Seminararbeit

Seminar: Spezielle Themen der Wirtschafts–und Sozialstatistik

Anonymisierungsverfahren

Autor: Antonija Tadic
Betreuer: Paul Fink

11. Juni 2016

Institut für Statistik
Ludwig–Maximilians–Universität
München

Inhaltsverzeichnis

1 Einleitung

1.1 Anonymität von Mikrodaten

Die wissenschaftliche Forschung hat ein besonderes Interesse an Mikrodaten bzw. Einzeldaten, da diese ein hohes Analysepotential besitzen. Mit den Einzeldaten können Verhaltensweisen einzelner Unternehmen und Betriebe sowie einzelner Haushalte und Personen beobachtet werden. Da diese Daten sensible Informationen der Auskunftgebenden betreffen, dürfen sie nur anonymisiert an Dritte weitergegeben werden. Um die Anonymität der einzelnen statistischen Objekte zu gewährleisten und einen Zugang zu den Daten für Analysezwecke zu ermöglichen, ist also eine Anonymisierung erforderlich.

Im Bundesdatenschutzgesetz (BDSG) ist die Anonymisierung folgendermaßen definiert:

> *„Anonymisieren ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können.“*

Die personenbezogenen Daten werden also soweit verändert, dass die Zuordnung nur mit einem unverhältnismäßig hohen Aufwand möglich ist. Diesen Vorgang bezeichnet man als *faktische Anonymisierung*. Die Identifizierung der Auskunftgebenden ist hier zwar nicht komplett ausgeschlossen, jedoch übersteigt der Aufwand den Nutzen, den man von einer erfolgreichen Zuordnung hätte.

Unterschieden wird ferner zwischen zwei weiteren Stufen der Anonymität:
Bei der *formalen Anonymisierung* werden die direkten (personenbezogenen) Identifikationsmerkmale aus dem Datensatz entfernt. Diese umfassen alle Angaben, mit denen eine Person direkt identifiziert werden kann. Solche Merkmale sind beispielsweise Namen und Adressangaben. Der Merkmalsumfang sowie die Gliederungen der Merkmale bleiben dagegen erhalten. Die formale Anonymisierung bietet unter Umständen keinen ausreichenden Schutz der Einzeldaten, da bei genügender Detailliertheit der Merkmalsausprägungen eine Reidentifizierung möglich ist.

Dagegen werden die Daten bei einer *absoluten Anonymisierung* so verändert, dass trotz beliebig viel Zusatzwissen eine Reidentifizierung Einzelner nicht möglich ist.

1.2 Nutzung personenbezogener Daten

Das Bundesdatenschutzgesetz (BDSG) und das Bundesstatistikgesetz (BStatG) regeln die mögliche Nutzung personenbezogener Daten. Nach dem Bundesstatistikgesetz dürfen *faktisch anonymisierte Daten* nur wissenschaftlichen Einrichtungen zugänglich gemacht werden. Diesen Einrichtungen stehen für ihre Analysen Scientific-Use-Files (SUF) zur

Verfügung.
Auf *formal anonymisierte Einzeldaten* kann man im Rahmen von kontrollierter Datenfernverarbeitung (KDFV) in Datenzentren zugreifen. Die Ergebnisse der Analyse werden nur in absolut anonymer Form an den Nutzer übermittelt.
Absolut anonyme Einzeldaten können an Einzeldatennutzer frei herausgegeben werden. Die Forschungsdatenzentren der amtlichen Statistik bieten absolut anonymisierte Mikrodaten in Form von Public-Use-Files (PUF) und CAMPUS-Files an, die allen interessierten Personen zur Verfügung gestellt werden.

Analysepotential
Die Analysen mit den anonymisierten Einzeldaten sollen möglichst zu den Originaldaten ähnliche Ergebnisse liefern. Bei einer Anonymisierung werden Daten verändert, wodurch Abweichungen von den Originaldaten entstehen. Eine stärkere Anonymisierung führt also zu einem geringeren statistischen Analysepotential. Daher besitzen formal anonymisierte Daten einen höheren Informationsgehalt als faktisch oder absolut anonymisierte Daten.
In der Praxis wird hauptsächlich die faktische Anonymisierung angewandt.

