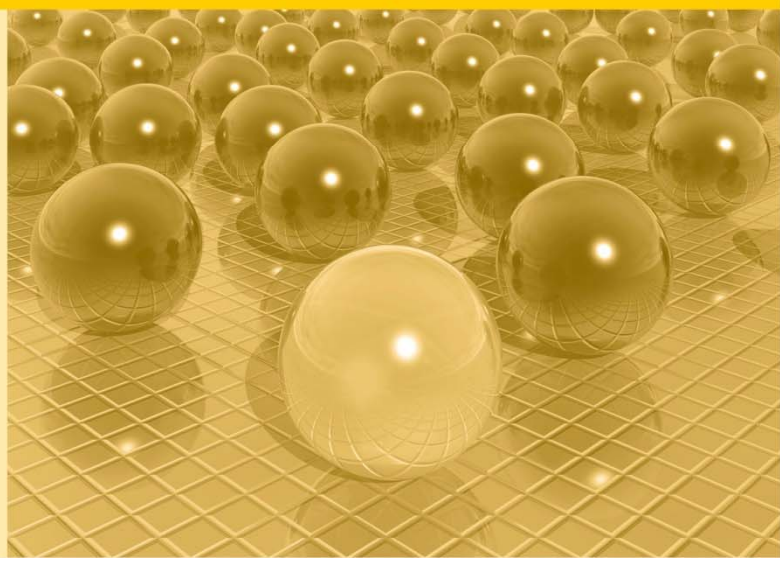
FDZ-Arbeitspapier Nr. 51



Record-Linkage bei fehlenden Betriebsnummern im Produzierenden Gewerbe

Lukas Mödl

2020

STATISTISCHE ÄMTER DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Impressum

Herausgeber: Statistische Ämter des Bundes und der Länder
Herstellung: Information und Technik Nordrhein-Westfalen
Telefon 0211 9449-01 • Telefax 0211 442006
Internet: www.forschungsdatenzentrum.de
E-Mail: forschungsdatenzentrum@it.nrw.de

Fachliche Informationen
zu dieser Veröffentlichung:

Forschungsdatenzentrum der
Statistischen Ämter der Länder
Standort Berlin-Brandenburg
Tel.: 030 9021-3300
Fax: 030 9028-4038
forschungsdatenzentrum@statistik-bbb.de

Informationen zum Datenangebot:

Statistisches Bundesamt
Forschungsdatenzentrum

Tel.: 0611 75-3277
Fax: 0611 72-3915
forschungsdatenzentrum@destatis.de

Forschungsdatenzentrum der
Statistischen Ämter der Länder
– Geschäftsstelle –
Tel.: 0211 9449-2883
Fax: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

Erschienen im September 2020

Diese Publikation wird kostenlos als **PDF-Datei** zum Download unter www.forschungsdatenzentrum.de angeboten.

FDZ-Arbeitspapier
Nr. 51

Record-Linkage bei fehlenden Betriebsnummern im Produzierenden Gewerbe

Lukas Mödl

2020

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Record-Linkage bei fehlenden Betriebsnummern im Produzierenden Gewerbe

Lukas Mödl

14. Juli 2020

Zusammenfassung

Im Rahmen eines Kooperationsprojektes zwischen dem Amt für Statistik Berlin-Brandenburg und der Humboldt-Universität zu Berlin wurde versucht, den Betrieben der Gehalts- und Lohnstrukturerhebung 1995 (im Bereich des Produzierenden Gewerbes) die jeweilige Betriebsnummer des Unternehmensregisters 95 zuzuordnen. Die Zuordnung wurde realisiert, indem die Gehalts- und Lohnstrukturerhebung 1995 auf Betriebsebene über ein deterministisches Record-Linkage mit dem Monatsbericht und der Kleinbetriebserhebung für Betriebe im Verarbeitenden Gewerbe 1995 verknüpft wurde. Obwohl nur vier Überschneidungsmerkmale existieren und es aufgrund struktureller Unterschiede zwischen den Erhebungen teils erhebliche Abweichungen in den Merkmalsausprägungen gibt, konnten Precision- und Recall-Werte von 97 % bzw. 87 % erzielt werden. Für bestimmte Beschäftigtengrößenklassen wurden Recall-Werte von über 90 % erzielt.

