Ludwig-Maximilians-Universität München
Institut für Statistik

Seminararbeit

Anonymisierungsverfahren

Seminar "Statistische Herausforderungen im Umgang mit fehlenden bzw. fehlerbehafteten Daten"

Autor: Ye Bin Park

Leitung: Prof. Dr. Augustin

15. März 2015

Zusammenfassung

Diese Arbeit beschäftigt sich mit den verschiedenen, gängigen Anonymisierungsverfahren der Mikrodaten. Nach Klärung allgemeiner Definitionen und Begrifflichkeiten werden zunächst die Gründe und Ziele, die bei der Anonymisierung von Mikrodaten zu betrachten und zu verfolgen sind, dargestellt. Anschließend werden auf verschiedenen Anonymisierungsverfahren, die hauptsächlich in zwei Gruppen, nämlich in Verfahren zur Informationsreduktion und Verfahren zur Datenveränderung, eingeteilt werden können, im Einzelnen aufgeführt und beschrieben.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Immer mehr an Bedeutung gewinnt in den heutigen Sozial- und Wirtschaftswissenschaften das Thema Datenschutz. Gerade in diesen Bereichen werden besonders viele personen-, haushalts- und unternehmensbezogene Daten erhoben und vielseitig, beispielsweise für verschiedene statistische Analysen, eingesetzt.
Um diese erhobenen Mikrodaten an das Dritte weitergeben zu können, müssen in Deutschland gesetzliche Beschränkungen eingehalten werden. Dies ist unter anderem durch Anonymisierung ermöglicht. Was genau versteht man aber unter *Anonymisierung*? Im *Bundesdatenschutzgesetz (BDSG)* ist sie 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.

- Vgl. BDSG §3 Abs.6

Nach dem *Bundesdatenschutzgesetz* ist Anonymisierung also ein Vorgang, bei dem die Daten so verändert werden, dass sie nicht mehr oder nur mit einem enormen Aufwand einer Person zugeordnet werden können. Die Daten, die durch den Einsatz von Anonymisierungsverfahren verändert wurden, werden dann als *anonym* bezeichnet.

1.1. Gründe und Ziele der Anonymisierung

Ein Grund, warum die Anonymisierung von Mikrodaten notwendig ist, liegt, wie bereits erwähnt, in der gesetzlichen Verpflichtung nach dem *Bundesdatenschutzgesetz* und dem *Bundeststatistikgesetz*. Im Bundesstatistikgesetz heißt es:

(1) Einzelangaben über persönliche und sachliche Verhältnisse [...] sind geheimzuhalten, soweit durch besondere Rechtsvorschrift nichts anderes bestimmt ist. Dies gilt nicht für
1. Einzelangaben, in deren Übermittlung oder Veröffentlichung der Befragte schriftlich eingewilligt hat,
2. Einzelangaben aus allgemein zugänglichen Quellen [...] auch soweit eine Auskunftspflicht [...] besteht,
3. Einzelangaben, die [...] mit den Einzelangaben anderer Befragter zusammengefaßt

und in statistischen Ergebnissen dargestellt sind,
4. Einzelangaben, wenn sie dem Befragten oder Betroffenen nicht zuzuordnen sind. [...]
[...]
(6) Für die Durchführung wissenschaftlicher Vorhaben dürfen [...] Einzelangaben [...] übermittelt werden, wenn die Einzelangaben nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können [...].

- Vgl. BStatG §16 Geheimhaltung

Demnach dürfen Daten, die dem Befragten oder Betroffenen, von dem die Daten stammen, nicht zugeordnet werden können, also wenn sie *anonym* sind, ohne weitere Maßnahmen veröffentlicht und verarbeitet werden.
Desweiteren spielt Anonymisierung eine wichtige Rolle für die Auskunftsbereitschaft der Befragten. Befragung ist die häufigste Art der Datenerhebung und die statistischen Analysen bedingen auf Daten, die von Befragten bereitgestellt werden. Ohne diese Daten gäbe es keine Informationsgrundlage, auf der die statistischen Analysen basieren sollen. Um so wichtiger ist es daher, die Auskunftsbereitschaft zu sichern und zu steigern. Als Maßnahme für die Sicherung kann die Anonymisierung gezählt werden. Durch Anonymisierung wird die Anonymität der Befragten gewährleistet und der denkbare Missbrauch, vor dem die Befragten womöglich zu befürchten haben, kann verhindert werden. Anonymisierung dient also in erster Linie dem Schutz der einzelnen Personen oder einzelnen Gruppen vor Re-Identifizierung.
Da Anonymisierung immer mit Veränderungen und Verluste von Daten verbunden ist, aber man dennoch versucht, bestmögliches Analysepotentials beizubehalten, soll bei der Durchführung im Auge behalten werden, dass man die Daten nur soweit verändert, wie es für die Erreichung der Anonymität erforderlich ist, um möglichst zu den Originaldaten ähnliche, sinnvolle Ergebnisse der statistischen Analyse zu erhalten. Dabei sind diejenigen Verfahren auszuwählen, die das Analysepotential möglichst wenig beeinflussen.