2 Anonymisierungsverfahren

Anonymisierungsverfahren reduzieren oder verändern die Informationen einer Mikrodatendatei. Sie können in zwei Gruppen eingeteilt werden:

- **Verfahren zur Informationsreduktion**
- **Verfahren zur Datenveränderung**

Die verschiedenen Anonymisierungsverfahren werden in den folgenden Abschnitten eingeführt und anhand eines fiktiven Datenbeispiels (Tabelle 1) veranschaulicht.

	Merkmal 1 **Name**	*Merkmal 2* **Alter**	*Merkmal 3* **Familienstand**	*Merkmal 4* **Wohnort**	*Merkmal 5* **Einkommen**
Datensatz 1	Müller	25	ledig	Bogenhausen	2700
Datensatz 2	Schmidt	29	verheiratet	Altstadt-Lehel	2800
Datensatz 3	Berger	43	ledig	Sendling	3200
Datensatz 4	Hofer	28	ledig	Moosach	2900
Datensatz 5	Lange	47	verheiratet	Maxvorstadt	6200
Datensatz 6	Meier	62	verheiratet	Schwabing	900

Tabelle 1: Beispieldatensatz vor der Anonymisierung.

In diesem Datenbeispiel muss das Merkmal *Name* aus dem Datensatz entfernt werden, damit formale Anonymität gewährleistet ist. Als sensible Information kann hier das Einkommen gesehen werden. Mögliche bekannte Zusatzinformationen könnten Familienstand, das Alter sowie der Stadtbezirk darstellen.

2.1 Verfahren zur Informationsreduktion

Verfahren zur Informationsreduktion werden in der Praxis häufig angewandt. Eine Informationsreduktion erfolgt durch die Unterdrückung oder Vergröberung von Informationen. Sie können grundsätzlich an Merkmalsträgern (*merkmalsträgerbezogene Verfahren*), an Merkmalen (*merkmalsbezogene Verfahren*) oder an einzelnen Ausprägungen (*ausprägungsbezogene Verfahren*) ansetzen.

2.1.1 Merkmalsträgerbezogene Verfahren

Merkmalsträgerbezogene Anonymisierungsverfahren schützen besonders sensible oder auffällige Merkmalsträger. Diese sind einem hohen Reidentifikationsrisiko ausgesetzt.

Entfernen auffälliger Merkmalsträger:
Besonders auffällige Merkmalsträger werden aus dem Datensatz entfernt. Diese Merkmalsträger zeichnen sich durch einzigartige oder seltene Merkmalskombinationen aus, wodurch die Gefahr der Reidentifizierung besteht. Die entfernten Ausreißer sind im Datensatz nicht mehr vorhanden und können somit nicht reidentifiziert werden. Die Entfernung der seltenen Beobachtungen führt allerdings zu Verzerrungen der Ergebnisse, da diese Ausreißer bei der Analyse nicht berücksichtigt werden.

Alter	**Familienstand**	**Wohnort**	**Einkommen**
25	ledig	Bogenhausen	2700
29	verheiratet	Altstadt-Lehel	2800
43	ledig	Sendling	3200
28	ledig	Moosach	2900
47	verheiratet	Maxvorstadt	6200
~~62~~	~~verheiratet~~	~~Schwabing~~	~~900~~

Tabelle 2: Merkmalsträgerbezogene Anonymisierung: *Entfernen auffälliger Merkmalsträger*

Systematische Einschränkung der Grundgesamtheit:
Systematisch abgrenzbare Teilgesamtheiten, die besonders reidentifikationsgefährdet sind, werden aus dem Datensatz entfernt. Dadurch sind sie zwar vor einer möglichen Zuordnung geschützt, allerdings sind für die entfernten Teilgesamtheiten keine Analysen mehr möglich.

