Seminararbeit

in Ausgewählte Aspekte der Wirtschafts- und Sozialstatistik

Anonymisierungsverfahren

Pham Son

Aufgabensteller:	Prof. Dr. Thomas Augustin
Betreuer:	Prof. Dr. Thomas Augustin
Abgabedatum:	23.09.2015

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Gesetzliche Grundlage

Im Bundesstatistikgesetz (BStatG) §16 zur Geheimhaltung wird genau beschrieben, welche Anforderungen Daten erfüllen müssen, bevor sie von den Statistischen Ämtern an Dritte weitergegeben werden können.

Bis 1987, als dieser Paragraph im Zuge einer Überarbeitung des BStatG verändert wurde, war es fast unmöglich Daten weiterzugeben. Es wurde im ersten Absatz von §16 verlangt, dass die Deanonymisierung der weitergegebenen Daten unmöglich sein muss. Das heißt Dritten darf es nicht möglich sein, die Daten bestimmten Personen oder Unternehmen zuzuordnen. Dies ist faktisch kaum möglich, da man nicht alle Möglichkeiten ausschließen kann und es theoretisch immer einen weg gibt, die Daten aufzuschlüsseln. (Vgl.: BStatG, 2014, S. 8 / Stat4, 2005, S. 123-124)

Durch die Veränderungen am BStatG 1987 wurde auch der §16 zu Geheimhaltung modifiziert. Der neue Absatz 6 erlaubt nun die Weitergabe von Datensätzen an wissenschaftliche Einrichtungen, ohne dabei die strengen Richtlinien des Absatz 1 zu fordern. Es wird stattdessen nur verlangt, dass die Daten nur unter gro§em Aufwand entschlüsselt werden können. Dabei soll der Aufwand deutlich höher sein, als die gewonnen Informationen für einen potentiellen Angreifer nützlich sind. (Vgl.: BStatG, 2014, S. 8 / Stat4, 2005, S. 123-124)

Aufwand bedeutet in diesem Fall Kosten, Zeit und eventuell auch die Verfügbarkeit anderer Quellen. So kann es sein, dass die Daten an die der Angreifer kommen will, einfacher aus anderen Quellen extrahiert werden können und sich somit ein Deanonymisierungsversuch der Daten nicht lohnt.

1.2 Arten der Anonymität

Grundsätzlich gibt es drei Arten von Anonymität die im folgendem genauer betrachtet werden.

1.2.1 Die absolute Anonymität

Diese Form der verschlüsselten Daten wird im Absatz 1 des §16 im BStatG gefordert. Wie bereits erwähnt darf es nicht möglich sein die Daten einzelnen Personen oder Unternehmen zuzuordnen. Falls dies garantiert ist, können die Datensätze als Public Use Files (PUF) öffentlich zugänglich gemacht werden.

Das Problem besteht darin dass man nie ausschlie§en kann, dass die Daten auch wirklich nicht entschlüsselt werden können. (Vgl.: Anonymität, 2011, S.1)

1.2.2 Die faktische Anonymität

Im Absatz 6 des §16 im BStatG wird verlangt, dass Daten nur unter unverhältnismä§ig gro§en Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können. Falls dies zutrifft, spricht man von faktisch anonymen Daten. Die Weitergabe dieser Daten ist jedoch nur an wissenschaftliche Institute möglich und auch nur zu wissenschaftlichen Zwecken. (Vgl.: Anonymität, 2011, S.1)

Mit dieser Art der Anonymität beschäftigt sich diese Seminararbeit. Das Hauptziel liegt darin, durch Veränderung und Entfernung von Information aus einem originalem Datensatz, Einen faktisch anonymen zu machen. Dabei sollen trotzdem so viele statistische Informationen wie möglich erhalten bleiben.

Es gibt dafür keinen pauschalen Lösungsansatz. Das hei§t dass man jeden Datensatz für sich betrachten muss und dann erst entscheiden kann, welche Methoden angewandt werden, um ihn faktisch anonym zu machen. Wie hoch ist der Aufwand der Deanonymisierung und der Nutzen der gewonnen Information für einen Angreifer? An wen wird der Datensatz weitergegeben? Welches Zusatzwissen kann erlangt werden, dass die Entschlüsselung erleichtert oder attraktiver macht? (Vgl.: Anonymität, 2011, S.1 / Stat16, 2010, S.11)

Diese Fragen muss man sich für jeden Datensatz beantworten und anhand dieser Erkenntnisse einen Lösungsansatz bilden. Dies ist jedoch nicht immer möglich da es auch sein kann, dass sich ein Datensatz nicht für die faktische Anonymisierung eignet. Ein weiteres Problem besteht im Analysepotential der Daten, da diese sonst kaum nützlich sind.

