

Die FuU-Teilnehmer-Datei

1976 -1997

Ruth Miquel, Conny Wunsch und Michael Lechner *

SIAW



Juli 2002

Kapitel 5: Qualität der Daten und Probleme

Korrespondenz-Adresse:

Ruth Miquel

Schweizerisches Institut für Ausenwirtschaft und Angewandte Wirtschaftsforschung (SIAW)

Universität St. Gallen

Dufourstr. 48, CH-9000 St. Gallen, Schweiz

ruth.miquel@unisg.ch

* Wir bedanken uns beim IAB (Projekt 6-531 A), Nürnberg für die finanzielle Unterstützung. Ausserdem danken wir Eva Poen für Ihre Unterstützung bei der Erstellung dieser Arbeit, Stefan Bender und Matthias Gehricke für ihre Informationen und Anmerkungen, Elizabeth Nagel und Heinz Gommlich für die Vorbereitung der Daten, und Bernd Fitzenberger, Stefan Speckesser und Annette Bergemann für ihre Kommentare. Alle verbleibenden Fehler und Ungenauigkeiten sind unsere eigenen.

5 Qualität der Daten und Probleme

In den vorherigen Abschnitten haben wir die auf den bisher verfügbaren Informationen basierenden Soll-Beschreibungen der Daten vorgestellt. Numerisch, d.h. in der "Realität" können die vorhanden, auswertbaren Daten jedoch davon abweichen. Um die Zuverlässigkeit der Daten zu überprüfen, haben wir Dummy- und qualitative Variablen generiert, die Aufschlüsse über mögliche Abweichungen geben. Für diese gibt es eine Reihe möglicher Erklärungen. Wie bereits beschrieben sind zwei Grössen für eine FuU-Meldung von zentraler Bedeutung: das Eingangsdatum der Meldung und das Anfangsdatum der Massnahme (Eintrittsdatum). Es ist möglich, dass das Anfangsjahr der Massnahme und das Eingangsjahr der Meldung verschieden sind, vor allem bei Massnahmeeintritten im letzten Quartal. Wenn nun Änderungen in den Codierungen genau zwischen diesen beiden Jahren erfolgt sind, ist apriori nicht klar, nach welcher Codierung die Daten erfasst wurden. Grundsätzlich wurden die Erfassungs- und Prüfprogramme jeweils zum Jahreswechsel geändert. Die Verarbeitungslogik hat dabei je nach Art der Änderung Altfälle gesondert berücksichtigt. Leider haben wir keine Informationen darüber, wie genau diese gesonderte Berücksichtigung dieser Fälle erfolgte. Auch ist in den Daten keine Systematik zu finden, so dass die Ausprägungen der Beobachtungen manchmal besser nach den Codierungen des Eingangsjahres, manchmal nach denen des Anfangsjahrs der Massnahme erklärbar sind. Erschwerend kommt hinzu, dass gleiche Ausprägungen in den Daten in unterschiedlichen Jahren eine unterschiedliche Bedeutung haben können. In diesem Fall ist es unmöglich zu entscheiden, nach welchem Jahr die Daten codiert wurden. Um diese Problematik transparenter zu machen, wurden binäre Variablen gebildet, die den Wert "Eins" annehmen, wenn das Anfangsjahr der Massnahme und das Eingangsjahr der Meldung verschieden ist und gleichzeitig unterschiedliche Bedeutungen in den Regelungen vorliegen, und "Null" sonst. Die Ergebnisse sind im Anhang, Tabelle A1, dargestellt. Die meisten Problemfälle finden sich in den 70er Jahren (bis 21% in 1979 für die "Stellung im Beruf" Variable).¹ Die Lehrgangskostenvariable weist z.B. aber auch in späteren Jahren hohe Anteile auf (1989: fast 10%).

Als nächstes haben wir eine Überprüfung zwischen der Soll-Codierung und der Codierung der tatsächlich vorliegenden Daten pro Jahr (Anfangsjahr der Massnahme und Eingangsjahr der Meldung) durchgeführt. Selbstverständlich wird für Variablen, die nach dem Endejahr erfasst werden (z.B. *FERGBNIS*), das Eingangsjahr der Austrittsmeldung und das Jahr des

¹ Das Datum, das verwendet wurde, um die Sätze pro Jahr zu berechnen, ist das Eintrittsdatum in die Massnahme (*AJAHR*).