1.2. Grad der Anonymisierung

Man unterscheidet zwischen drei verschiedenen *Stärkegraden der Anonymisierung*:

(1) Formale Anonymisierung

Formal anonymisierte Daten liegen dann vor, wenn die direkten Identifizierungsmerkmale, wie z.B. der Name, vom Datensatz entfernt werden. Dabei bleiben alle andere

Merkmale unverändert enthalten.

(2) Faktische Anonymisierung

Bei *faktisch anonymisierten* Daten ist die Identifizierung der Auskunftsgebenden zwar nicht komplett ausgeschlossen, aber die Daten werden soweit verändert, dass die Zuordnung nur mit einem *unverhältnismäßig hohen Zeit-, Kosten- und Arbeitsaufwand* möglich ist.

(3) Absolute Anonymisierung

Daten werden als *absolut anonym* bezeichnet, wenn die Identifizierung der Auskunftsgebenden gänzlich ausgeschlossen werden kann. Aufgrund der Existenz vieler Faktoren, die die Identifizierung der Auskunftsgebenden ermöglicht, wie z.B. die zur Verfügung stehenden technischen Möglichkeiten der Datenverarbeitung und das mögliche Zusatzwissen, kommen absolut anonymisierte Daten in der Praxis äußerst selten vor.

In der folgenden Abbildung ist der Zusammenhang zwischen dem Grad der Anonymisierung und dem Grad des Analysepotentials veranschaulicht.

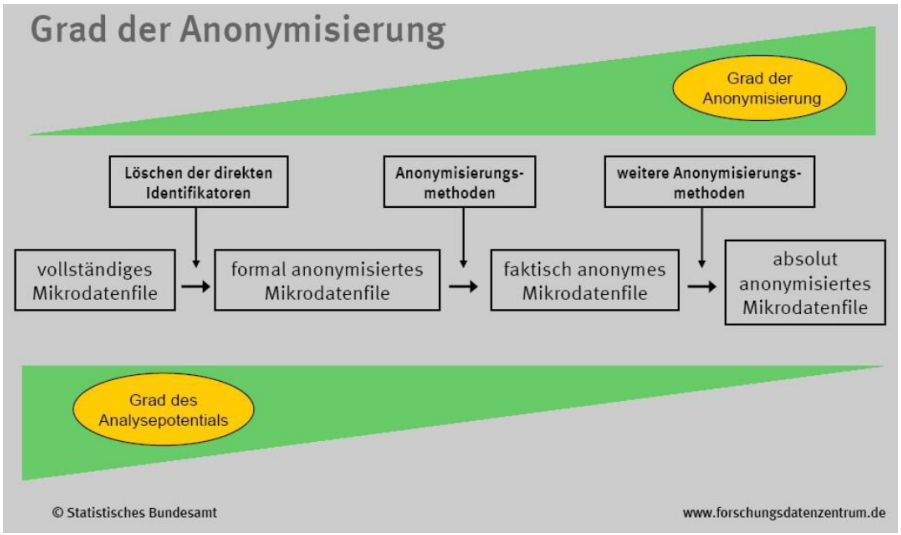

Abbildung 1.1: Grad der Anonymisierung (Quelle: Forschungsdatenzentrum)

Aus der Abbildung kann man entnehmen, dass das Analysepotential mit steigendem

Anonymisierungsgrad abnimmt. Dieser Zusammenhang ist selbstverständlich, denn bei der Anonymsierung werden Daten entweder entfernt, ersetzt oder verändert, wodurch Abweichungen von den Originaldaten unvermeidlich sind. Dies hat zur Folge, dass mit steigender Stärke der Anonymisierung die Aussagekraft der Daten beeinträchtigt wird.

2. Anonymisierungsverfahren

In der Praxis wird hauptsächlich die *faktische Anonymisierung* vorgenommen. Daher werden im Folgenden Anonymisierungsverfahren vorgestellt, die der faktischen Anonymität dienen. Anonymisierungsverfahren können im Großen und Ganzen in zwei Verfahrensgruppen eingeteilt werden:

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

In den folgenden Abschnitten werden die verschiedenen Anonymisierungsverfahren im Einzelnen aufgeführt und beschrieben.