1.2.3 Die formale Anonymität

Bei der formalen Anonymität werden nur die Direkten Identifikatoren wie der Name entfernt. Dadurch bleiben Merkmalsumfang, regionale und fachliche Gliederung erhalten. Diese Daten können jedoch nicht weitergegeben werden. Damit Externe trotzdem diese Daten nutzen können, gibt es die Möglichkeit des Fernrechnens.(Vgl.: Anonymität, 2011, S.1)

Beim Fernrechnen werden Prozeduren an die statistischen Ämter geschickt. Die Syntax dieser Prozeduren führen zum Beispiel statistische Test durch und geben die Ergebnisse dieser Tests zurück. Die Ämter lassen die Syntax durchlaufen und prüfen die Ergebnisse die ausgegeben werden. Falls diese Ergebnisse freigegeben werden, wird den Externen nur das Ergebnis zurückgeschickt.

Kapitel 2

Anonymisierungsverfahren

2.1 Unterscheidung der Methoden

Zuerst haben alle Verfahren gemein, dass sie Veränderungen an den Einzeldaten durchführen. Unterteilen kann man sie in die Kategorien Informationsreduktion und Informationsveränderung. Man versucht zu verhindern, dass Informationen eindeutig zugeordnet werden können. Au§erdem soll der mögliche Informationsgewinn für einen Angreifer reduziert werden, im besten Fall sogar unmöglich sein. Au§erdem wird unterschieden ob die Methode für kategoriale oder für metrische Variablen eingesetzt wird. (Vgl.: Stat4, 2005, S.53-54)

Eine weitere Unterscheidung erfolgt dann, in welchem Umfang Daten verändert werden. Man trennt je nach dem ob nun alle Einzelwerte, Ausprägungen eines Merkmals/ Merkmalsträgers oder auffällige Einzelwerte bearbeitet werden sollen.

2.2 Verfahren zur Informationsreduktion

2.2.1 Merkmalsträgerbezogene Verfahren

Entfernen auffälliger Merkmalsträger: Sollte es besonders auffällige Merkmalsträger im Datensatz geben, werden diese entfernt. Dadurch können sie später nicht identifiziert werden, was aber auch dazu führt, dass Ergebnisse statistischer Analysen und Tests verzerrt bzw. verändert werden. Diese Verfahren wird meist bei den Personen- und Haushaltsdaten verwendet, da Ausrei§er bei Unternehmensdaten meist auch mehr Einfluss auf die Analysen haben. (Vgl.: Stat4, 2005, S.55-56)

Systematische Einschränkung der Grundgesamtheit: Hier wird statt einem Merkmalsträger, eine ganze Gruppe von Merkmalsträgern entfernt, die gewisse Kriterien erfüllen. Wie diese Gruppen gewählt werden, hängt von dem Ziel der Analyse ab. Sollen zum Beispiel Umsatzzahlen analysiert werden, wobei man eher auf die kleinen Unternehmen schau will, so kann man z.B die Gruppe der Gro§unternehmen entfernen. (Vgl.: Stat4, 2005, S.56)

Substichprobenziehung: Aus der Grundgesamtheit der Daten wird eine Stichprobe gezogen. Dadurch wird die Zuordnungsmöglichkeit verringert. Der Angreifer wei§nicht ob sich der Merkmalsträger den er sucht, überhaupt in der Stichprobe befindet. Dieses Verfahren eignet sich besser für Haushaltsdaten, da hier die Grundgesamtheit grö§er ist als bei Unternehmensdaten. Au§erdem sind die Verteilungen bei Unternehmensdaten in der Regel schiefer. (Vgl.: Stat4, 2005, S.56-57)

2.2.2 Merkmalsbezogene Verfahren

Entfernen/Ersetzen/Zusammenfassen von Merkmalen: Merkmalsentfernung geschieht zum Beispiel bei direkten Identifikatoren. Man kann aber auch mehrere Merkmale entfernen und durch ein neues Merkmal ersetzen. So kann man beispielsweise Verhältniszahlen, Summen oder Scores bilden. Das Ersetzen/Zusammenfassen von Merkmalen funktioniert aber nur bei metrischen Variablen. (Vgl.: Stat4, 2005, S.57-58)

Vergröberung von Merkmalen: Dies kann geschehen wenn man metrische Merkmale gruppiert. Als Beispiel kann man statt der Einwohnerzahl einer Stadt zu verwenden, diese in verschiedene Stadtgrö§en unterteilen. Man kann auch einfach metrische Variablen runden. Statt der genauen Einwohnerzahl nimmt man in unserem Beispiel die Einwohnerzahl in Tausenderbeträgen. Alternativ kann man auch existierende Kategorien zu einer kategorialen Variable zusammenfassen. Die Vergröberung der Merkmale hat den Effekt, dass es weniger mögliche Merkmalskombinationen gibt. Dadurch überschneiden sich auch mehr Merkmalsträger was die Zuordnungsmöglichkeit erschwert. (Vgl.: Stat4, 2005, S.58-59)

2.2.3 Ausprägungsbezogene Verfahren

Da man hier nur Einzelwerte behandelt, gibt es in der Regel nur die Möglichkeit auffällige Einzelwerte zu entfernen. Der Vorteil besteht darin, dass keine ganzen Merkmale verändert oder angepasst werden müssen. (Vgl.: Stat4, 2005, S.59-60)

2.3 Verfahren zur Informationsveränderung

Hier unterscheiden wir zwischen Methoden die auf kategoriale oder metrische Variablen anwendbar sind. Bei der Informationsreduktion ist das meist nicht nötig da beim Entfernen nicht darauf geachtet werden muss.