Massnahmeendes betrachtet. Der als erstes geprüfte Fehler betrifft die in den Regelungen zulässigen Ausprägungen: Abweichungen zwischen den Werten in den Regelungen und den in den Daten beobachteten Werten definieren wir als einen Fehler vom Typ 1. Die Häufigkeit dieses Fehlers variiert über die verschiedenen Variablen und Jahre. Über den gesamten betrachteten Zeitraum (Anfangsdatum der Massnahme) haben nur drei Variablen mehr als 10% Fehler vom Typ 1: die Arbeitslosigkeitsdauer in Monaten (*FALODAU*) 16.5%, die Kinderbetreuungskosten (*FKIDKOST*) 14.2% und die Rehabilitand/Familienstand-Variable (*FREHA*) 10.1%. Die übrigen Variablen haben durchschnittlich 3.3% Fehler. Der Anteil der Fehler ist jedoch in einigen Jahren dramatisch hoch. Für die 70er Jahre kann die Fehlerquote sogar 100% erreichen (*FALODAU*, *FLEHKOST* und *FUNTART*). Die meisten dieser Fehler erscheinen in den Jahren, in denen Änderungen in den Ausprägungen und/oder Vercodungen vorgenommen wurden. Für Variablen, die neu eingeführt wurden, treten die meisten Fehler vor dem Jahr ihrer Einführung auf. Ab 1993 finden sich kaum noch Fehler vom Typ 1. Die Ergebnisse nach dem Eintrittsjahr in die Massnahme (*AJAHR*) sind im Anhang, Tabelle A2, enthalten.²

Die Tatsache, dass für Variablen, die neu eingeführt wurden, die meisten Fehler vor dem Jahr ihrer Einführung auftreten, kann dadurch erklärt werden, dass bei Einführung eines neuen Feldes in den Datensatz ein altes, nicht mehr benötigtes Feld verwendet wurde, um eine sogenannte Bandsatzerweiterung zu vermeiden. War nun ein solches Feld vor Einführung der neuen Variable nicht leer bzw. Null oder Missing, können beispielsweise für diesen Zeitraum aus Sicht der neuen Variable unerklärte Werte auftreten. Hiermit lässt sich z.B. der hohe Fehleranteil in dem Merkmal "Arbeitslosigkeitsdauer in Monaten" (*FALODAU*) erklären. Diese Variable sollte ab 1993 existieren. Wir finden jedoch auch vor 1993 verschiedene Werte in der Variablen. Insbesondere finden wir vor 1986 Kategorien, für die wir keine Erklärung haben.

Ein ähnliches Phänomen zeigt sich bei der Kinderbetreuungskosten-Variable (*FKIDKOST*). Diese hat Fehler zwischen 1976 und 1992. Da diese Variable jedoch erst ab 1993 existiert, sollte sie leer sein (Null). Wir finden jedoch Werte ungleich Null, die wir somit als Fehler betrachten. Auch hier ist der höchste Fehlerwert wiederum unmittelbar vor der Einführung des Merkmals (1992: 64.3%). Zuvor war der Fehleranteil durchschnittlich nur bei 14%. Der hohe Fehleranteil in 1992 ist vermutlich darauf zurückzuführen, dass in 1992 bereits die erst ab 1993 abgefragten Informationen zu einem hohen Anteil enthalten sind. Denn die in 1992 als Fehler

² Die Ergebnisse nach dem Eingangsjahr der Eintrittsmeldung können von den Autoren bezogen werden.

betrachteten Werte in den Daten entsprechen den ab 1993 gültigen Ausprägungen "Kinderbetreuungskosten ja" und "Kinderbetreuungskosten nein".

Etwas anders ist es bei der Rehabilitand-Variable (*FREHA*). Hier treten Fehler nur von 1988 bis 1992 auf, mit einer starken Häufung in 1991/92. Diese Fehlerhäufung ist durch einen Bedeutungswechsel der Variable erklärbar. Vor 1993 enthält diese Variable Informationen über die Familienstand und nach 1993 Informationen über den Rehabilitationsstand. Hier wurde also anscheinend dasselbe Feld erst für die Variable "Familienstand", die bis 1992 abgefragt und erfasst wurde, und dann ab 1993 für die neue Rehabilitand-Variable (*FREHA*) verwendet. Dabei tritt bereits in 1991 und 1992 die spätere Ausprägung "Rehabilitand nein" sehr häufig auf, die jedoch aus Sicht der Familienstandsvariable in diesen beiden Jahren noch als Fehler betrachtet wird und somit für den hohen Fehleranteil in diesen beiden Jahren verantwortlich ist.