2.1. Verfahren zur Informationsreduktion

Verfahren zur Informationsreduktion zählen zu den *traditionellen Verfahren* und finden in der Praxis häufig Anwendung. Es werden Informationen reduziert, in dem sie unterdrückt oder vergröbert werden. Dabei ist zu unterscheiden, ob sie *merkmalsträger-, merkmals- oder ausprägungsbezogen* eingesetzt werden.

2.1.1. Merkmalsträgerbezogene Informationsreduktion

Merkmalsträgerbezogene Anonymisierungsverfahren werden eingesetzt, um Merkmalsträger, die besonders auffällig oder hohem Reidentifikationsrisiko ausgesetzt sind, zu schützen.

- **Entfernen auffälliger Merkmalsträger**
  Besonders auffällige Merkmalsträger, d.h. Merkmalsträger mit ausgeprägten oder seltenen Merkmalsausprägungen, werden aus dem Datensatz entfernt. Der Vorteil an diesem Verfahren ist, dass die entfernten Ausreißer nicht mehr im Datensatz enthalten sind und somit nicht reidentifiziert werden können. Dies bringt jedoch einen erheblichen Nachteil für die späteren Analysen mit sich, denn diese Ausreißer sind gerade die Merkmalsträger, die einen größeren Einfluss auf Ergebnisse statistischer Analysen ausüben und demnach am interessantesten zu untersuchen wären. Durch Entfernung solcher Ausreißer bleiben sie bei der Analyse unberücksichtigt, welches zu Verzerrungen der Ergebnisse führt.
- **Systematische Einschränkung der Grundgesamtheit**

Teilgesamtheiten des Datenbestandes, welche systematisch abgegrenzt werden können und besonders reidentifikationsgefährdet sind, werden aus dem Datensatz entfernt.

- **(Sub-)Stichprobenziehung**
  Aus dem Datenbestand wird eine Zufallsstichprobe gezogen. Möglich wären auch Ziehungen mit vorgegebenen Auswahlwahrscheinlichkeiten oder Ziehungen mit Zurücklegen. Durch Ziehung von Stichproben wird die Teilnahmewahrscheinlichkeit eines Merkmalsträgers verringert und die Unsicherheit, ob das gesuchte Objekt im Datensatz enthalten ist, hervorgerufen.

2.1.2. Merkmalsbezogene Informationsreduktion

Merkmalsbezogene Anonymisierungsverfahren werden zur Behandlung einzelner oder mehrerer Merkmalsträger eingesetzt. Dieses Verfahren wird in der Regel auf *Überschneidungsmerkmale* verwendet um eine eindeutige Zuordnung zu verhindern oder auf besonders sensible Merkmale, um die wahren Werte vor Reidentifikation zu schützen.

- **Beseitigung, Ersetzung oder Zusammenfassung von Merkmalen**
  Merkmale werden entweder vollständig aus dem Datenbestand entfernt (*Unterdrückung einzelner Variablen*) oder duch Linearkombinationen, Indizes oder Verhältniszahlen als neues Merkmal ersetzt.

- **Vergröberung von Merkmalsausprägungen**
  Es wird unterschieden, auf welche Arten von Merkmalsausprägungen das Verfahren anzuwenden ist. Bei *metrischen* Merkmalen 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.

2.1.3. Ausprägungsbezogene Informationsreduktion

Unter ausprägungsbezogene Anonymisierungsverfahren versteht man in der Regel die *Local Suppression*. Bei Local Suppression werden Merkmale mit seltenen oder einzigartigen Ausprägungen oder Ausprägungskombinationen unterdrückt. Dadurch entstehen fehlende Werte (*Missing Values*) im Datensatz. Durch diese Unterdrückung sind seltene, auffällige Ausprägungskombinationen nicht mehr im Datensatz enthalten und sind

somit nicht reidentifizierbar.

2.2. Verfahren zur Datenveränderung

Bei datenverändernden Verfahren wird die Anonymität durch Veränderungen der Einzeldaten sichergestellt. Im Wesentlichen wird hierbei unterschieden, ob diese auf *kategorialen* oder auf *metrischen* Variablen angewendet werden.