2.3.1 Verfahren für kategoriale Variablen

Swapping: Man vertauscht Werte zwischen Merkmalsträgern getrennt nach Variablen. Man kann dies für alle Merkmalsträger vornehmen oder auch nur für bestimmte Gruppen. Dadurch erschwert man die Zuordnungsmöglichkeit für potentielle Angreifer, jedoch verändert man sehr stark die Daten. (Vgl.: Stat4, 2005, S.61)

Post-Randomisierung: Bei dieser Methode werden die Werte nur zu festgelegten Wahrscheinlichkeiten verändert. (Vgl.: Stat4, 2005, S.61)

2.3.2 Verfahren für metrische Variablen

Swapping: Diese Methode funktioniert wie bei der kategorialen Variante. Hier kann jedoch auch die Rangstatistiken weitestgehend erhalten bleiben. Man kann beispielsweise die Merkmalsausprägungen sortieren und grenzt das Swapping für jeden Merkmalsträger auf einen Bereich um ihn herum ein. Auch bivariate Verteilungen können erhalten bleiben und es gibt noch andere Möglichkeiten das Analysepotential zu erhöhen. (Vgl.: Stat4, 2005, S.65)

Imputationsverfahren: Die ursprüngliche Verwendung von Imputationsverfahren lag darin, fehlende Werte mit Hilfe von Regressionsmodellen zu schätzen und somit unvollständige Datensätze zu vervollständigen. Bei der Anonymisierung ersetzt das Verfahren stattdessen aber auffällige Werte oder schätzt sogar ganze Merkmale neu. (Vgl.: Stat4, 2005, S.66)

Stochastische überlagerung: Diese Verfahrensgruppe kann in zwei gro§e Felder unterteilt werden. Zum einen die additiven und zum anderen die multiplikativen Methoden. Die Verfahren addieren/multiplizieren die Merkmalswerte mit einem Zufallsfehler, der zum Beispiel mit der Normalverteilung ermittelt wird. Dies ist jedoch ein sehr umfangreiches Thema, was in dieser Seminararbeit auch nicht weiter vertieft wird. (Vgl.: Stat4, 2005, S.67-80)

Simulationsverfahren: Wie der Name bereits vermuten lässt, erzeugt man bei diesem Verfahren einen neuen Datensatz, der eine ähnliche empirische Verteilung aufweisen soll wie das Original. Möglich soll dies durch eine Schätzung der Kerndichte eines Datensatzes sein. Diese Schätzung ist aber nur für niedrigdimensionale Datensätze möglich. (Vgl.: Stat4, 2005, S.86-92)

Ausprägungsbezogene Verfahren: Um einzelne Merkmalsträger besser zu schützen kann man diese Klonen oder Zerlegen. Dadurch kommen sie öfter bzw. gar nicht im Datensatz vor. Ein weiteres Mittel ist die Beschränkung des Wertebereichs. Die Ausrei§er werden auf das Maximum bzw. Minimum des Wertebereichs gesetzt. Alternativ zu Minimum und Maximum kann man auch das arithmetische Mittel aller Werte, die über/unter dem Bereich liegen, verwenden. (Vgl.: Stat4, 2005, S.93)

Kapitel 3

Mikroaggregation

3.1 Grundidee

Die Mikroaggregation ist ein datenveränderndes Verfahren für metrische Variablen. Man teilt den Datensatz in Gruppen ein. Die Werte aller Merkmalsträger in der Gruppe werden dann durch das arithmetische Mittel der Gruppe ersetzt.

Dabei gibt es verschiedene Ansätze. Deterministische Methoden fassen ähnliche Werte zusammen, während stochastische die Gruppen zufällig bilden. Die Verfahren unterscheiden sich hauptsächlich in der Bildung der Gruppen und ob nun Gruppen für mehrere oder einzelne Merkmale gelten.

Es muss jedoch immer eine Gruppengrö§e von mindestens drei Merkmalsträgern sein. Meist sind es nicht mehr als fünf da man eine Sechsergruppe durch zwei Dreiergruppen ersetzen kann. Man braucht mindestens drei, weil man bei zwei sonst bei Kenntnis der Werte eines Merkmalsträgers, die Werte des Anderen einfach errechnen kann. (Vgl.: Stat16, 2010, S.45-47)

Durch das Mikroaggregationsverfahren existiert jeder Wert mindestens drei mal, somit ist es kaum möglich Merkmalsträger eindeutig zuzuordnen. Durch die Verwendung der Mittelwerte wird auch der mögliche Informationsgewinn reduziert.