Einen vergleichsweise hohen Fehleranteil (Typ 1) weist mit 6.7% auch die Variable "Art des Unterrichts" *FUNTART* auf, wobei Fehler ausschliesslich in den Jahren 1982 bis 1986 auftreten mit einer Häufung in 1985 und 1986. Dies ist vermutlich auf einen Wechsel in der Vercodung von 1985 auf 1986 zurückzuführen. Denn in den Jahren 1983-1985 treten als Fehler ausschliesslich die ab 1986 zulässigen Werte auf und umgekehrt in 1986. Werden die parallel auftretenden unterschiedlichen Vercodungen als äquivalent angenommen, so ergibt sich aus diesen Fehlern kein Informationsverlust.

Was den sehr hohen Fehleranteil in den 70er Jahren betrifft, so haben verschiedene Plausibilitätsprüfungen zu der Vermutung geführt, dass die Daten für diese Zeitperiode nach der in 1980 erlaubten Vercodung erfasst wurden. Werden die Fehler vom Typ 1 mit der für 1980 zulässigen Vercodung betrachtet, so reduziert sich beispielsweise bei den Lehrgangskosten der Fehleranteil von 100% auf 60% bis 87%. Der Fehleranteil bleibt jedoch insgesamt auch mit der 1980 gültigen Vercodung relativ hoch.

Wie bereits in Abschnitt 4 beschrieben, impliziert die Eingabe bestimmter Massnahmen, Beschränkungen im Merkmalskanon anderer Variablen. Finden sich in diesen Variablen andere Merkmale als die Zulässigen, werden sie als Fehler vom Typ 2 (bei Einarbeitung), Typ 3 (für Deutsch-Sprachlehrgang) und Typ 4 (für §41a-Massnahmen: Verbesserung der Vermittlungsaussichten) erfasst. Zusätzlich implizieren einige Variablen – unabhängig von der Art der Massnahme - Beschränkungen in den möglichen Ausprägungen anderer Variablen. Verletzungen dieser Beschränkungen definieren wir als Fehler vom Typ 5. Für jede Variable ist zusätzlich eine Gesamtfehlerquote berechnet, die im Anhang, Tabelle A4 enthalten ist.

Der Fehler vom Typ 2 kommt selten vor. Nur die Variable bezüglich der Massnahmeträger (*FTRAEGER*) hat einen Fehleranteil von insgesamt 5.4%. Hierbei sind allerdings Schwankungen im Fehleranteil über die Jahre festzustellen: 7.5% zwischen 1987 und 1991, 10.2% in 1980 und ungefähr 4% für die anderen Jahre (ab1980). Auch einige andere Variablen haben Fehler vom Typ 2, allerdings mit weniger als 2% Fehleranteil. Diese Resultate verändern sich auch nicht, wenn man anstelle des Anfangsjahres der Massnahme, das Eingangsjahr der Meldung betrachtet.

Die Fehler vom Typ 3 (durchschnittlich unter 0.1%) und Typ 4 (maximaler Fehleranteil 0.8%) kommen seltener als der Fehler vom Typ 2 vor. Trotz dieser relativ geringen Fehleranteile, gibt es bei einigen Merkmalen für einige Jahre hohe Fehleranteile: beispielsweise hat die Variable "Art des Unterrichts" (*FUNTART*) 1982 7% und 1984 12.2% Fehler. Allgemein können wir jedoch sagen, dass die Fehlertypen (2-4) kein wirkliches Problem für Analysen mit den Daten darstellen.

Zwei Variablen, die Leistungsvariable und die Lehrgangskostensvariable, leiden besonders unter dem Fehler vom Typ 5. Obwohl deren Fehleranteil insgesamt nicht grösser als 3.5% ist, liegt er in 1980 und 1981 zwischen 41% und 45%.