2.2.1. Swapping-Verfahren

Das Swapping-Verfahren basiert auf der *Vertauschung* von existierenden Merkmalsausprägungen zwischen verschiedenen Merkmalsträgern. Sind nicht nur eine, sondern mehrere Variablen vorhanden, findet die Vertauschung für jede Variable getrennt statt. Man unterscheidet zwischen zwei verschiedene Swapping-Verfahren:

- **Einfaches Data-Swapping**
  Das Data-Swapping-Verfahren kommt bei *kategorialen* Variablen zum Einsatz. Zuerst werden die Merkmalsträger anhand ausgewählter kategorialer Merkmale gruppiert. Innerhalb dieser Gruppen werden dann die Merkmalswerte zufällig getauscht.
- **Rank-Swapping**
  Dieses Verfahren findet bei *metrischen* Variablen die Anwendung. Es werden zunächst die Merkmalswerte für jede einzelne Variable nach ihrer Größe sortiert. Die Werte werden innerhalb der vorher festgelegten Nachbarschaftsbereiche zufällig getauscht.

Die Schutzfunktion des *Swapping-Verfahrens* besteht in der Erschwerung von Zuordnung durch Änderung der Merkmalswerte. Durch den Einsatz dieses Verfahrens kommt es aber zur starken Informationsveränderung. Es ist daher ratsam, Veränderungen der Merkmalsausprägungen nur mit einer festgelegten Wahrscheinlichkeit vorzunehmen. Dies geschieht bei folgendem Verfahren, nämlich bei *Post-Randomisierung*.

2.2.2. Post-Randomisierung (PRAM)

Das Verfahren Post-Randomisierung wird für *kategoriale* Variablen eingesetzt. Dabei werden die vorhandenen Merkmalswerte mit einer vorher festgelegten *Übergangswahrscheinlichkeit* p randomisiert, d.h. die Merkmalswerte werden bewusst zu falschen Ka-

tegorien zugeordnet. Die *Übergangswahrscheinlichkeit p* für *dichotome* Variablen lässt sich folgendermaßen darstellen (Ronning et al. 2002):

$$p = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} \tag{1}$$

Die *Übergangswahrscheinlichkeitsmatrix* kann beliebige Strukturen aufnehmen, wobei die Einträge auf positiven Werte eingeschränkt sind. Ist dem Datennutzer die *Übergangswahrscheinlichkeitsmatrix p* bekannt und es existiert die *Inversematrix* p^{-1}, dann besteht für ihn die Möglichkeit, die mit PRAM behandelten Daten zu korrigieren und an die wahren Werte heranzukommen.

2.2.3. SAFE-Verfahren

Das SAFE-Verfahren ist ein Verfahren zur ***s**icheren **A**nonymisierung **f**ür **E**inzeldaten,* welches ursprünglich von Mitarbeitern des Statistischen Landesamtes Berlin-Brandenburg entwickelt wurde. Das Verfahren gehört zur *Mikroaggregation* und wird zur Behandlung *kategorialer* Variablen eingesetzt. Beim SAFE-Verfahren wird ein Datensatz erzeugt, in dem jeder Merkmalsträger mindestens zwei weitere Merkmalsträger, die bezüglich aller beobachteten Merkmale identische Ausprägungen aufweisen, besitzt. Dies wird durch gezielte Veränderung der Merkmalswerte realisiert. Dabei soll die Abweichung in den Häufigkeitsverteilungen möglichst minimal gehalten werden, also die ursprüngliche Häufigkeitsverteilung wenn möglich erhalten bleiben.

2.2.4. Mikroaggregation

Mikroaggregation, eines der wichtigsten Anonymisierungsverfahren, wird auf *metrischen* Variablen angewendet und dient zur Reduzierung der eindeutigen Zuordnungsmöglichkeit der Merkmalsträger. Es basiert auf der Idee, ähnliche Objekte zu kleinen Gruppen zusammenzufassen und die ursprünglichen Werte durch das jeweilige Gruppenmittel zu ersetzen. Um die Reidentifikationsrisiko zu verringern, sollen dabei die Gruppen aus mindestens drei Objekten bestehen.

Es existieren zwei Arten von Mikroaggregation:

- *Deterministische Mikroaggregation*
  Bei der deterministischen Mikroaggregation werden möglichst ähnliche Merkmalsträger zu Gruppen zusammengefasst und ihre ursprünglichen Werte durch

das arithmetische Gruppenmittel ersetzt. Dabei wird eine zusätzliche Unterscheidung darüber getroffen, ob die Mikroaggregation für alle Variablen gemeinsam erfolgt (*gemeinsame Mikroaggregation*) oder die Variablen einzeln (*getrennte Mikroaggregation*) mikroaggregiert werden.

