**: jene indirekte Identifizierungsvariablen für welche anzunehmen ist, dass Datenangreifer Informationen besitzen.
- **typische Schlüsselvariablen**: Nationalität, Beruf, NACE, Forschungsausgaben von Unternehmen, ...

$\longrightarrow$ **Wichtig:** durch Verkreuzung von Schlüsselvariablen entsteht das Geheimhaltungsproblem

Geheimhaltung von Mikrodaten / k-Anonymität

- Merkmalsets von Schlüsselvariablen werden als **Keys** (f_k) bezeichnet

- Merkmalsets von Schlüsselvariablen werden als **Keys** (f_k) bezeichnet

- **Uniqueness**: gilt für ein Individiuum $f_k = 1$, so ist dieses eindeutig im Datensatz identifiziert
- k-**Anonymität**: jeder Kombination (jedem Key) können zumindest k Beobachtungen zugeordnet werden (3-Anonymität $\rightarrow f_k \geq 3$)

- Merkmalsets von Schlüsselvariablen werden als **Keys** (f_k) bezeichnet
- **Uniqueness**: gilt für ein Individiuum $f_k = 1$, so ist dieses eindeutig im Datensatz identifiziert
- k-**Anonymität**: jeder Kombination (jedem Key) können zumindest k Beobachtungen zugeordnet werden (3-Anonymität $\rightarrow f_k \geq 3$)
- Erreichen von k-Anonymität durch
  - Löschen einzelner Werte
  - Umkodieren und vergröbern einzelner Variablen

Geheimhaltung von Mikrodaten / Risiko

- Unterscheidung zwischen individuellen- und globalen Risikomaßen
- Risikomaße hängen ab von der Verteilung der Keys in
  - der Stichprobe $\longrightarrow f_k$
  - der Grundgesamtheit $\longrightarrow F_k$

- Unterscheidung zwischen individuellen- und globalen Risikomaßen
- Risikomaße hängen ab von der Verteilung der Keys in
  - der Stichprobe $\longrightarrow f_k$
  - der Grundgesamtheit $\longrightarrow F_k$

- F_k nicht beobachtbar, muss durch $\hat{F}_k$ geschätzt werden
- langes Formelwerk mit einigen Verteilungsannahmen, aber es gilt:

$$F_k | f_k \sim negBIN(\hat{p}_k, f_k)$$

$\rightarrow \hat{p}_k$ hängt dabei von Stichprobengewichten w_i und f_k ab!

Geheimhaltung von Mikrodaten / Umkodieren

- für kategorische Variablen: Vereinigung mehrerer Kategorien in eine neue oder bereits bestehende Kategorie

Geheimhaltung von Mikrodaten / Umkodieren

- für kategorische Variablen: Vereinigung mehrerer Kategorien in eine neue oder bereits bestehende Kategorie

- für stetige Variablen: Diskretisieren, Umkodieren in Kategorien (z.B. Einkommensklassen)

- für kategorische Variablen: Vereinigung mehrerer Kategorien in eine neue oder bereits bestehende Kategorie
- für stetige Variablen: Diskretisieren, Umkodieren in Kategorien (z.B. Einkommensklassen)
- Spezialfall: Top-/Bottom-Coding
  - $\longrightarrow$ kleinste und größte Werte oftmals selten und werden daher in einer 'top' btw. 'bottom' Kategorie zusammengefügt (z.B extrem hohe Einkommen)
- typischerweise wird das Umkodieren einer Variable für alle Units im Datensatz angewendet

- Idee: suche m ähnliche Beobachtungen, aggregiere diese und ersetzte die Werte mit dieser Aggregation

- Idee: suche m ähnliche Beobachtungen, aggregiere diese und ersetzte die Werte mit dieser Aggregation

- viele Möglichkeiten um 'ähnliche' Beobachtungen zu finden (mit und ohne Clustering, Projektionen, unterschiedliche Distanzmaße)
- unterschiedliche Möglichkeiten die Werte zu aggregieren (typisch ist das arithmetische Mittel)

- Idee: suche m ähnliche Beobachtungen, aggregiere diese und ersetzte die Werte mit dieser Aggregation

- viele Möglichkeiten um 'ähnliche' Beobachtungen zu finden (mit und ohne Clustering, Projektionen, unterschiedliche Distanzmaße)
- unterschiedliche Möglichkeiten die Werte zu aggregieren (typisch ist das arithmetische Mittel)

- Wahl von m? $\longrightarrow$ oft wird $m = 3$ gesetzt
- Wichtig: geeignete (robuste) Verfahren bei Ausreißern verwenden