Insgesamt gibt es vier Variablen, die, was die Fehlerquote betrifft, problematisch sind. Wenn wir alle möglichen Fehlertypen zusammen betrachten, haben diese Variablen eine Fehlerquote von über 7%: "Arbeitslosigkeitsdauer in Monaten" *FALODAU* 16.5%, "Kinderbetreuungskosten" *FKIDKOST* 15%, "Rehabilitand" *FREHA* 10% und "Art des Unterrichts" *FUNTART* 7.5%. Bei der Arbeitslosigkeitsdauer- und der Kinderbetreuungskostenvariable handelt es sich um Variablen, die erst ab 1993 erfasst wurden und bei denen Fehler ausschliesslich vor diesem Zeitpunkt auftreten. Ihre Verwendbarkeit für Analysen ist damit nicht eingeschränkt, sofern sie ausschliesslich ab 1993 verwendet werden. Bei der Rehabilitand-Variable ist der Bedeutungswechsel von 1992 (Familienstand) auf 1993 (Rehabilitand ja/nein) zu beachten. Problematisch sind die Jahre 1991 und 1992, in denen die Merkmale anscheinend gemischt erfasst wurden, wodurch der Fehleranteil sehr hoch ist. Zudem ist die Aussagekraft der (echten) Rehabilitand-Variable für 1993/1994 gering, da eigentlich nur Nicht-Rehabilitanden im Datensatz enthalten sein sollten. Die Variable sollte daher in Analysen nur für die Jahre bis einschliesslich 1990 als Indikator für den Familienstand verwendet werden. Bezüglich der Variable "Art des Unterrichts" *FUNTART* wurde bereits festgestellt, dass sich, wenn die parallel auftretenden unterschiedlichen Vercodungen als

äquivalent angenommen werden, kein Informationsverlust aus den beobachteten Fehlern vom Typ 1 ergibt.

Ein Erfassungsproblem besteht darüber hinaus bei der Variable "Massnahmedauer" (*FTATADUR*). Ein- und Austrittsdatum werden nur monatsgenau erfasst und daraus wird dann die Dauer der Massnahme berechnet. Somit sind Abweichungen von der tatsächlichen Dauer von bis zu zwei Monaten denkbar. Darüber hinaus wird im Falle, dass ein Antragsteller in einem Monat zwei Massnahmen beginnt, das Eintrittsdatum der zweiten Massnahme fiktiv auf den nächstfolgenden Monat festgelegt, wodurch weitere Ungenauigkeiten entstehen. Ausserdem sind generell nur Zeiten erfasst, in denen ein Teilnehmer eine Förderung erhält.

Die Zuverlässigkeit der Daten kann ergänzend durch einen Vergleich der verschiedenen Datumsangaben im Datensatz geprüft werden. Es bieten sich hierbei folgende Vergleiche an:

- Anfangsdatum der Massnahme mit Eingangsdatum der Meldung,
- Endedatum der Massnahme mit Eingangsdatum der Austrittsmeldung,
- Endedatum der Massnahme mit dem voraussichtlichen Ende der Massnahme.

Wenn das Endedatum der Massnahme mehr als zwei Monate nach dem voraussichtlichen Ende der Massnahme liegt, ist das Endedatum der Massnahme wahrscheinlich nicht zuverlässig und es ist besser, das voraussichtliche Ende der Massnahme als Ende der Massnahme zu betrachten. Dieser Fall tritt bei 2.4% der Beobachtungen auf.

Diese "Zwei-Monatsregel" wird auch beim Vergleich zwischen dem Anfangs- und Endedatum der Massnahme und den jeweiligen Eingangsdatum der Meldungen angewendet. Es wird angenommen, dass diese zeitliche Verzögerung beim Eingang der Meldung im zentralen Archiv der Bundesanstalt für Arbeit auf einen Fehler in der Meldung hindeutet. 20% der Meldungen haben dieses Problem bezüglich des Anfangsdatums und 41% bezüglich des Austrittsdatums (siehe Anhang, Tabelle A3). Diese hohen Werte implizieren, dass zumindest die Informationen der Variablen "Ergebnis der Massnahme" *FERGBNIS* und "Ende der Massnahme" *FMASENDE*, die aus der Austrittsmeldung stammen, wahrscheinlich eher fragwürdig sein dürften.