- **Gemeinsame Mikroaggregation nach einer Variable**
  Zuerst wird eine *dominierende Variable* festgelegt und die Merkmalsträger nach ihr sortiert. Danach werden aus jeweils drei benachbarten Merkmalsträgern Gruppen gebildet und alle ihre metrischen Merkmalswerte durch das arithmetische Mittel innerhalb der jeweiligen Gruppen ersetzt.
- **Gemeinsame Mikroaggregation nach allen metrischen Variablen**
  Die Gruppenbildudng basiert auf der *euklidischen Distanz*

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

  Es werden zwei Merkmalsträger, die am weitesten entfernt liegen, herausgesucht. Sie bilden jeweils mit ihren zwei dichtesten Merkmalsträgern eine Gruppe. Dieser Vorgang wird solange wiederholt, bis alle Merkmalsträger einer Gruppe zugeordnet sind. Anschließend werden die Merkmalswerte durch das Gruppenmittel ersetzt.
- **Getrennte Mikroaggregation**
  Die Funktionsweise ist identisch wie bei der gemeinsamen Mikroaggregation, bis auf den Unterschied, dass nicht alle metrischen Variablen gleichzeitig durch das arithmetische Gruppenmittel ersetzt werden, sondern der Vorgang für einzelne, zu anonymisierende Variablen getrennt vorgenommen wird.

- **Stochastische Mikroaggregation**
  Anders als bei der deterministischen Mikroaggregation beruht die Gruppenbildung der stochastischen Mikroaggregation auf *Zufälligkeit*. Diese lässt sich in zwei Methoden unterteilen:
  - **Zufällige Mikroaggregation**
    Das Prinzip der zufälligen Mikroaggregation ist identisch wie bei deterministischen Mikroaggregation und kann entweder für alle Variablen *gemeinsam* oder für alle Variablen *getrennt* erfolgen. Der Unterschied zur determinis-

tischen Mikroaggregation liegt darin, dass bei zufälliger Mikroaggregation die Gruppen rein zufällig und nicht nach Ähnlichkeit der Merkmalsträger gebildet werden.

- **Bootstrap-Mikroaggregation**
  Bei Bootstrap-Mikroaggregation erfolgt die Auswahl der Merkmalsträger durch *Zufallsziehungen mit Zurücklegen*. Dabei werden für jeden Merkmalsträger zwei weitere Merkmalsträger gezogen, welche zusammen eine Gruppe bilden.

2.2.5. Stochastische Überlagerung

Unter der stochastischen Überlagerung versteht man das *Hinzufügen eines zufälligen Messfehlers* zu den *metrischen* Variablen. Diese Überlagerung erfolgt entweder *additiv* durch Addierung von Zufallszahlen oder *multiplikativ* durch Multiplizierung von Zufallszahlen. Es werden in der Regel folgende Annahmen getroffen: (I) die zu beobachtenden Variablen sind normalverteilt, (II) W, eine Matrix aus Zufallszahlen, enthält nur positive Elemente, (III) der Originalwert X ist unabhängig von W.

- **Additive stochastische Überlagerung**
  Additive Überlagerung lässt sich als

$$X^a = X + W \tag{3}$$

  formal darstellen. Dem *Originalwert X* wird eine Matrix aus Zufallszahlen W hinzuaddiert. Es gilt dabei: *E(W) = 0*

- **Multiplikative stochastische Überlagerung**
  Bei *multiplikativer Überlagerung* wird der Wert X mit der Zufallszahlenmatrix W multipliziert, mit der Annahme, dass *E(W) = 1*:

$$X^a = X * W \tag{4}$$

Diese zusätzlichen Annahmen sind notwendig, um das Erwartungswertverhältnis

$$E(W) = E(W^a) \tag{5}$$

beizubehalten. Probleme ergeben sich jedoch unter der Normalverteilungsannahme, da der größte Teil der Merkmalswerte um den Erwartungswert liegt. Daher bietet es sich an, andere Verteilungsannahmen zu treffen. Eine mögliche, in Frage kommende Verteilung wäre die *gestutzte Normalverteilung*. Bei gestutzter Normalverteilung werden Bereiche nahe dem Erwartungswert und extrem außerhalb des Erwartungswertes ausgeschlossen. Somit kann auf die unerwünschten Eigenschaften des Erwartungswertes verzichtet werden. Eine Alternative besteht in der *Mischverteilung mehrerer Normalverteilungen*. Dabei werden mehrere Normalverteilungen so zusammengestellt, bis die gewünschte Eigenschaft bezüglich des Erwartungswertes - beispielsweise, dass die Werte nicht dicht am Erwartungswert liegen - erreicht wird.
Eine Korrektur der stochastisch überlagerten Merkmalswerte bzw. eine Schätzung der Originalwerte ist unter anderem durch die Likelihood-Methode für Messfehler möglich.