(Sub-)Stichprobenziehung:
Ziel ist es den Schutz des gesamten Datenbestandes zu erhöhen. Durch die (Sub-)Stichprobenziehung wird die Teilnahmewahrscheinlichkeit eines Merkmalsträgers verringert. Durch eine Ziehung werden die Merkmalsträger zufällig entfernt. Somit weiß der Angreifer nicht, ob das gesuchte Objekt noch im Datensatz enthalten ist. Einen besseren Schutz vor Reidentifizierung bieten Ziehungen mit Zurücklegen.

2.1.2 Ausprägungsbezogene Verfahren

Bei der Verwendung ausprägungsbezogener Anonymisierungsverfahren werden seltene Werte oder Merkmalskombinationen entfernt. Durch diese Unterdrückung seltener Ausprägungen entstehen fehlende Werte. Dadurch sind auffällige Ausprägungskombinationen nicht mehr im Datensatz enthalten und können somit nicht mehr zugeordnet werden.

Alter	Familienstand	Wohnort	Einkommen
25	ledig	Bogenhausen	2700
29	verheiratet	Altstadt-Lehel	2800
43	ledig	Sendling	3200
28	ledig	Moosach	2900
47	verheiratet	Maxvorstadt	NA ~~6200~~
62	verheiratet	Schwabing	NA ~~900~~

Tabelle 3: Ausprägungsbezogene Anonymisierung

2.1.3 Merkmalsbezogene Verfahren

Beim merkmalsbezogenen Vorgehen zur Informationsreduktion können Merkmale entfernt, ersetzt oder zusammengefasst werden. Dieses Verfahren wird in der Regel auf Überschneidungsmerkmale angewandt, um eine eindeutige Zuordnung zu verhindern. Wird es auf besonders sensible Merkmale angewandt, werden die Anreize des Angreifers verringert eine Reidentifikation vorzunehmen, da der Nutzen einer Zuordnung sinkt.

Beseitigung, Ersetzung oder Zusammenfassung von Merkmalen:
Die Merkmale können aus dem Datensatz entfernt werden oder durch Linearkombinationen, Beziehungs-und Verhältniszahlen oder durch Indexbildung auf einer plausiblen Basis als neues Merkmal ersetzt werden (nur auf metrischen Variablen anwendbar). Bei der Ersetzung von Merkmalen muss die Interpretierbarkeit der neu gebildeten Variablen sichergestellt werden.

Wird eine *Vergröberung von Merkmalsausprägungen* auf metrische Merkmale angewandt, so können Merkmalsausprägungen zu Kategorien zusammengefasst oder die Merkmalswerte durch gerundete Werte ersetzt werden. Für kategoriale Merkmale kann eine Zusammenfassung von bereits vorhandenen Kategorien vorgenommen werden. Bei diesem Vorgehen sinken die Anreize des Angreifers, da mit der Vergröberung Informationen verloren gehen. Dies stellt allerdings auch einen Nachteil dar, da der Informationsgehalt der Merkmale (je nach Grad der Vergröberung) erheblich verringert wird.

Alter	**Familienstand**	**Wohnort**	**Einkommen**
~~25~~	ledig	~~Bogenhausen~~	~~2700~~
~~29~~	verheiratet	~~Altstadt-Lehel~~	~~2800~~
~~43~~	ledig	~~Sendling~~	~~3200~~
~~28~~	ledig	~~Moosach~~	~~2900~~
~~47~~	verheiratet	~~Maxvorstadt~~	~~6200~~
~~62~~	verheiratet	~~Schwabing~~	~~900~~

Alter	**Familienstand**	**Wohnort**	**Einkommen**
0–30	ledig	München-Ost	0–2800
0–30	verheiratet	München-Zentrum	0–2800
> 30	ledig	München-Süd	> 2800
0–30	ledig	München-Nord	> 2800
> 30	verheiratet	München-West	> 2800
> 30	verheiratet	München-West	0–2800

Tabelle 4: Merkmalsbezogene Anonymisierung

2.2 Datenverändernde Verfahren

Bei datenverändernden Anonymisierungsverfahren erfolgt die Anonymisierung durch Veränderungen der Einzeldaten. Unterschieden werden dabei datenverändernde Verfahren für *kategoriale* Variablen bzw. für *metrische* Variablen.
Die verschiedenen Verfahren werden im Folgenden vorgestellt.