Inhaltsverzeichnis

Abkürzungsverzeichnis

AGS	Amtlicher Gemeindeschlüssel
FDZ	Forschungsdatenzentrum der Statistischen Landesämter
GLS	Gehalts- und Lohnstrukturerhebung im Produzierenden Gewerbe und im Dienstleistungsbereich
KBE	Erhebung für industrielle Kleinbetriebe im Bereich Verarbeitendes Gewerbe, Bergbau und Gewinnung von Steinen und Erden
MBB	Monatsbericht im Bereich Verarbeitendes Gewerbe, Bergbau und Gewinnung von Steinen und Erden
NACE	Systematik der Wirtschaftszweige in der Europäischen Gemeinschaft
SYPRO	Systematik der Wirtschaftszweige 1979, in der Fassung für das Produzierende Gewerbe
URS	Unternehmensregister-System 95
WZ93	Systematik der Wirtschaftszweige 1993

1 Einleitung

„Linked-employer-employee“-Datensätze erlauben differenzierte Analysen von Verdienststrukturen. Sie sind deshalb vor allem in der Arbeitsmarktforschung von Bedeutung. Einer der wichtigsten „Linked-employer-employee“-Datensätze in Deutschland war bis 2006 die Gehalts- und Lohnstrukturerhebung im Produzierenden Gewerbe und im Dienstleistungsbereich (GLS).[1]

Seit der GLS 2001 ist es möglich, das Datenmaterial der GLS auf Betriebsebene vollständig mit dem Unternehmensregistersystem 95 (URS) zu verknüpfen. Durch die Betriebsnummer im URS können weitere Datenbestände an die GLS angespielt werden – das Forschungspotenzial der GLS wird dadurch enorm erhöht. Aufgrund fehlender Identifikationsmerkmale ist eine direkte Verknüpfung mit dem URS bei früheren Erhebungen (GLS 1990 und GLS 1995) allerdings nicht möglich. Ziel dieses Projektes ist es, die GLS 1995 auf Betriebsebene mit dem URS (damals noch „Kartei im Produzierenden Gewerbe“) zu verknüpfen. Da ca. 61 % aller Betriebe und rund 70 % aller Beschäftigten in der GLS 1995 im Produzierenden Gewerbe (Abteilungen 10 bis 37 der Systematik der Wirtschaftszweige 1993, WZ93) tätig sind, wird sich, aufgrund der zeitlichen Beschränkung des Projektes, auf diese Betriebe beschränkt.

Da die Betriebe in der GLS 1995 nur mit einer systemfreien Betriebsnummer identifiziert werden, ist eine direkte Verknüpfung mit dem URS nicht ohne weiteres möglich. Es wird daher ein Umweg über den Monatsbericht und die Kleinbetriebserhebung im Bereich Verarbeitendes Gewerbe, Bergbau und Gewinnung von Steinen und Erden (MBB bzw. KBE) 1995 gegangen. MBB und KBE 1995 enthalten die originale Betriebsnummer des URS und bilden zusammen eine Vollerhebung aller 1995 in den WZ93-Abteilungen 10 bis 37 tätigen Betriebe. Es sollte daher möglich sein, für jeden Betrieb in der GLS 1995 den entsprechende Betrieb im MBB oder in der KBE 1995 zu finden, d. h. die GLS 1995 sollte auf Betriebsebene vollständig mit dem MBB und der KBE 1995 verknüpfbar sein.[2] Aufgrund fehlender Identifikationsmerkmale kann die Verknüpfung aber nur auf Basis von Merkmalen realisiert werden, die sich zwischen den Erhebungen überschneiden. Es existieren vier Überschneidungsmerkmale: i) der Standort, an dem ein Betrieb angesiedelt ist, ii) der Wirtschaftszweig (im Sinne der WZ93), in dem ein Betrieb tätig ist, iii) der Handwerksrolleneintrag eines Betriebs und iv) die Anzahl der Beschäftigten/tätigen Personen eines Betriebs.

