Advanced Survey Statistics: Disclosure Control

Part 2: Basics

Matthias Templ

Institut für Datenanalyse und Prozessdesign
School of Engineering
Zürcher Hochschule für Angewandte Wissenschaften

FU-Berlin, 2019

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering
IDP Institut für Datenanalyse und Prozessdesign

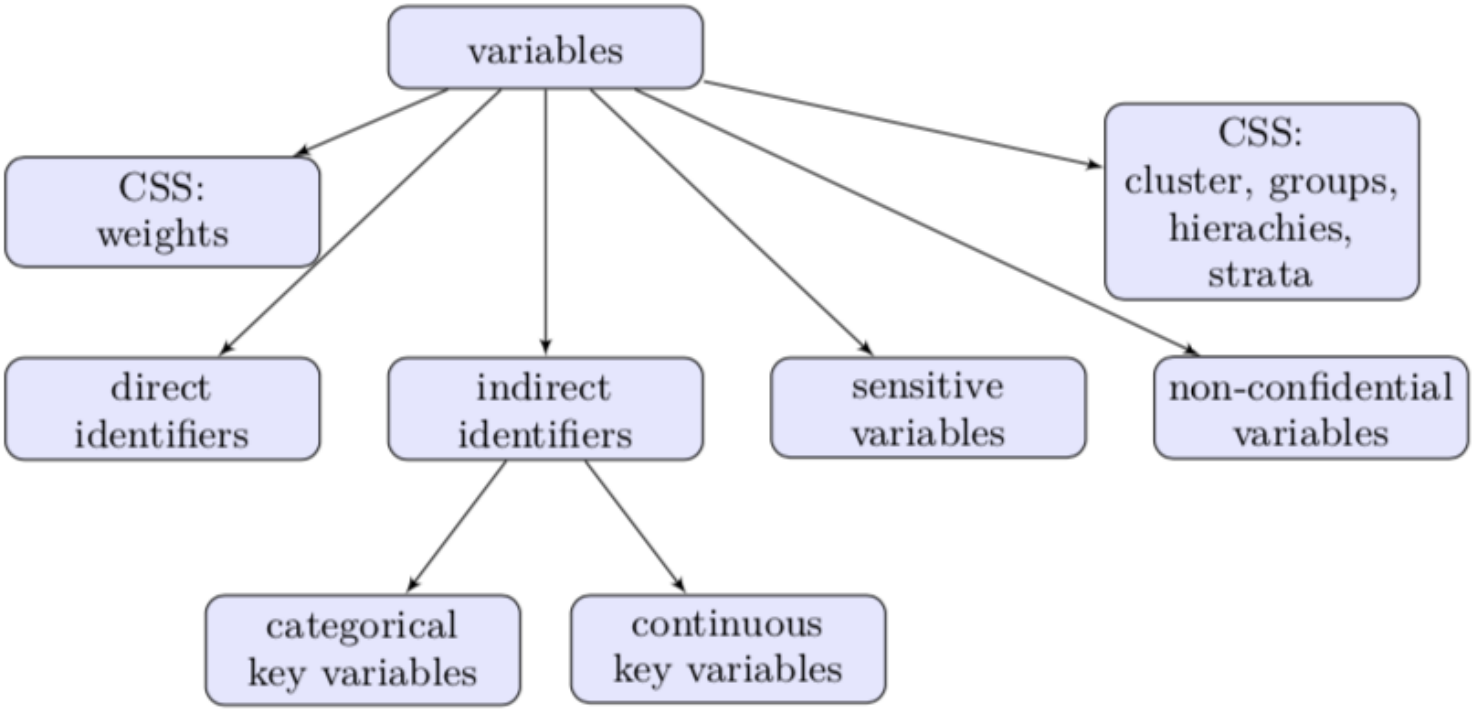
Überblick

Im Folgenden werden vor allem Begriffe eingeführt.