2.2.1 Swapping-Verfahren (Vertauschungsverfahren)

Bei diesem Verfahren werden die Werte zwischen den Merkmalsträgern zufällig vertauscht. Befinden sich mehrere sensible Merkmale im Datensatz, so findet die Vertauschung für jedes Merkmal getrennt statt. Das Swapping-Verfahren ist sowohl für kategoriale als auch metrische Variablen durchführbar. Es werden zwei verschiedene Swapping-Verfahren unterschieden:

Einfaches Data-Swapping — Verfahren für kategoriale Variablen
Das Data-Swapping-Verfahren kommt bei kategorialen Variablen zum Einsatz. Die Merkmalsträger werden anhand ausgewählter kategorialer Merkmale gruppiert und anschließend werden innerhalb dieser Gruppen die Merkmalswerte zufällig vertauscht. Ein Nachteil dieses Verfahrens stellt die starke Informationsveränderung dar.

Die Vorgehensweise des einfachen Data-Swapping-Verfahrens wird nun anhand der fiktiven Beispieldaten verdeutlicht. Als Gruppierungsvariable wird das Merkmal *"Familienstand"* herangezogen.

- *Sortierung nach dem Merkmal "Familienstand":*

Alter	**Familienstand**	**Wohnort**	**Einkommen**
25	ledig	Bogenhausen	2700
43	ledig	Sendling	3200
28	ledig	Moosach	2900
29	verheiratet	Altstadt-Lehel	2800
47	verheiratet	Maxvorstadt	6200
62	verheiratet	Schwabing	900

- *Für jede übrige Variable werden die Werte zufällig innerhalb der Gruppe vertauscht:*

Alter	**Familienstand**	**Wohnort**	**Einkommen**
43	ledig	Moosach	2700
28	ledig	Bogenhausen	2900
25	ledig	Sendling	3200
62	verheiratet	Altstadt-Lehel	6200
29	verheiratet	Schwabing	900
47	verheiratet	Maxvorstadt	2800

Rank-Swapping — Verfahren für metrische Variablen
Bei diesem Verfahren werden zunächst die Merkmalswerte jeder Variablen der Größe nach sortiert. Anschließend werden die Werte innerhalb der festgelegten Gruppen zufällig vertauscht.

Bei den Swapping-Verfahren bleiben die univariaten Verteilungen erhalten, während sich die gemeinsame Verteilung der Merkmale verändert.

2.2.2 Post-Randomisierung (PRAM)

Dieses Verfahren wird für kategoriale Variablen eingesetzt. Dabei werden die Merkmalsausprägungen nur mit einer bestimmten Wahrscheinlichkeit verändert. Die Merkmalswerte werden durch eine festgelegt Übergangswahrscheinlichkeit p randomisiert und somit mit einer bestimmten Wahrscheinlichkeit anderen Kategorien zugeordnet. Durch die Post-Randomisierung entsprechen die Werte des anonymisierten Datensatzes den Werten im Originaldatensatz nur noch mit der in dem Verfahren festgelegten Wahrscheinlichkeit.

2.2.3 Stochastische Überlagerung

Unter der Methode der stochastischen Überlagerung versteht man das Hinzufügen von Zufallszahlen zu den metrischen Variablen. Die ursprünglichen Werte werden durch die Werte, die durch diese Überlagerung entstehen, ersetzt. Die Überlagerung kann durch Addierung oder Multiplizierung von Zufallszahlen erfolgen.