Darüber hinaus haben auch einige Meldungen kein gültiges Ende- oder Eingangsdatum in der Austrittsmeldung. Hierbei sind folgende Möglichkeiten aufgetreten:

- kein Eintrag im Austrittsdatum,
- das Endedatum früher als 1975,
- das Endedatum ist zeitlich vor dem Anfangsdatum der Massnahme,

- das Eingangsdatum der Austrittsmeldung ist zeitlich vor der Eintrittsmeldung.

Da diese Punkte allerdings weniger als 1% der Meldungen sind, sind wir der Auffassung, dass diese Beobachtungen gelöscht werden könnten.

Insgesamt kann gesagt werden, dass Angaben nach 1993 zuverlässiger sind als frühere Meldungen. Betrachtet man alle Meldungen im gesamten Zeitraum (22 Jahre) so sind nur 56% vollständig fehlerfrei (vgl. Anhang A5). Führt man diese Betrachtung allerdings pro Jahr durch, so erhöht sich der Prozentsatz der fehlerfreien Meldungen auf über 80% in 1993 und ab 1994 auf ungefähr 92%. Im Gegensatz dazu sind z.B. in 1992 nur 30% der Beobachtungen ohne Fehler.³ Extrem problematische Jahre sind auch 1984, 1985 und 1986, wo der Anteil an fehlerfreien Meldungen unter 17% (mit weniger als 3% in 1984!) liegt. In den Jahren vor 1984 haben alle Beobachtungen mindestens einen Fehler, d.h. vor 1984 ist kein Eintrag der FuU-Teilnehmerstatistik fehlerfrei.

Trotz dieser ernüchternden Fehlerbilanz kommen wir zu dem Schluss, dass diese Daten nützlich und anwendbar für die Forschung sind. Nicht zuletzt daher, weil das Ausmass der Fehlerhäufigkeit nicht allzu gravierend ist. So haben die meisten Spells (ca. 88%) keine Fehler (56%), nur einen (15%) oder zwei Fehler (17%) in der gesamten Meldung (siehe Anhang, Tabelle A5). Des weiteren sind einige der für die Forschung wichtigen Merkmale, beispielsweise die Angaben zur Ausbildung oder die Berufserfahrung der Teilnehmer - sehr detailliert und vergleichsweise zuverlässig.

Obwohl die Informationen über die Massnahmen sehr detailliert sind, können diese Informationen für Forschungsfragen nicht direkt verwendet werden. Beispielsweise die in der Variable "Art der Massnahme" *FMASART* erfassten Massnahmen geben aufgrund des hohen Aggregationsgrades kaum Aufschluss über den konkreten Charakter und Inhalt einer Massnahme: Im wesentlichen kann nur zwischen Einarbeitungen, Umschulungen, §41a-Massnahmen, Aufstiegsmassnahmen, Massnahmen in einer Übungsfirma oder -werkstatt sowie sonstigen Fortbildungsmassnahmen unterschieden werden. Letztere, die Massnahmeart mit dem geringsten Informationsgehalt, umfasst zudem noch die meisten Beobachtungen. Gleiches gilt auch für die Variable "Träger der Massnahme" *FTRAEGER*, wo sich die Beobachtungen bei der Ausprägung "sonstiger Träger" häufen.

³ Wir haben auch geprüft, ob die relativ hohen Fehlerquoten in 1991/1992 eventuell auf das Hinzukommen der neuen Bundesländer in die Statistik zurückzuführen ist. Hierfür konnten jedoch keine Anzeichen gefunden werden.

Ausserdem können sogenannte "gesplittete" Massnahmen, d.h. Massnahmen, die in mehreren Teilen durchgeführt werden, nicht identifiziert werden.⁴ Die im Datensatz enthaltenen Informationen über die Massnahmen stellen sich als ungenügend heraus, um zu identifizieren, ob es sich bei einer Meldung um eine Teilnahme an verschiedenen Massnahmen oder um verschiedene Teile einer einzigen Massnahme handelt. Es ist beispielsweise möglich, dass eine berufsqualifizierende Massnahme in mehreren Teilen mit unterschiedlichen inhaltlichen Schwerpunkten durchgeführt wurde. Auch legen Vermutungen nahe, dass solche gesplitteten Massnahmen in einigen Fällen jeweils als mehrere Einzelmassnahmen erfasst worden sein könnten; manchmal sogar als zeitlich parallel verlaufende Massnahmen.