Im Rahmen dieser Seminararbeit wurde auch ein R-Skript erstellt. Dieses enthält verschiedene Funktionen für einige Mikkroaggregationsverfahren, die im folgenden genannt werden.

3.2 Deterministische Mikroaggregation

3.2.1 Gemeinsame Mikroaggregation

Mikroaggregation nach einer dominierenden Variablen: Man ermittelt eine dominierende Variable und sortiert den Datensatz absteigend danach. Im Anschluss werden drei aneinanderliegende Werte zu einer Gruppe zusammengefasst und ihre Werte durch das arithmetische Mittel ersetzt. Die Gruppen bleiben dabei für jedes Merkmal gleich. (Vgl.: Stat4, 2005, S.82)

Für die Funktion im R-Skript wurde diese Variable ermittelt, in dem eine Korrelationsmatrix erstellt wird. Das Merkmal, welches als Summe aller absoluten Werte einer Zeile/Spalte den höchsten Wert hat, wurde als dominierende Variable deklariert und der Datensatz wurde nach ihr sortiert. Eine Alternative wäre die Variable zu verwenden, die die höchste Anzahl an absoluten Werten in der Korrelationsmatrix aufweist, welche eine bestimmte Grenze überschreiten.

Mikroaggregation nach einer Hilfsvariablen: Man kann auch statt einer dominierenden Variable, versuchen eine Hilfsvariable zu erstellen. Diese sollte eine möglichst hohe Korrelation mit den anderen Variablen aufweisen und kann zum Beispiel ein Score oder eine durch Transformation gebildete Variable sein. Es wurde zu dieser Methode keine Funktion geschrieben, da diese Hilfsvariable für jeden Datensatz unterschiedlich sein muss und au§er diesem Aspekt unterscheidet sich diese Methode nicht von der Mikkroaggregation nach einer dominierenden Variable. (Vgl.: Stat4, 2005, S.82)

Mikroaggregation nach der euklidischen Distanz: Man ermittelt für alle Merkmalsträger die euklidische Distanz zu allen Anderen. Man Nimmt nun zwei Merkmalsträger mit der höchsten euklidischen Distanz. Zu diesen nimmt man je zwei MerkmalstrÃger mit der jeweils geringsten Distanz und bildet somit zwei Dreiergruppe. Dies wiederholt man so oft, bis alle Merkmalsträger gruppiert sind. (Vgl.: Stat4, 2005, S.82)

Die euklidische Distanz d(x,y) wird wie folgt berechnet. Seien x und y zwei gleichlange Vektoren, dann gilt:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.1}$$

Für die Funktion im R-Skript wurde eine externe Prozedur für die euklidische Distanz erstellt. Diese erzeugt eine Matrix mit der Breite und Länge, die

gleich der Anzahl der Merkmalsträger ist. In jeder Zelle dieser Matrix steht die euklidische Distanz zwischen dem x-ten und dem y-ten Merkmalsträger im Datensatz. x und y sind dabei die Zeile bzw. Spalte in der sich diese Zelle befindet. Somit ergibt sich eine Diagonale aus Nullen von links oben nach links unten. In der Hauptfunktion wird mit Hilfe dieser Matrix das Verfahren durchgeführt.

3.2.2 Getrennte Mikroaggregation

Bei dieser Methode der Mikroaggregation sortiert man für eine Variable den Datensatz. Anschlie§end bildet man für diese Variable Gruppen und führt die Mikroaggregation durch. Dies wiederholt man für alle metrischen Variablen. Ein Problem dieser Methode liegt darin, dass die Werte sich kaum von den Originalwerten unterscheiden, wenn die Datenpunkte sich in einem Bereich häufen. Au§erdem kann ein potentieller Angreifer den Bereich, in dem die Originalwerte liegen, einfach einschränken. Dieser liegt zwischen den Mittelwerten der benachbarten Gruppen. (Vgl.: Stat4, 2005, S.82)

Im R-Skript wurde der Datensatz zuerst nach der ersten metrischen Variable sortiert und die Mikroaggregation für diese durchgeführt. Mit einer Schleife wurde der Vorgang für alle weiteren metrischen Variablen wiederholt.

3.2.3 Gruppierte Mikroaggregation

Bei dieser Methode werden stark korrelierte Variablen zu Gruppen zusammengefasst, für die eine gemeinsame Mikroaggregation innerhalb der Gruppe durchgeführt wird. Dies wird für alle Gruppen wiederholt. (Vgl.: Stat4, 2005, S.83)
//
Dieses Verfahren wurde nicht im R-Skript umgesetzt, da nicht ganz klar war, nach welchem Kriterium die Variablengruppen gebildet werden sollen. Die Umsetzung der nachfolgenden Prozedur wäre eine Kombination der bereits geschriebenen Funktionen zur getrennten und gemeinsamen Mikroaggregation.