Geheimhaltung von Mikrodaten / Noise

- Überlagern von Variablen mit Zufallsvariablen
- oftmals normalverteilt mit $\mu = 0$ und fixer Varianz σ^2

Geheimhaltung von Mikrodaten / Noise

- Überlagern von Variablen mit Zufallsvariablen
- oftmals normalverteilt mit $\mu = 0$ und fixer Varianz σ^2

- kleine Werte werden verhältnismäßig stark verändert
- Ausreißer werden oft nicht genug modifiziert

Geheimhaltung von Mikrodaten / Noise

- Überlagern von Variablen mit Zufallsvariablen
- oftmals normalverteilt mit $\mu = 0$ und fixer Varianz σ^2

- kleine Werte werden verhältnismäßig stark verändert
- Ausreißer werden oft nicht genug modifiziert

- korrelierter Noise: es wird die geschätzte Varianz/Kovarianzmatrix der Originaldaten bei der Erzeugung des Noise-Terms berücksichtigt
- wesentliche Statistiken (z.B Varianzen, Korrelationen) können (asymptotisch) bewahrt werden

Geheimhaltung von Mikrodaten / Noise

- Überlagern von Variablen mit Zufallsvariablen
- oftmals normalverteilt mit $\mu = 0$ und fixer Varianz σ^2

- kleine Werte werden verhältnismäßig stark verändert
- Ausreißer werden oft nicht genug modifiziert

- korrelierter Noise: es wird die geschätzte Varianz/Kovarianzmatrix der Originaldaten bei der Erzeugung des Noise-Terms berücksichtigt
- wesentliche Statistiken (z.B Varianzen, Korrelationen) können (asymptotisch) bewahrt werden

- Noise kann auch nur für selektierte Units hinzugefügt werden

- **Data swapping**: ein gewisser %-Satz der Beobachtungen wird modifiziert, indem die Werte für einige Variablen zwischen zwei Units getauscht (geswappt) werden

- **Data swapping**: ein gewisser %-Satz der Beobachtungen wird modifiziert, indem die Werte für einige Variablen zwischen zwei Units getauscht (geswappt) werden

- **Rank swapping**: zuerst werden die Werte einer Variable sortiert. Jeder Wert wird dann (zufällig) mit einem anderen Wert getauscht, der in einer gewissen Range liegen muss

- **Data swapping**: ein gewisser %-Satz der Beobachtungen wird modifiziert, indem die Werte für einige Variablen zwischen zwei Units getauscht (geswappt) werden

- **Rank swapping**: zuerst werden die Werte einer Variable sortiert. Jeder Wert wird dann (zufällig) mit einem anderen Wert getauscht, der in einer gewissen Range liegen muss

- **PRAM** ist ein Spezialfall von Data-Swapping (randomisiertes Swapping) basierend auf einer Übergangsmatrix

Geheimhaltung von Mikrodaten / Utility

- **Data-Utility**: Beurteilung (vorallem von stetig skalierten Variablen) basiert auf (robusten) Distanzmaßen

- **Data-Utility**: Beurteilung (vorallem von stetig skalierten Variablen) basiert auf (robusten) Distanzmaßen

- **Benchmarking**: Vergleich von Indikatoren auf Original- und anonymisiertem Datensatz und Evaluierung der Qualität anhand der Unterschiede

- **Data-Utility**: Beurteilung (vorallem von stetig skalierten Variablen) basiert auf (robusten) Distanzmaßen

- **Benchmarking**: Vergleich von Indikatoren auf Original- und anonymisiertem Datensatz und Evaluierung der Qualität anhand der Unterschiede

- **Modellrechnungen**: Berechnen von (Regressions)Modellen auf Original- und anonymisierten Daten und Vergleich der Ergebnisse

- **Data-Utility**: Beurteilung (vorallem von stetig skalierten Variablen) basiert auf (robusten) Distanzmaßen
- **Benchmarking**: Vergleich von Indikatoren auf Original- und anonymisiertem Datensatz und Evaluierung der Qualität anhand der Unterschiede
- **Modellrechnungen**: Berechnen von (Regressions)Modellen auf Original- und anonymisierten Daten und Vergleich der Ergebnisse
- **Einfluß** von Anonymisierungsmaßnahmen auf Modellergebnisse (Parameter) vergleichen

- **Remote Execution (kontrolliertes Fernrechnen)**:
  - Forscher sieht keine Originaldatenwerte. Ev. Zugriff auf künstliche Daten. Anwendung von Code auf Originaldaten und (wiederholter!) Check der Ergebnisse notwendig.