2.2.6. Simulationsverfahren

Durch den Einsatz des Simulationsverfahrens werden mit Hilfe stochastischer Verfahren synthetische Merkmalsträger erzeugt. Dabei entspricht die Anzahl der Merkmalsträger im Originaldatensatz in den meisten Fällen nicht der Anzahl der Merkmalsträger von synthetisch erzeugtem Datensatz. Dies führt dazu, dass kleinere bzw. größere Datensätze entstehen.

- **Resampling**
  Das Resampling-Verfahrens ist eine statistische Methode zur Erzeugung synthetischer Datensätze, welche auf der Schätzung der mehrdimensionalen Kerndichte basiert. Dabei wird zuerst die mehrdimensionale Kerndichte geschätzt und unter Verwendung dieser Schätzung werden Merkmalsträger synthetisch erzeugt.
- **Latin Hypercube Sampling (LHS)**
  Beim Latin Hypercube Sampling werden Merkmalswerte unter Verwendung der geglätteten empirischen Verteilungsfunktion oder einer theoretischen Verteilungsfunktion synthetisch erzeugt. Anschließend erfolgt eine Umordnung mit Hilfe eines Swapping-Verfahrens mit dem Ziel, möglichst Originaldaten ähnliche Rangkorrelationen zu generieren.

2.2.7. Imputationsverfahren

Imputationsverfahren werden üblicherweise zur Behandlung fehlender Daten (*Missing Values*) eingesetzt. Im Gegensatz zu ihrem klassischen Einsatz werden sie hier nicht für fehlende Werte, sondern für besonders auffällige und daher anonymisierungsbedürftige Merkmalswerte eingesetzt. Man unterscheidet zwischen zwei Imputationsverfahren:

- **Single Imputation (Einfache Imputation)**
  Single Imputation ersetzt für jeden zu anonymisierenden Wert einen durch das Regressionsmodell geschätzten Wert. Aufgrund der einmaligen Schätzung wird hier die Unsicherheit, die durch Imputation entstehen, nicht berücksichtigt. Um dieser Unsicherheit entgegenzukommen, wird multiple Imputation eingesetzt.
- **Multiple Imputation**
  Multiple Imputation ist eine Erweiterung der Single Imputation. Für jeden zu ersetzenden Wert erfolgen mehrere Regressionsschätzungen und liefern somit plausiblere Schätzwerte als bei Single Imputation.

2.3. Auswahl der Verfahren

Die in dieser Arbeit vorgestellten Anonymisierungsverfahren behandeln die zu anonymisierende Variablen verschiedenartig und üben somit unterschiedlichen Einfluss auf die Ergebnisse der statistischen Analyse aus. Um möglichst sinnvolle Ergebnisse zu erzielen, ist es daher angebracht, je nach Zielsetzung, bei der Verfahrensauswahl verschiedene Kriterien in Betracht zu ziehen. *Das statistische Bundesamt* schlägt folgende Kriterien zur Verfahrensauswahl vor *(Vgl. Statistik und Wissenschaft - Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten)*:

- **Leichte Handhabbarkeit des Verfahrens**
  Dieses Kriterium ist notwendig, da die Verfahren nicht nur theoretisch, sondern auch praktisch durch das Personal durchgeführt werden müssen.
- **Erfolgsaussichten des Verfahrens**
  Die Vor- und Nachteile und die praktische Einsetzbarkeit des jeweiligen Verfahrens sind zu überprüfen, um die Anonymisierung erfolgreich zu vollbringen.
- **Repräsentative Vertretung der Verfahrensgruppen**
  Da jede Verfahrensgruppe einen anderen Ansatz der Anonymisierung repräsentiert, sollen bei der Verfahrensauswahl alle Verfahrensgruppen in Erwägung ge-

zogen werden.

- **Methodenmix von Verfahren**
  Eine sinnvolle Anonymisierung ist oftmals nur unter Verwendung mehrerer Verfahren zu erreichen. Daher sollen alle Verfahren, die zusammengesetzt werden können, berücksichtigt werden.

3. Fazit

Eine immer größere Rolle spielt die Anonymisierung der Daten in unserer Gesellschaft. Es ist daher wichtig, die verschiedenen Anonymisierungsmethoden zu kennen und auf sie einzugehen.
Die Anonymisierungsverfahren lassen sich grundsätzlich in zwei Verfahrensgruppen einteilen, zum einen in Verfahren, die die Informationen des vorhandenen Datenbestandes reduzieren und zum anderen in Verfahren, bei denen die Daten verändert werden. Alle Verfahren, die zu diesen Verfahrensgruppen gehören, weisen jeweils Vor- und Nachteile auf. Es existieren bereits einige Ansätze zur Aufhebung solcher Nachteile. Manche Verbesserungsvorschläge existieren erst in der Theorie und müssen weiter ausgearbeitet werden. Um der Zielsetzung des jeweils vorliegenden Anonymisierungsfalles möglichst nachzukommen bietet es sich an, verschiedene Kriterien zur Auswahl des geeigneten Verfahrens in Erwägung zu ziehen.