3.3 Stochastische Mikroaggregation

Zufällige Mikroaggregation: Statt die Variablen nach dem Abstand zueinander zu sortieren, werden hier zufällig Merkmalsträger zusammengefasst. Dies ist sowohl als getrennte oder gemeinsame Mikroaggregation möglich.

Beim R-Skript wurden die Funktionen realisiert, in dem man den Datensatz

zufällig sortiert. Der Rest der Prozedur ist vergleichbar mit den deterministischen Varianten.

Bootstrap Mikroaggregation: Bei dieser Methode werden für jeden Merkmalsträger im Datensatz, zwei Andere mit Zurücklegen gezogen. Diese Drei bilden eine Gruppe und das arithmetische Mittel ersetzt den Wert des Merkmalsträgers im Originaldatensatz. Auch diese Methode kann als eine getrennte oder gemeinsame Mikroaggregation durchgeführt werden. (Vgl.: Stat4, 2005, S.84)

3.4 Vor- und Nachteile der Mikroaggregation

Bei der Mikkroaggregation werden die arithmetischen Mittel beibehalten, jedoch wird die Varianz verringert. Dies führt zu verzerrten Ergebnissen bei vielen statistischen Tests, da die Varianz fast immer eine Rolle spielt. (Vgl.: Stat4, 2005, S.85-86)

Ein Lösungsansatz wäre die Bildung von Vierergruppen. Man ersetzt die Werte der Merkmalsträger in der Gruppe nicht durch das arithmetische Mittel. Stattdessen erhalten zwei Merkmalsträger das arithmetische Mittel plus die Standardabweichung. Die zwei Anderen erhalten das arithmetische Mittel minus die Standardabweichung. Auf diese Art und Weise kann man auch die Varianz erhalten und verbessert somit das Analysepotential im Idealfall. (Vgl.: Stat4, 2005, S.83)

3.5 Allgemeines Mikroaggregationsverfahren im R-Skript

Bei der Vorstellung der verschiedenen Mikroaggregationsverfahren wurde bereits darauf eingegangen wie Einige umgesetzt wurden. Hier wird nochmal der allgemeine Aufbau der Funktionen erläutert.

Die Funktionen nehmen als Eingabe einen Datensatz an sowie die gewünschte Gruppengrö§e. Zuerst werden die metrischen Variablen des Datensatzes von den kategorialen getrennt. Dann wird die Anzahl der zu bildenden Gruppen festgestellt und eine Restgruppe erstellt, falls dies nötig sein sollte. Als nächstes wird der Datensatz sortiert, wobei die Art der Sortierung von der Mikkroaggregationsmethode abhängt.

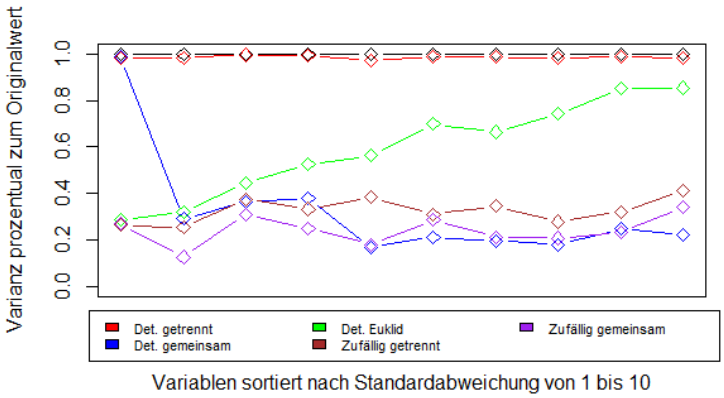

Abbildung 3.1: Plot der die relativen Varianzen zwischen den verschiedenen Verfahren vergleicht

Im folgenden durchläuft eine Schleife eine Spalte und ersetzt die Werte der Gruppen durch ihre arithmetischen Mittel. Diese liegt in einer Schleife die diesen Spaltendruchlauf für alle Spalten durchführt. Bei getrennter Mikroaggregation wird nach jedem Spaltendurchlauf neu sortiert. Das so entstandene Dataframe wird mit den row.names ausgegeben.

3.6 Simulation

Zusätzlich zu den Funktionen für die Mikroaggregation wurde im R-Skript eine Simulation durchgeführt. Dazu wurde ein Datensatz mit 100 Beobachtungen erstellt, bestehend aus zehn metrischen und einem kategorialen Merkmal. Die metrischen Variablen sind dabei normalverteilt mit einer Standardabweichung von 1 bis 10 sowie einem Mittel von 0 bis 9.

Auf diesen Datensatz wurden dann die fünf Funktionen, die erstellt wurden, angewandt. Es wurden die Mittelwerte verglichen. Diese sollten gleich sein, da sonst die Methode fehlerhaft bei der Gruppierung vorgeht. Au§erdem wurden die Varianzen verglichen.