- **Remote Execution (kontrolliertes Fernrechnen)**:
  - Forscher sieht keine Originaldatenwerte. Ev. Zugriff auf künstliche Daten. Anwendung von Code auf Originaldaten und (wiederholter!) Check der Ergebnisse notwendig.

- **Lab (Safe Center)**:
  - Forscher bekommt Arbeitsplatz und Vertrag in der Statistik Austria. Kann nur auf speziell eingerichteten PCs auf Originaldaten zugreifen. Outputkontrolle notwendig!

- **Remote Execution (kontrolliertes Fernrechnen)**:
  - Forscher sieht keine Originaldatenwerte. Ev. Zugriff auf künstliche Daten. Anwendung von Code auf Originaldaten und (wiederholter!) Check der Ergebnisse notwendig.
- **Lab (Safe Center)**:
  - Forscher bekommt Arbeitsplatz und Vertrag in der Statistik Austria. Kann nur auf speziell eingerichteten PCs auf Originaldaten zugreifen. Outputkontrolle notwendig!
- **Remote Access**:
  - Forscher hat Fernzugriff auf Daten und kann mit Daten arbeiten. Generierter Output wird kontrolliert und an den Forscher gesendet.
  - resourcenschonend da nur finale Geheimhaltung notwendig und optimal für Forscher (Sehen von Echtdaten)
  - rechtlich in Österreich nicht möglich.

Johannes Gussenbauer
Qualitätsmanagement und
Methodik (QM)

Wien
February 2020

Geheimhaltung

Geheimhaltung von Tabellen

- **Grundlage** für alle Tabellen sind Mikrodaten
- **Unterscheidung** von Wertetabellen und Häufigkeitstabellen

- **Grundlage** für alle Tabellen sind Mikrodaten
- **Unterscheidung** von Wertetabellen und Häufigkeitstabellen

- **Häufigkeitstabelle**: Anzahl der beitragenden Einheiten für jede Zelle der Tabelle wird ausgewiesen
- **Wertetabelle**: für eine erhobene Variable wird die Summe dieser Variable über alle beitragenden Einheiten in jeder Zelle der Tabelle ausgewiesen

- **Grundlage** für alle Tabellen sind Mikrodaten
- **Unterscheidung** von Wertetabellen und Häufigkeitstabellen

- **Häufigkeitstabelle**: Anzahl der beitragenden Einheiten für jede Zelle der Tabelle wird ausgewiesen
- **Wertetabelle**: für eine erhobene Variable wird die Summe dieser Variable über alle beitragenden Einheiten in jeder Zelle der Tabelle ausgewiesen

- **Wichtig**: lineare Abhängigkeiten zwischen Tabellenzellen (Zeilen-/Spaltensummen im 2-dimensionalen Fall)
- statistische Tabellen können ein- oder mehrdimensional, hierarchisch und/oder verlinkt sein.

Geheimhaltung von Tabellen / Formalisierung

- was ist eigentlich eine Tabelle?
- eine allgemeine statistische Tabelle ist gegeben durch:
  - einen Datenvektor: $a = [a_1, ..., a_n]$
  - lineare Einschränkungen der Form: $M \cdot a = b$

- Bemerkungen:
  - M ist eine Matrix mit $M_{ij} \in \{-1, 0, 1\}$
  - b ist ein Vektor mit allen $b_i = 0$
  - jede Zeile des Gleichungssystems $M \cdot a = b$ entspricht hier der Einschränkung einer Zeilen-/ oder Spaltensumme.
  - die Zellen sind durch ihren (Spalten)Index: $j = 1, ..., n$ festgelegt

- zur Beurteilung, ob eine Tabellenzelle als ‘unsicher’ (und daher schützenswert) gelten soll, kann eine der folgenden Regeln herangezogen werden:
  - **Fallzahlregel**: die Anzahl der zu einer Zelle beitragenden Einheiten ist $<$ einem festgesetzten Wert (oftmals 3 oder 4)

- zur Beurteilung, ob eine Tabellenzelle als ‘unsicher’ (und daher schützenswert) gelten soll, kann eine der folgenden Regeln herangezogen werden:
  - **Fallzahlregel**: die Anzahl der zu einer Zelle beitragenden Einheiten ist $<$ einem festgesetzten Wert (oftmals 3 oder 4)
  - **(n,k)-Dominanzregel**: eine Zelle muss geschützt werden, wenn der Gesamtwert der n größten Beitragenden k% des gesamten Zellwertes überschreitet