A. Anhang - Verfahren zur Informationsreduktion

Zur vereinfachten Veranschaulichung der verschiedenen Anonymisierungsverfahren wird folgendes Anwendungsbeispiel betrachtet:
Gegeben sei ein Datensatz von sechs Beschäftigten eines Münchener Unternehmens mit Angaben zu *Wohnhort, Familienstand, Einkommen* und *Freizeitausgaben*. Dabei sei *Einkommen* die sensible Information.

A.1. Anwendungsbeispiel

WOHNORT	**FAMILIENSTAND**	**EINKOMMEN***	**FREIZEIT-AUSGABEN***
Sendling	ledig	3600	700
Maxvorstadt	ledig	2900	350
Bogenhausen	verheiratet	5700	500
Schwabing-West	ledig	3420	442
Au-Haidhausen	verheiratet	3700	590
Altstadt-Lehel	verheiratet	3300	210

*Euro/Monat

Tabelle A.1: Anwendungsbeispiel

A.2. Entfernung auffälliger Merkmalsträger

WOHNORT	**FAMILIENSTAND**	**EINKOMMEN***	**FREIZEIT-AUSGABEN***
Sendling	ledig	3600	700
Maxvorstadt	ledig	2900	350
~~Bogenhausen~~	~~verheiratet~~	~~5700~~	~~500~~
Schwabing-West	ledig	3420	442
Au-Haidhausen	verheiratet	3700	590
Altstadt-Lehel	verheiratet	3300	210

*Euro/Monat

Tabelle A.2: Entfernung auffälliger Merkmalsträger

A.3. Systematische Einschränkung der Grundgesamtheit

WOHNORT	FAMILIENSTAND	EINKOMMEN*	FREIZEIT-AUSGABEN*
~~Sendling~~	~~ledig~~	~~3600~~	~~700~~
Maxvorstadt	ledig	2900	350
~~Bogenhausen~~	~~verheiratet~~	~~5700~~	~~500~~
Schwabing-West	ledig	3420	442
~~Au-Haidhausen~~	~~verheiratet~~	~~3700~~	~~590~~
Altstadt-Lehel	verheiratet	3300	210

*Euro/Monat

Tabelle A.3: Systematische Einschränkung der Grundgesamtheit

A.4. (Sub-)Stichprobenziehung

WOHNORT	FAMILIENSTAND	EINKOMMEN*	FREIZEIT-AUSGABEN*
Sendling	ledig	3600	700
~~Maxvorstadt~~	~~ledig~~	~~2900~~	~~350~~
Bogenhausen	verheiratet	5700	500
~~Schwabing-West~~	~~ledig~~	~~3420~~	~~442~~
~~Au-Haidhausen~~	~~verheiratet~~	~~3700~~	~~590~~
Altstadt-Lehel	verheiratet	3300	210

WOHNORT	FAMILIENSTAND	EINKOMMEN*	FREIZEIT-AUSGABEN*
Sendling	ledig	3600	700
Bogenhausen	verheiratet	5700	500
Altstadt-Lehel	verheiratet	3300	210

*Euro/Monat

Tabelle A.4: (Sub-)Stichprobenziehung

A.5. Zusammenfassung von Kategorien

WOHNORT	FAMILIENSTAND	EINKOMMEN*	FREIZEIT-AUSGABEN*
~~Sendling~~	ledig	~~3600~~	700
~~Maxvorstadt~~	ledig	~~2900~~	350
~~Bogenhausen~~	verheiratet	~~5700~~	500
~~Schwabing-West~~	ledig	~~3420~~	442
~~Au-Haidhausen~~	verheiratet	~~3700~~	590
~~Altstadt-Lehel~~	verheiratet	~~3300~~	210

WOHNORT	FAMILIENSTAND	EINKOMMEN*	FREIZEIT-AUSGABEN*
München-Süd	ledig	$>$ 3500	700
München-West	ledig	0 - 3500	350
München-Ost	verheiratet	$>$ 3500	500
München-West	ledig	0 - 3500	442
München-Ost	verheiratet	$>$ 3500	590
München-Zentrum	verheiratet	0 - 3500	210