Bei dem getrennten Deterministischen Verfahren sind die Varianzen relativ nah bei den Originalwerten. Beim deterministischen Ansatz mit Sortierung nach der dominierenden Variable, sieht man deutlich, dass die erste Variable dominant war. Die Varianz für die anderen Variablen weicht sehr stark von

den Originalwerten ab. Bei der Methode mit dem euklidischen Abstand sieht man, dass die Methode für Variablen mit höherer Standardabweichung besser funktioniert. Ein Grund dafür kann die Formel sein, mit der man den Abstand berechnet. Variablen mit höheren Varianzen fallen bei der Abstandsbestimmung stärker ins Gewicht. Bei den zufälligen Verfahren sind die Ergebnisse wie erwartet, da die Sortierung zufällig passiert fallen die Ergebnisse in der Regel unterschiedlich aus.

Kapitel 4

Das SAFE-Verfahren

4.1 Grundidee

Das SAFE-Verfahren ist eine Kombination aus verschiedenen Verfahren. Das Ziel dieser Methode besteht darin folgende Probleme zu eliminieren:

Fallzahlproblem: Dieses Problem tritt auf wenn es für Ausprägungen in den Daten nur ein oder zwei Fälle gibt. So kann es passieren, dass es beispielsweise in einer Stadt nur einen Betrieb einer bestimmten Grö§enordnung in einer Branche existiert. Durch diese Information kann man aus dieses Unternehmen leicht zuordnen. (Vgl.: SAFE, 2003, S.97 / Stat16, 2010, S.77-78)

Randsummenproblem: Es kann vorkommen, dass einige Merkmalskombinationen nur eine Ausprägung besitzen. Hat man nun die Information, dass die Ausprägungen von drei Merkmalen für Teile der Merkmalsträger gleich sind, so kann man mit Hilfe von zwei Merkmalen auf das dritte schlie§en. Zum Beispiel sind alle Schüler einer Klasse, die älter als 16 sind, bei einem Test durchgefallen. Es reicht zu wissen ob der Schüler in die Klasse geht und über 16 Jahre alt ist, um zu wissen, dass er durchgefallen ist. (Vgl.: SAFE, 2003, S.97 / Stat16, 2010, S.77-78)

Dominanzproblem: Es entsteht wenn die Summe eines metrischen Merkmals zu gro§en Teilen durch die Werte von einem oder zwei Merkmalsträgern gebildet wird. Ob ein Dominanzproblem vorliegt, hängt davon ab welche Regel man verwendet, um dieses festzustellen. (Vgl.: SAFE, 2003, S.97-98 / Stat16, 2010, S.77-78)

Matching: Ein weiteres Problem das entsteht, ist das Matching. Dabei versuchen potentielle Angreifer mit Hilfe externer Daten, Merkmalsträger zu identifizieren. Dadurch erhalten sie Informationen zu diesem Merkmalsträger, die im Datensatz enthalten sind. (Vgl.: SAFE, 2003, S.98 / Stat16, 2010, S.77-78)

4.2 Die SAFE-Methode

Die Lösung die SAFE bietet, ist eine Kombination aus Mikroaggregation und Einzeldatenanonymisierung. Durch die Mikroaggregation wird das Fallzahlproblem eliminiert, da es für jede Merkmalsausprägung mindestens drei Fälle gibt.Dadurch ist auch das Matching nicht mehr sinnvoll für den Angreifer. Die Randsummen werden durch die Mikroaggregation verändert bzw. das Problem entsteht erst dadurch. Somit kann ein Angreifer sich nicht sicher sein ab das Randsummenproblem nicht erzeugt wurde. (Vgl.: SAFE, 2003, S.98-99 / Stat16, 2010, S.78)

Das Dominanzproblem wird mit der Einzeldatenanonymisierung eliminiert. Man legt Unzulässigkeitsintervalle fest und bestimmt somit alle Felder die ein Dominanzproblem aufweisen. Die Werte dieser Felder werden dann verändert. (Vgl.: SAFE, 2003, S.98-99 / Stat16, 2010, S.78)

4.2.1 Begriffserklärung

Um die Methodik zu erklären, müssen zuerst die Begriffe erklärt werden (Vgl.: SAFE, 2003, S.98-99 / Stat16, 2010, S.78-82):

X^0 : Ist der originale Datenbestand der metrischen Werte. Die Zeilen entsprechen den Merkmalsträgern und die Spalten den Merkmalen. Es gilt dass n die Anzahl der Merkmalsträger ist.

Z_j^0 : Dies sind die Zuordnungsmatrizen der kategorialen Variablen aus dem originalen Datenbestand. Dabei steht j für die Anzahl der kategorialen Merkmale. Die Matrix besteht aus n Zeilen. Die Anzahl der Spalten hängt von der Anzahl der Merkmalsausprägungen einer jeden kategorialen Variable ab. Jede Zeile kann dabei nur eine 1 enthalten und zwar bei der Merkmalsausprägung die der Merkmalsträger hat. Die restlichen Felder der Zeile bestehen aus Nullen.