- zur Beurteilung, ob eine Tabellenzelle als 'unsicher' (und daher schützenswert) gelten soll, kann eine der folgenden Regeln herangezogen werden:
  - **Fallzahlregel**: die Anzahl der zu einer Zelle beitragenden Einheiten ist $<$ einem festgesetzten Wert (oftmals 3 oder 4)
  - **(n,k)-Dominanzregel**: eine Zelle muss geschützt werden, wenn der Gesamtwert der n größten Beitragenden k% des gesamten Zellwertes überschreitet
  - **p-% Regel**: der Totalwert minus der 2 größten Beitragenden ist geringer als p% des größten Beitrages (typischerweise liegt p zwischen 5 und 15)

- Primärsperrung alleine oft nicht ausreichend (z.B Rückrechnung wg. linearer Abhängigkeiten)

- Primärsperrung alleine oft nicht ausreichend (z.B Rückrechnung wg. linearer Abhängigkeiten)

- es gibt verschiedene Möglichkeiten um primär unsichere Tabellenzellen zu schützen, z.B:
  - Zellsperrung/-unterdrückung
  - Runden
  - Zellanpassung

- Primärsperrung alleine oft nicht ausreichend (z.B Rückrechnung wg. linearer Abhängigkeiten)
- es gibt verschiedene Möglichkeiten um primär unsichere Tabellenzellen zu schützen, z.B:
  - Zellsperrung/-unterdrückung
  - Runden
  - Zellanpassung
- Zellsperrung ist die (in der Statistik Austria) am häufigsten verwendete Methode

Sicherung sensitiver Zellen (Beispiel)

W	A	B	C	Total
x	20	50	10	**80**
y	8	19	22	**49**
z	17	32	12	**61**
Total	**45**	**101**	**44**	**190**

Sicherung sensitiver Zellen (Beispiel)

W	A	B	C	Total
x	20	50	10	**80**
y	8	19	22	**49**
z	17	32	12	**61**
Total	**45**	**101**	**44**	**190**

- Sei Zelle y/C ($PS = \{7\}$) sensibel und muss unterdrückt werden.

Sicherung sensitiver Zellen (Beispiel)

W	A	B	C	Total
x	20	50	10	**80**
y	8	19	NA	**49**
z	17	32	12	**61**
Total	**45**	**101**	**44**	**190**

Sicherung sensitiver Zellen (Beispiel)

W	A	B	C	Total
x	20	50	10	**80**
y	8	19	NA	**49**
z	17	32	12	**61**
Total	**45**	**101**	**44**	**190**

- Wegen linearer Zusammenh"ange ist es nicht ausreichend, nur geheimzuhaltende Zellen alleine zu unterdr"ucken (Prim"arsperrung).

Sicherung sensitiver Zellen (Beispiel)

- alternative Sperrmuster

Sicherung sensitiver Zellen (Beispiel)

- alternative Sperrmuster

W	A	B	C	Total
x	20	50	10	**80**
y	S	19	NA	**49**
z	S	32	S	**61**
Total	**45**	**101**	**44**	**190**

Sicherung sensitiver Zellen (Beispiel)

- alternative Sperrmuster

W	A	B	C	Total
x	20	50	10	**80**
y	S	19	NA	**49**
z	S	32	S	**61**
Total	**45**	**101**	**44**	**190**

W	A	B	C	Total
x	S	50	S	**80**
y	S	19	NA	**49**
z	17	32	12	**61**
Total	**45**	**101**	**44**	**190**

Geheimhaltung von Tabellen / Zellunterdrückung

Geheimhaltung von Tabellen / Zellunterdrückung

- Unterdrückung der primär gesperrten Werte nicht ausreichend
- was ist ein ausreichendes Sperrmuster?

Geheimhaltung von Tabellen / Zellunterdrückung

- Unterdrückung der primär gesperrten Werte nicht ausreichend
- was ist ein ausreichendes Sperrmuster?

- primär unsichere Zellen dürfen nicht innerhalb eines gegebenen Intervals berechenbar sein
- gibt es optimale Sperrmuster und wenn ja, was charakterisiert ein optimales Muster?

Geheimhaltung von Tabellen / Zellunterdrückung

- Unterdrückung der primär gesperrten Werte nicht ausreichend
- was ist ein ausreichendes Sperrmuster?

- primär unsichere Zellen dürfen nicht innerhalb eines gegebenen Intervals berechenbar sein
- gibt es optimale Sperrmuster und wenn ja, was charakterisiert ein optimales Muster?