*Euro/Monat

Tabelle A.5: Zusammenfassung von Kategorien

A.6. Entfernung von seltenem, auffälligem Wert

WOHNORT	FAMILIENSTAND	EINKOMMEN*	FREIZEIT-AUSGABEN*
Sendling	ledig	3600	700
Maxvorstadt	ledig	2900	350
Bogenhausen	verheiratet	NA ~~5700~~	500
Schwabing-West	ledig	3420	442
Au-Haidhausen	verheiratet	3700	590
Altstadt-Lehel	verheiratet	3300	210

*Euro/Monat

Tabelle A.6: Entfernung von seltenem, auffälligem Wert

B. Anhang - Verfahren zur Datenveränderung

B.1. Data-Swapping

WOHNORT	FAMILIENSTAND	EINKOMMEN*	FREIZEIT-AUSGABEN*
Sendling	ledig	3600	700
Maxvorstadt	ledig	2900	350
Bogenhausen	verheiratet	5700	500
Schwabing-West	ledig	3420	442
Au-Haidhausen	verheiratet	3700	590
Altstadt-Lehel	verheiratet	3300	210

WOHNORT	FAMILIENSTAND	EINKOMMEN*	FREIZEIT-AUSGABEN*
Sendling	ledig	2900	700
Maxvorstadt	ledig	3420	350
Bogenhausen	verheiratet	3300	500
Schwabing-West	ledig	3600	442
Au-Haidhausen	verheiratet	5700	590
Altstadt-Lehel	verheiratet	3700	210

*Euro/Monat

Tabelle B.1: Data-Swapping

Literatur- und Quellenverzeichnis

Höhne, J. (2010). *Statistik und Wissenschaft - Verfahren zur Anonymisierung von Einzeldaten*. Statistisches Bundesamt, Wiesbaden.

Lenz, R. (2010). *Statistik und Wissenschaft - Methoden der Geheimhaltung wirtschaftsstatistischer Einzeldaten und ihre Schutzwirkung*. Statistisches Bundesamt, Wiesbaden.

Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M. und Vorgrimler, D. (2005). *Statistik und Wissenschaft - Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten*. Statistisches Bundesamt, Wiesbaden.

Bundesministerium der Justiz (2009). *Bundesdatenschutzgesetz (BDSG)*. http://www.gesetze-im-internet.de/bundesrecht/bdsg_1990/gesamt.pdf

Statistisches Bundesamt (2008). *Bundesstatistikgesetz (BStatG)*. https://www.destatis.de/DE/Methoden/Rechtsgrundlagen/Statistikbereiche/Inhalte/010_BStatG.pdf?__blob=publicationFile

Augustin, T. and Wiencierz, A. (2012). *Wirtschafts- und Sozialstatistik Foliensatz 4.4.*. http://www.statistik.lmu.de/institut/ag/agmg/lehre/2011_WiSe/wiso/WiSo_folien_Kap_4.4_20120102.pdf

Rosemann, M. (2007). *Auswirkungen von stochastischer Überlagerung und Mikroaggregation auf die Schätzung linearer und nichtlinearer Modelle* https://www.destatis.de/DE/Publikationen/WirtschaftStatistik/AllgemeinesMethoden/GastbeitraegeHistorisch0407a.pdf?__blob=publicationFile

Statistische Ämter des Bundes und der Länder. *Anonymität von Mikrodaten*. http://www.forschungsdatenzentrum.de/anonymisierung.asp

Statistisches Bundesamt. *Kapitel III. Anonymisierung von Mikrodaten*. https://www.empiwifo.uni-freiburg.de/lehre-teaching-1/winter-term-08-09/materialien-wirtschaftsstatistik III_1_Anonymisierung

Statistisches Bundesamt. *Kapitel V. Anonymisierung von Mikrodaten*. http://www.empiwifo.uni-freiburg.de/lehre-teaching-1/Summer-term-10/Mat-Wirt-Sta/anonym

Wikipedia - Die freie Online-Enzyklopädie (2015, März). *Anonymität*. http://de.wikipedia.org/wiki/Anonymit%C3%A4t

Wikipedia - Die freie Online-Enzyklopädie (2015, März). *Imputation (Statistik)*. http://de.wikipedia.org/wiki/Imputation_(Statistik)

Statisches Amt Mecklenburg-Vorpommern. *Geheimhaltungsverfahren der Zensusergebnisse - SAFE.* http://www.statistik-mv.de/cms2/STAM_prod/STAM/de/zs/Ergebnisse/_Pressekonferenz/SAFE_Versand.pdf

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Seminararbeit selbstständig und lediglich unter Benutzung der angegebenen Quellen und Hilfsmittel verfasst habe.

München, 15.03.2015

Ye Bin Park