Diese Matrizen können auch mit dem Kroneckerprodukt kombiniert werden: $Z_{i,jl} = Z_{i,j} \otimes Z_{i,l}$
Die so erhaltene Matrix hat n Zeilen. Das Produkt der Spaltenanzahl beider

$$\mathrm{X}^0 = \begin{pmatrix} 1 & 2 & 0 \\ 3 & 1 & 4 \\ 0 & 7 & 2 \end{pmatrix} \qquad \mathrm{Z}_1^0 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mathrm{T}_1^0 = \begin{pmatrix} 4 & 3 & 4 \\ 0 & 7 & 2 \end{pmatrix} \qquad \mathrm{A}_1^0 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

Abbildung 4.1: Beispiele für die Matrizen

ursprünglichen Matrizen ist die Spaltenanzahl der Neuen. Auch hier gibt es in jeder Zeile nur eine 1 als Ausprägung.

$T_j^0 = (Z_j^0)X^0$**:** Diese Matrix enthält die Summen aller metrischen Werte, die durch die Spalten der Zuordnungsmatrix definiert sind. Dabei bleibt die Spaltenreihenfolge wie im Originaldatensatz erhalten.

$A_j^0 = (Z_j^0)Z_j^0$**:** Hier wird auf der Hauptdiagonale der resultierenden Matrix, die Anzahl der jeweiligen Merkmalsausprägung angegeben. Somit gibt die Zahl in der i-ten Spalte und Zeile der Matrix an, wie oft die i-te Merkmalsausprägung vorkommt.

X^a **und** Z_j^a**:** Sie sind gleich aufgebaut wie X^0 und Z_j^0, jedoch haben sie maximal $n/3$ Zeilen.

H^a**:** Dies ist eine Diagonalmatrix mit ganzen Zahlen grö§er drei. Die Dimension der Matrix entspricht der Anzahl der Zeilen in x^A. Die Werte auf der Diagonal geben an wie oft die Zeile aus x^A existiert. Dadurch werden die folgenden Matrizen anders gebildet.

$T_j^a = (Z_j^a)H^a X^a$

$A_j^a = (Z_j^a)H^a Z_j^a$

G_j^0**:** Dies ist eine Matrix mit denselben Dimensionen wie T_j^0. Sie gibt an, ob in dem Tabellenfeld ein Dominanzproblem vorhanden ist oder nicht.

$$F_H = \min\left(\max_{j\in(1,t)}\left(\max_{i\in(1,s_j)} \left(\left| a^a_{j,i,i} - a^o_{j,i,i} \right| \right) \right) \right)$$

$$F_T = \sum_{j=1}^{t}\sum_{i=1}^{s_j}\sum_{l=1}^{m} \left| t^a_{j,i,l} - t^o_{j,i,l} \right| (1 - g^o_{j,i,l})$$

Abbildung 4.2: Formel der Minimierungsaufgaben aus Stat16, 2010, S.81)

4.2.2 Lösungsansatz

Die Idee ist nach X^a, Z^a_j und H^a zu suchen, damit T^a_j und A^a_j möglichst ähnlich zum Original sind. Das hei§t wir haben zwei Minimierungsaufgaben vor uns, die in Abbildung 4.2 dargestellt sind.

Es müssen somit die metrischen Werte (X^a) bestimmt werden sowie die kategorialen Zuordnungsmatrizen (Z^a_j) für höchstens $n/3$ Werte. Au§erdem braucht man noch eine Häufigkeitstabelle (H^a). Dies stellt oft ein Problem da, weil die Dimensionen der originalen Daten sehr gro§sind. Da au§erdem nach den zwei Kriterien F_H und F_T minimiert werden muss, steigt der zeitliche Aufwand sehr stark mit jeder Dimension. Es ist somit nur möglich mit der SAFE-Methode eine Annäherung an das optimale Ergebnis zu erreichen. (Vgl.: Stat16, 2010, S.83)

Dies gelingt wie folgt (Vgl.: SAFE, 2003, S.101 / Stat16, 2010, S.88):

1: Man bearbeitet zuerst die kategorialen Variablen. Man bestimmt Z^j und H^a sodass F_H minimal ist.

2: Die Zeilen von X^0 werden zu den Z^a_j zugeordnet. Das Ziel dabei ist den Abstand der Zeilen von Z^a_j und Z^0_j zu minimieren.

3: Man verändert nun X^0 sodass alle Kriterien der Geheimhaltung erfüllt werden. Man bestimmt alle Lösungen die zulässig sind.

4: X^0 wird weiter verändert um F_T zu minimieren. Dabei werden nur Lösungen zugelassen, die die Kriterien von 3. erfüllen.

5: Anschlie§end wird eine Mikroaggregation durchgeführt.

Kapitel 5

Analysepotential

Bis hierhin wurden nur die Verfahren beschrieben, welche einen Datensatz anonymisieren sollen. Dabei ist es aber auch wichtig, dass die Daten noch geeignet für Analysen sind.