- Variablentypen bzgl. Statistischer Geheimhaltung
- Arten der Re-identifizierung
- Disclosure Risk vs. Data Utility

Buch Kapitel 2

Variablentypen



Arten von Merkmalen in Daten (1/2)

1. Global-eindeutige (z.B. Versicherungsnummer) und **direkte Identifikatoren** (z.B. genaue Adresse) unbedingt löschen oder pseudo-anonymisieren.
2. **Quasi-Identifikatoren** (z.B. PLZ, Alter, Geschlecht), abgekürzt QIDs: Attribute die für eine Re-Identifikation genutzt werden können; heissen auch Key Variables/**Schlüsselvariablen**, Indirect Identifiers oder Implicit Identifiers. Salopp: Jene Variablen die sich mit anderen am Markt erhältlichen Populationen (oder Stichproben) überschneiden.
3. **Sensible Attribute** (z.B. Krankheitsstatus, Kosten, ...): Schützenswerte Informationen mit denen Individuen nicht assoziiert werden wollen.
4. **Nicht-Sensible Attribute**: Informationen deren Kenntnis keinen Datenschutzverstoss darstellt.

Arten von Merkmalen in Daten (2/2)

5. Linked (auch manchmal Ghost-Variablen genannt): Wenn Region suppressed wird, sollte auch Gemeinde-ID suppressed werden.
6. Gewichtsvektor: Stichprobengewichte müssen bei SDC berücksichtigt werden. Risiko einer Identifizierung ist höher, wenn man die Daten der Population hat, als nur von einer Stichprobe.
7. Hierachien/Cluster: Beispiel ist European Structural Earnings Statistics. Hier werden in der ersten Stufe Firmen gezogen, in der 2. Stufe Personen aus den gezogenen Firmen.
   Clusterbespiel: European Statistics on Income and Living Conditions: hier werden Haushalte gezogen und von allen Personen im Haushalt Informationen abgefragt.

Arten der Re-identifizierung (disclosure)

- **Identity disclosure**. Link des Datensatzes mit externen Daten sodass Person identifiziert wird.
  - Beispiel: Versichertendaten über Personen inklusive Krankheitsbilder. Match der Quasi-Identifiers (Alter, Geschlecht, Wohngemeinde) mit Daten von GfK welche Namen enthält. Falls Link für eine Person erfolgreich, weiss der Datenangreifer nun welche Krankheiten diese Person hatte.

Arten der Re-identifizierung (disclosure)

- **Attribute Disclosure**.
  - Beispiel 1: DeStatis gibt Daten an Externe in denen **alle** dunkelhäutigen Menschen zwischen 50 und 60 in Region 1234 röm/kath. sind.

	key variables			sensitive variable
obs	race	age	region	religion
1	black	50–60	1234	roman/catholic
2	black	50–60	1234	roman/catholic
3	black	50–60	1234	roman/catholic
4	black	50–60	1234	roman/catholic

$\longrightarrow$ nun weiss man von diesen **Individuen** eine sensitive Information.

Arten der Re-identifizierung (disclosure)

- **Inferencial Disclosure**. Wenn die modellbasierte einer sensitiven Variable den wahren Wert schätzt.

Beispiel: Wir verwenden das Buch, Seite 40.

Übungen

Exercises Buch Seite 41

Machen Sie sich mit den Daten aus Question 2.1. und 2.2. vom Buch vertraut.

- 2.1: Folgen Sie den Anweisungen
- Diskutieren Sie mit Ihrem Sitznachbarn a-d (Question 2.1 und 2.2)