2.2.4 Imputationsverfahren

Imputationsverfahren werden üblicherweise zur Behandlung fehlender Werte eingesetzt. Nun werden nicht "Missing Values", sondern die zu anonymisierenden Merkmalswerte durch die eingeschätzten Werte ersetzt. Man unterscheidet die folgenden zwei Imputationsverfahren:
Bei der *einfachen Imputation* werden die zu anonymisierenden Merkmalswerte durch den geschätzten Wert eines Regressionsmodells ersetzt. Die *multiple Imputation* schätzt mehrere Regressionsmodelle für die zu ersetzenden Werte und liefert somit plausiblere Schätzwerte.

2.2.5 SAFE-Verfahren

Das SAFE-Verfahren ist ein Verfahren der Mikroaggregation. Die Werte werden so vertauscht bzw. verändert, dass jeder Merkmalsträger mit mindestens zwei weiteren Merkmalsträgern bezüglich aller beobachteten Merkmale identische Ausprägungen aufweist. Da mehrere gleiche Objekte vorhanden sind wird das Risiko einer eindeutigen Zuordnung reduziert.

3 Mikroaggregationsverfahren

Die Mikroaggregation gehört zu den datenverändernden Anonymisierungsverfahren und wird auf metrische Variablen angewandt. Zunächst werden Objekte zu Gruppen zusam-

mengefasst und anschließend die ursprünglichen Werte durch das jeweilige arithmetische Gruppenmittel ersetzt. Besteht eine Gruppe nur aus zwei Objekten, kann man bei der Kenntnis der Werte eines Merkmalsträgers auf den anderen Merkmalsträger schließen. Um also das Risiko einer eindeutigen Zuordnung zu verringern, sollten die Gruppen mehr als zwei Objekte enthalten.

Man unterscheidet folgende zwei Arten der Mikroaggregation:
Die *deterministische Mikroaggregation* fasst ähnliche Objekte zu Gruppen zusammen, während bei der *stochastischen Mikroaggregation* die Gruppenbildung zufällig erfolgt.
Sie unterscheiden sich also nach der Bestimmung der Gruppen. Im Folgenden werden diese zwei Typen der Mikroaggregation vorgestellt.

3.1 Deterministische/Abstandsorientierte Mikroaggregation

Bei dieser Mikroaggregation werden zunächst möglichst ähnliche Merkmalsträger zu Gruppen zusammengefasst und anschließend die ursprünglichen Werte durch das jeweilige arithmetische Gruppenmittel ersetzt. Man unterscheidet dabei folgende Arten:

- *Mehrdimensionale Mikroaggregation (gemeinsame Mikroaggregation):* Die Gruppierung der Variablen erfolgt gemeinsam.
- *Eindimensionale Mikroaggregation (getrennte Mikroaggregation):* Die Variablen werden getrennt mikroaggregiert.
- *Gruppierte Mikroaggregation:* Stark korreliere Variablen werden zu Gruppen zusammengefasst.

3.1.1 Gemeinsame Mikroaggregation

Gemeinsame Mikroaggregation nach einer Variablen:
Bei dieser Methode wird eine dominierende Variable festgelegt und der Datensatz absteigend danach sortiert. Die dominierende Variable sollte dabei mit möglichst vielen weiteren Merkmalen stark korreliert sein. Anschließend werden immer drei benachbarte Merkmalsträger in einer Gruppe zusammengefasst und ihre stetigen Merkmalswerte durch das arithmetische Mittel innerhalb der Gruppen ersetzt.

Gemeinsame Mikroaggregation nach einer Hilfsvariablen:
Die Gruppenbildung erfolgt hier anhand der absteigenden Sortierung einer Hilfsvariablen. Diese kann eine durch Transformation gebildete Variable sein und sollte eine möglichst hohe Korrelation zu den anderen Variablen aufweisen.

Gemeinsame Mikroaggregation nach allen p metrischen Variablen:

Zunächst wird die euklidische Distanz zwischen den Merkmalsträgern ermittelt.