- möglichst wenige zusätzliche Sperrungen
- möglichst geringe Wertesumme bei zusätzlichen Sperrungen

Geheimhaltung von Tabellen / Zellunterdrückung

- Unterdrückung der primär gesperrten Werte nicht ausreichend
- was ist ein ausreichendes Sperrmuster?

- primär unsichere Zellen dürfen nicht innerhalb eines gegebenen Intervals berechenbar sein
- gibt es optimale Sperrmuster und wenn ja, was charakterisiert ein optimales Muster?

- möglichst wenige zusätzliche Sperrungen
- möglichst geringe Wertesumme bei zusätzlichen Sperrungen

- Problem der sekundären Unterdrückung ist komplex
- Lösungsalgorithmen basieren auf linearer Optimierung
- Zellunterdrückung ist in Wahrheit eine Form von Intervallpublikation.

Geheimhaltung von Tabellen / Runden

- normales Runden:
  - Runden des Zellwertes zum nächsten Vielfachen der Basis
  - bietet kaum Schutz

Geheimhaltung von Tabellen / Runden

- normales Runden:
  - Runden des Zellwertes zum nächsten Vielfachen der Basis
  - bietet kaum Schutz

- zufälliges Runden:
  - Zellwerte werden unabhängig voneinander zufällig auf- oder abgerundet
  - Vielfache der Basis werden nicht verändert
  - Wahl unterschiedlicher Gewichtungsschemata möglich
  - Verlust der Additivitätseigenschaft

Geheimhaltung von Tabellen / Runden

- normales Runden:
  - Runden des Zellwertes zum nächsten Vielfachen der Basis
  - bietet kaum Schutz
- zufälliges Runden:
  - Zellwerte werden unabhängig voneinander zufällig auf- oder abgerundet
  - Vielfache der Basis werden nicht verändert
  - Wahl unterschiedlicher Gewichtungsschemata möglich
  - Verlust der Additivitätseigenschaft
- kontrolliertes Runden:
  - Additivität der Tabelle soll nach dem Runden gewahrt bleiben
  - Vielfache der Basis werden (grundsätzlich) nicht verändert
  - ein (komplexes) lineares Problem (nicht notwendigerweise lösbar)

- Idee der Zellanpassung:
  - jeder primär gesperrte Zellwert wird durch einen ‘sicheren’ Wert am oberen oder unteren Rand eines fixen Sicherheitsintervals ersetzt.
  - andere Zellen werden so adjustiert, dass eine neue, additive Tabelle entsteht.

- Idee der Zellanpassung:
  - jeder primär gesperrte Zellwert wird durch einen ‘sicheren’ Wert am oberen oder unteren Rand eines fixen Sicherheitsintervals ersetzt.
  - andere Zellen werden so adjustiert, dass eine neue, additive Tabelle entsteht.

- Ergebnis: vollständige Tabellen (ohne Lücken)
  - meist geringe Anpassungen notwendig
  - optimale Algorithmen nur brauchbar für sehr kleine Tabellen
  - Heuristiken existieren, garantieren aber keine Lösung

Johannes Gussenbauer
Qualitätsmanagement und Methodik (QM)

Wien
February 2020

Geheimhaltung

Software

- `sdcMicro` (im Haus entwickeltes, freies `R`-Paket):
  - S4 basierend, rechenintensive Methoden in C/C++ implementiert
  - zusätzliche Methoden (z. B robuste Mikroaggregation, robuste Risikomaße) vorhanden
  - Anwendung der Methoden mit CLI und GUI (Paket {`sdcMicroGUI`})
  - `automatisches Berechnen von Häufigkeiten und Risikomaßen`
  - `Reproduzierbarkeit (Skript, Report)`
  - `Link:` https://github.com/sdcTools/sdcMicro

- `sdcTable` (im Haus entwickeltes, freies `R`-Paket):
  - automatische Modellierung beliebig komplexer, hierarchischer Tabellenstrukturen
  - unterschiedliche Methoden zur Identifizierung primär unsicherer Zellen
  - optimale und heuristische Unterdrückungsalgorithmen implementiert
  - Wahl unterschiedlicher LP-Solver möglich
  - derzeit keine grafische Benutzeroberfläche vorhanden
  - strikte S4-Klassenprogrammierung, flexible Anpassungen möglich
  - Link `https://github.com/sdcTools/sdcTable`

Rückfragen bitte an:
Johannes Gussenbauer

Kontakt:
Guglgasse 13, 1110 Wien
Tel: +43 (1) 71128–7934
Gregor.deCillia@statistik.gv.at

Geheimhaltung

Statistische Geheimhaltung in der Statistik Austria