Dabei gibt es drei wesentliche Ansätze, um das Analysepotential zu messen:

Der ma§zahlorientierte Ansatz: Bei diesem Ansatz werden Ma§zahlen wie Mittelwerte, Mediane, Streuungsma§e oder Korrelationsma§e verglichen. Dabei wurden verschiedene Scores entwickelt, die aussagen, wie sehr das Analysepotential durch die Anonymisierungsverfahren reduziert wurde. Probleme bei diesem Ansatz bestehend darin, dass es keine Gewichtung der einzelnen Variablen gibt und es auch keine allgemeingültige Grenze gibt ab der ein Datensatz zu wenig Analysepotential hat. Man hat mit dieser Methode höchstens einen Ansatz, um Ergebnisse verschiedener Verfahren miteinander zu vergleichen. (Vgl.: Stat4, 2005, S.153-157)

Der anwendungsorientierte Ansatz: Mit Hilfe eines Katalogs an Analysen, vergleicht man deren Ergebnisse vor den Anonymisierungsverfahren und danach. Der Katalog soll verschiedene Arten von Analysen enthalten und möglichst vielfältig sein. Problematisch ist, dass die Ergebnisse der einzelnen Analysen nur beispielhaft sind und somit nicht für alle Analysen repräsentativ.(Vgl.: Stat4, 2005, S.157-158)

Der theorieorientierte Ansatz: Man befasst sich bei diesem Ansatz nicht nur mit den Ergebnissen eines Verfahrens, sondern man versucht dieses zu korrigieren. Ein Beispiel aus der Mikkroaggregation ist die Korrektur zum Erhalt der Varianz aus Kapitel 3.4. Man bewertet das Analysepotential danach ob

die statistischen Ma§e erwartungstreu bleiben oder das Verfahren so angepasst werden kann, damit dies der Fall ist. (Vgl.: Stat4, 2005, S.158-159)

Jedoch sind die korrigierten Methoden nicht unbedingt fehlerfrei und kÃ¶nnen trotzdem verzerrte Schätzer erzeugen. Man muss jedoch bei Erhalt von komplexeren Sachverhalten auch deutlich mehr Modelle und Ma§zahlen untersuchen. Ein weiteres Problem ist, dass diese Methode nicht auf alle Anonymisierungsverfahren anwendbar sind. (Vgl.: Stat4, 2005, S.158-159)

Kapitel 6

Schlusswort

Festzuhalten bleibt, dass es nicht den pauschalen Weg beim Anonymisieren von Datensätzen gibt. Jeder Datensatz muss individuell geprüft werden. Auch die Fragestellung spielt eine gro§e Rolle bei der Auswahl der Verfahren. Einige Methoden erschweren es potentiellen Angreifern mehr als Andere, doch sie verringern oft auch das Analysepotential deutlich stärker. Hier gilt, welche Ma§zahlen und welche Variablen sind überhaupt wichtig für die Fragestellung? So kann das Analysepotential bei einem Verfahren schlechter aussehen. Doch wenn man genauer hinschaut stellt sich heraus, dass nur irrelevante Ma§zahlen verzerrt werden.

Es ist möglich für gleich strukturierte Datensätze, die beispielsweise jährlich erhoben werden, eine optimale Kombination aus Verfahren zu erstellen. Doch dies kann sich schnell Ändern wenn neue Fragen beantwortet werden sollen oder neue Verfahren entwickelt werden.

Abbildungsverzeichnis

Literaturverzeichnis

[BStatG, 2014] "**Gesetz über die Statistik für Bundeszwecke BstatG**'", Bundesministerium der Justiz, juris GmbH, *https://www.destatis.de*

[Abmd, 2014] "**Anonymising business micro data - results of a German project**'", Rainer Lenz, Martin Rosemann, Daniel Vorgrimler, Roland Sturm, *https://www.destatis.de*

[Anonymität, 2011] "**Datenzugang — Anonymität von Mikrodaten**'", Statistische Ämter des Bundes und der Länder, *http://www.forschungsdatenzentrum.de/anonymisierung.asp*

[Stat4, 2005] "**Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten - Statistik und Wissenschaft Band 4**'", Gerd Ronning, Roland Sturm, Jörg Höhne, Rainer Lenz,Martin Rosemann, Michael Scheffer, Daniel Vorgrimler - Statistisches Bundesamt

[Stat16, 2010] "**Verfahren zur Anonymisierung von Einzeldaten - Statistik und Wissenschaft Band 16**'", Jörg Höhne - Statistisches Bundesamt

[SAFE, 2003] "**SAFE - ein Verfahren zur Geheimhaltung und Anonymisierung statistischer Einzelangaben**'", Jörg Höhne, Berliner Statistik Monatsschrift 3/03, *https://www.statistik-berlin-brandenburg.de/publikationen/aufsaetze/2003/MS-BE200303-01:pdf*