RU-Maps

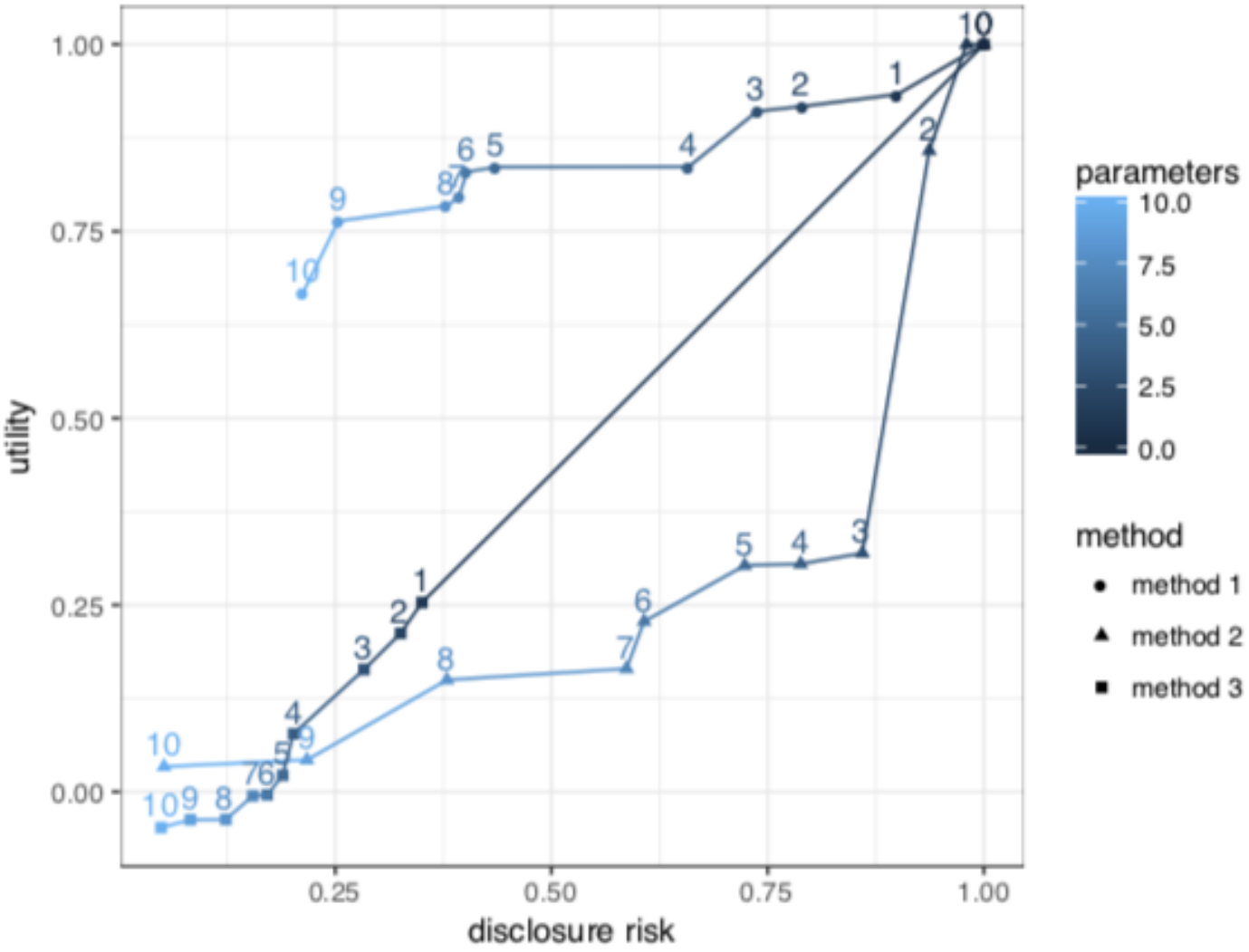
Wir wissen:

- je höher das Re-identifizierungsrisiko von Personen, desto schlechter
- je massiver in die Daten eingegriffen wird um das Risiko zu senken, desto schlechter, da die Datenqualität leidet.

Optimal ist also eine Methode die es schafft, das Risiko zu senken und die Datenqualität hoch hält.

- Um Methoden zu vergleichen und Parametereinstellungen von Methoden zu vergleichen, eignen sich **Risk-Utilitiy Maps**.

RU-Maps


utility
disclosure risk
1.00
0.75
0.50
0.25
0.00
0.25
0.50
0.75
1.00
parameters
10.0
7.5
5.0
2.5
0.0
method
method 1
method 2
method 3

R'U'-Maps (eigentlich R-IL Maps)

- Statt Utility kann der Informationsverlust bewertet werden.