Um diesen Sachverhalt zu verdeutlichen, sollen nachfolgend zwei Beispiel-Personen betrachtet werden: Die erste weist drei direkt aufeinanderfolgende FuU-Spells beim selben Arbeitsamt auf: 1/80-3/80, 4/80-10/80 und 10/80-4/81. Beim ersten Spell handelt es sich um eine "sonstige Anpassung der beruflichen Kenntnisse und Fertigkeiten" ohne Prüfung, beim zweiten Spell ebenfalls um eine "sonstige Anpassung der beruflichen Kenntnisse und Fertigkeiten", jedoch mit Prüfung. Die Berufskennziffer des Schulungsziels unterscheidet sich geringfügig. Der dritte Spell ist eine "sonstige Aufstiegsfortbildung" mit Prüfung, die Berufskennziffer des Schulungsziels ist jedoch dieselbe wie beim zweiten Spell. Unterrichtsart, Lernort und Trägerkennung sind in allen drei Fällen gleich. Bezüglich der beiden letzten Spells könnte man nun vor allem aufgrund desselben Schulungsziels vermuten, dass es sich um eine einzige, in zwei Teilen durchgeführte Aufstiegsmassnahme mit Zwischen- und Abschlussprüfung gehandelt hat. Anhand der Daten lässt sich dies jedoch weder bestätigen noch falsifizieren.

Die zweite Person weist drei Spells auf. Der erste Spell läuft von 1/88-9/90 und der zweite ab 9/89-9/90, also parallel zum ersten, sowie direkt im Anschluss der dritte Spell von 9/90-3/91. Während der letzte Spell eine "sonstige Anpassung der beruflichen Kenntnisse und Fertigkeiten" ist, handelt es sich bei den beiden ersten Spells jeweils um eine Umschulung. Da es nun weder sinnvoll, noch aus gesetzlicher Sicht möglich ist, zwei geförderte Umschulungen gleichzeitig zu besuchen, liegt die Vermutung nahe, dass es sich eigentlich um eine einzige Umschulungsmassnahme handelt. Bei einer näheren Analyse der Spells zeigt sich, dass es sich beim ersten Spell um eine rein schulische Massnahme unter der Trägerschaft der Bundesanstalt für Arbeit handelt. Beim zweiten dagegen um eine rein betriebliche Massnahme unter der Trägerschaft eines Betriebes. Dies lässt vermuten, dass die Umschulung von 1/88 bis 8/89 aus einem rein theoretischen/schulischen und ab 9/89 aus einem zusätzlichen bzw. parallelen

⁴ Dies betrifft vor allem Aufstiegsmassnahmen und hier vor allem die sogenannten Meisterkurse.

praktischen/betrieblichen Teil durchgeführt wurde. Auch hier enthalten die Daten keinerlei Informationen, welche diese These untermauern könnten. Daher bemühen wir uns durch eine Befragung von Arbeitsämtern über ihr Vorgehen in solchen Fällen, zusätzliche Informationen zu den Daten zu erhalten.

Um diese Daten trotzdem bestmöglich nutzen zu können, werden auch ergänzende Informationen aus anderen Datenquellen benötigt. Als eine Hauptquelle sind hier Informationen über den Leistungsbezug von Arbeitslosengeld-, Hilfe und Unterhaltsgeld zu nennen. Diese Informationen liegen beispielsweise in Form der sogenannten Leistungsempfängerdatei vor. Allein die Informationen über die Art des Leistungsbezugs kann sehr aufschlussreich für eine Interpretation der dargestellten Beispiele sein: Findet sich beispielsweise für unsere zwei Personen jeweils über den gesamten Zeitraum (4/80-4/81 im ersten Beispiel bzw. 1/88-9/90 im zweiten Beispiel) ein parallel verlaufender ununterbrochener Unterhaltsgeld-Spell in der Leistungsempfängerdatei, so handelt es sich in beiden Fällen höchstwahrscheinlich um eine einzige Massnahme, die in zwei Teilen durchgeführt wurde. Wichtig für eine Zusammenführung dieser Datenquellen ist allerdings ein eindeutiger Personenidentifikator, der nicht immer in den Daten der Bundesanstalt für Arbeit zur Verfügung steht.