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^{p} (x_{i,j} - x_{k,j})^2}$$

x_i, x_k: Datenvektoren von zwei Merkmalsträgern.
Die Merkmalsträger mit dem größten Abstand werden mit den zwei Merkmalsträgern mit der jeweils geringsten Distanz gruppiert. Analog erfolgt die Gruppenbildung der noch nicht gruppierten Merkmalsträger.

3.1.2 Getrennte Mikroaggregation

Hier geht man ähnlich vor wie bei der gemeinsamen Mikroaggragation: Zunächst wird der Datensatz nach der zu anonymisierenden Variable sortiert, in Gruppen von mindestens drei Merkmalsträgern eingeteilt und anschließend die stetigen Merkmalswerte durch die Gruppenmittelwerte ersetzt. Die metrischen Variablen werden jedoch nicht gleichzeitig durch das arithmetische Gruppenmittel ersetzt, sondern dieser Vorgang wird für die einzelnen zu anonymisierenden Variablen getrennt vorgenommen.

3.1.3 Gruppierte Mikroaggregation

Dieses Verfahren gruppiert stark korrelierte Variablen. Innerhalb der Gruppen erfolgt anschließend eine gemeinsame Mikroaggregation.

- *Sortierung nach dem Merkmal Ëinkommen":*

Alter	**Familienstand**	**Wohnort**	**Einkommen**
47	verheiratet	Maxvorstadt	6200
43	ledig	Sendling	3200
28	ledig	Moosach	2900
29	verheiratet	Altstadt-Lehel	2800
25	ledig	Bogenhausen	2700
62	verheiratet	Schwabing	900

- *Bildung der Gruppenmittelwerte:*

Alter	Familienstand	Wohnort	Einkommen
39	verheiratet	Maxvorstadt	4100
39	ledig	Sendling	4100
39	ledig	Moosach	4100
38	verheiratet	Altstadt-Lehel	2133
38	ledig	Bogenhausen	2133
38	verheiratet	Schwabing	2133

Tabelle 5: Anonymisierung des Datensatzes mit der Methode der deterministischen Mikroaggregation nach der Variable *Einkommen*

3.2 Stochastische Mikroaggregation

Anders als bei den deterministischen Mikroaggregationsverfahren werden bei bei der stochastischen Mikroaggregation die Gruppen zufällig gebildet. Man unterscheidet dabei folgende Arten:

Zufällige Mikroaggregation:
Das Verfahren ist identisch zu der Vorgehensweise der deterministischen Mikroaggregation und kann entweder für alle Variablen gemeinsam oder für alle Variablen getrennt erfolgen. Der einzige Unterschied liegt bei der Gruppenbildung: hier erfolgt die Gruppenbildung nicht nach Ähnlichkeit der Merkmalsträger, sondern nach Zufälligkeit.

Bootstrap-Mikroaggregation:
Die Gruppenzuteilung für einen Merkmalsträger erfolgt durch das zufällige Ziehen zweier weiterer Objekte. Dabei können auch die gleichen Merkmalsträger gezogen werden, da die Ziehung mit Zurücklegen erfolgt. Die Merkmalswerte der jeweiligen Gruppen werden durch die Werte des ersten Objektes in der Gruppe ersetzt.

4 Fazit

Die verschiedenen Anonymisierungsmethoden lassen sich unterteilen in Verfahren zur Informationsreduktion bzw. datenverändernde Verfahren. Je nach Datensatz muss geprüft werden welches Verfahren sich am besten eignet. Die Auswahl der Verfahren spielt dabei eine große Rolle für den Informationsgehalt der Mikrodaten, denn ihre Anwendung hat Auswirkungen auf deren Analysepotential. Methoden, bei denen durch eine Anonymisierung die Daten stark verändert werden verringern das Analysepotential deutlich. Es soll ein ausreichender Schutz vor einer Reidentifizierung erreicht werden aber gleichzeitig ist auch zu beachten, dass der Informationsgehalt dieser Daten durch eine Anonymisierungsmethode nicht zu stark eingeschränkt wird.