Datenfusion: Theoretische Implikationen und praktische Umsetzung

***Dr. Florian Meinfelder**, geb. 1975, Studium der Sozialwissenschaften an der Friedrich-Alexander-Universität Erlangen-Nürnberg sowie Statistik an der University of St. Andrews (1996 bis 2002). Von 2002 bis 2010 Mitarbeiter bei der GfK SE, Bereich Fernsehforschung. Promotion 2009 zum Thema Multiple Imputation für diskrete Daten. Seit April 2010 Akademischer Rat am Lehrstuhl für Statistik und Ökonometrie der Otto-Friedrich-Universität Bamberg.*

Datenfusion bezeichnet ein spezifisches Datenausfallmuster das entsteht, wenn (mindestens) zwei unabhängig entstandene Datenquellen „übereinandergestapelt" werden, so dass es eine Gruppe an Variablen (X) gibt, die in beiden Datenquellen vorkommen sowie eine Gruppe an Variablen (Y), die nur in der ersten Datenquelle und eine Gruppe an Variablen (Z), die nur in der zweiten Datenquelle existieren. Zudem bezieht sich das generelle Analyseziel auf die gemeinsame Verteilung von Y und Z – den Variablen, die nicht gemeinsam beobachtet wurden. Häufig wird Datenfusion fälschlicherweise gleichgesetzt mit *Nearest-Neighbour*-Algorithmen – der in der englischsprachigen Literatur häufig für „Datenfusion" verwendete Begriff *Statistical Matching* hat dazu beigetragen; allerdings handelt es sich hierbei nur um eine Gruppe zur Verfügung stehender Methoden, um die oben genannten Analyseziele zu erreichen. Dieser Vortrag beschreibt die einzelnen Teilschritte in der praktischen Umsetzung einer Datenfusion und vermittelt einen Bezug zu den entsprechenden Konzepten aus der statistischen Missing-Data Literatur.

Die zentrale Annahme einer Datenfusion ist die so genannte *Conditional Independence Assumption*, die besagt, dass Y und Z gegeben die gemeinsamen Merkmale X unabhängig sein müssen. Diese Annahme ist (so gut wie immer) nicht testbar, weswegen sich die Evaluierung einer Datenfusion in der Regel auf den Verteilungserhalt von Y|X beziehungsweise Z|X vor und nach Fusion beschränkt. Es ist jedoch möglich, Grenzen für die Zusammenhänge zwischen Y und Z auf Grund deren jeweiliger Zusammenhänge zu den gemeinsamen Merkmalen X zu bestimmen. Während es zur Evaluierung der Datenfusion einige Beiträge in der Statistischen Literatur gibt, werden geeignete Methoden zur Durchführung eher selten behandelt.

Dieser Vortrag behandelt das Thema „Datenfusion" als imaginäres Projekt, indem in chronologischer Reihenfolge alle Teilschritte beschrieben werden. Hierbei sind zwei Situationen zu unterscheiden:

a) Die beteiligten Studien existieren bereits
b) Eine der beiden Studie ist noch in der Konzeptionsphase

In letzterem Falle kann noch Einfluss auf die Abfrage der gemeinsamen Variablen X genommen werden. Fall a) ist der Regelfall und zieht einen sehr arbeitsaufwendigen Arbeitsschritt nach sich, der in Publikationen – zumindest im Kontext „Datenfusion" – bislang weitgehend ignoriert worden ist: Je nach Umfang der Studien ist der Aufwand der Identifizierung gemeinsamer Merkmale nicht zu unterschätzen. Und selbst wenn Variablen identifiziert wurden, deren Beschreibung scheinbar dasselbe misst, besteht dennoch die Möglichkeit sehr unterschiedlicher Kategorisierungen. Dies kann man sich am Beispiel „persönliches Netto-Einkommen" vor Augen führen, dass z.B. in Studie A offen und in Studie B in Klassen abgefragt wird. In diesem Fall ist die Angleichung der Information sehr einfach.[1] In anderen Fällen ist eine Angleichung der Information schwieriger, z.B. wenn in beiden Studien gefragt wird, wie viel Sport man treibt und die Studien unterschiedlich viele Kategorien hierfür ausweisen. Diesen Prozess der Informationsangleichung nennen wir „Harmonisierung" der gemeinsamen Merkmale.

Der nächste Projektschritt bezieht sich – wie alle nachfolgenden auch – bereits auf die späteren Analysen: Die Auswahl der Fusionsmethode. Neben den bereits erwähnten Nearest-Neighbour-Algorithmen können auch rein verteilungs- bzw. modellbasierte Algorithmen zum Einsatz kommen. Zudem sollte feststehen, ob auf den fusionierten Daten nur deskriptiv oder auch induktiv gearbeitet werden soll. In letzterem Falle wird ein Verfahren benötigt, dass die durch die Ergänzung (die Fusion) induzierte Unsicherheit korrekt abbildet, wie etwa die ‚Multiple Imputation' (MI), bei der fehlende Information m > 1-mal ergänzt wird. „Zeilenweise" Nearest-

[1] Oder auch nicht: Eine ambitionierte Variante wäre tatsächlich, das klassierte Einkommen zu Einzelbeobachtungen aufzulösen.

Neighbour-Verfahren werden in Fusionsszenarien häufig verwendet, da bei einer simultanen Übertragung mehrerer Variablen gewährleistet ist, dass der Vektor der Ausprägungen in sich konsistent bleibt. Um zudem sicherzustellen, dass die „gespendete“ Information des „Donors“ zu der des „empfangenden“ „Rezipienten“ konsistent ist, kann das Nearest-Neighbour-Matching in Segmenten (Unterstichproben) vollzogen werden. Ein weiterer Vorteil dieser Nearest-Neighbour-Verfahren in der praktischen Umsetzung ist, dass die Fusion mit verdichteten Hilfsvariablen durchgeführt werden kann (wenn die eigentlichen spezifischen Variablen sehr zahlreich und/oder komplex sind) und man anschließend als Ergebnis eine Paarliste mit den Identifikatoren der Studien erhält, die es theoretisch erlaubt, weitere Information zu übertragen, die nur indirekt in das Modell eingeflossen ist.

Bei der Fusionsdurchführung selbst ist im Grunde nur zu unterscheiden, ob Y (bzw. Z) oder beide Gruppen Y und Z der spezifischen Merkmale ergänzt werden, ob die Fusion mit einem MI-Verfahren durchgeführt wurde und ob ein (zeilenweises) Nearest-Neighbour-Verfahren zum Einsatz kam, da diese Faktoren auf die Datenhaltung Einfluss haben (das konkrete Ergebnis einer Fusion mit Nearest-Neighbour-Verfahren, bei der multipel in beide Richtungen ergänzt wird, wären m×2 Paarlisten).

Die abschließende Evaluierung einer Datenfusion ist – wie eingangs bereits erwähnt – der wichtigste Schritt, zu welchem auch die meisten Publikationen vorliegen. Manche der Evaluierungsmöglichkeiten beschäftigen sich mit einer möglichen Verletzung der Conditional Independence Assumption, andere mit dem Verteilungserhalt vor und nach Fusion und wieder andere beziehen sich auf vorab durchgeführte „Testfusionen“, bei denen ein künstliches Datenausfallmuster erzeugt und anschließend ergänzt wird. Wir geben zum Abschluss des Vortrags eine Einschätzung bezüglich der Anwendbarkeit dieser Evaluierungen vor dem Hintergrund des Analyseziels einer Datenfusion.