Für die Verknüpfung wird ein deterministischer Record-Linkage-Ansatz vorgeschlagen. Wie in einem ähnlichen Projekt von Hafner [1] wird davon ausgegangen, dass sich eine potenzielle Zuordnung genau dann auf denselben Betrieb bezieht, wenn Standort, Wirtschaftszweig, Handwerksrolleneintrag übereinstimmen (Blocking-Schritt) und die Differenz aus der Anzahl der Beschäftigten/tätigen Personen minimal ist (Linkage-Schritt). Durch eine gelungene Zuordnung kann den Betrieben in der GLS 1995 die originale Betriebsnummer zugewiesen werden. Für etwa ein Drittel der Betriebe in der GLS 1995 ist die originale Betriebsnummer bereits bekannt, da die originalen Betriebsnummern für einige Bundesländer noch im Datenmaterial des Forschungsdatenzentrums der Statistischen Landesämter (FDZ) für die GLS 1995 vorliegen.

[1] Die GLS wurde 2006 von der Verdienststrukturerhebung abgelöst.

[2] Rechtlich gesehen ist eine solche Verknüpfung unproblematisch. Paragraph 13a des Bundesstatistikgesetzes erlaubt das Zusammenführen von Daten aus verschiedenen Wirtschaftsstatistiken.

Aufgrund von strukturellen Unterschieden zwischen den Erhebungen (abweichende Erhebungszeiträume und Merkmalsdefinition) werden neben dem Verknüpfungsvorschlag auch die geblockten Zuordnungen zusammen mit weiteren Merkmalen bereitgestellt. Dies erlaubt in späteren Datennutzungen gegebenenfalls gezielt auf bestimmte Anwendungen ausgerichtete Verknüpfungen zu realisieren. Die geringe Anzahl von Überschneidungsmerkmalen und die strukturellen Unterschiede bleiben aber bei allen Verknüpfungen ein grundsätzliches Problem.

In Abschnitt 2 wird zunächst ein kurzer Überblick über das Datenmaterial der Erhebungen gegeben. Abschnitt 3 geht im Detail auf die vorhandenen Überschneidungsmerkmale ein und bewertet deren Qualität im Hinblick auf die Verknüpfung. Der vorgeschlagene Record-Linkage-Ansatz wird in Abschnitt 4 erläutert und anschließend anhand derjenigen Betriebe evaluiert, für die die Umsteigeschlüssel zwischen der originalen Betriebsnummer des URS und der in der GLS 1995 verwendeten systemfreien Betriebsmummer vorhanden sind. Einige abschließende Bemerkungen bezüglich der Verknüpfung finden sich in Abschnitt 5. Appendix A enthält eine Übersicht über die bereitgestellten Daten inklusive einer Beschreibung der enthaltenen Merkmale. In Appendix B ist der verwendete Code dokumentiert.

2 Beschreibung des Datenmaterials

Das verwendete Datenmaterial wurde im Rahmen eines Kooperationsprojektes zwischen dem Amt für Statistik Berlin-Brandenburg und der Humboldt-Universität zu Berlin im FDZ bereitgestellt. Aus Geheimhaltungsgründen ist das Datenmaterial nur mit einem Zugang über die Forschungsdatenzentren verfügbar. Zur Wahrung der Geheimhaltung muss für veröffentlichte Werte eine Fallzahlenuntergrenze von mindestens drei Fällen pro Wert eingehalten werden. Ist dies nicht gewährleistet, darf der entsprechende Wert nicht veröffentlicht werden und muss entweder gesperrt oder durch das Bilden eines Durchschnitts aus zwei weiteren Werten pseudo-anonymisiert werden. Im Folgenden sind gesperrte Werte mit dem Symbol „≤ 3“ und pseudo-anonymisierte Werte mit dem hochgestellten Suffix „$\varnothing$“ gekennzeichnet.

2.1 GLS 1995

Die GLS 1995 war eine zweistufige Stichprobenerhebung, die dezentral von den Statistischen Landesämtern durchgeführt wurde. In der ersten Stufe wurde eine nach 17 Regionen, 47 WZ93-Gruppen und sechs Beschäftigtengrößenklassen geschichtete Stichprobe gezogen [2]. Zum Berichtkreis gehörten alle Betriebe in den WZ93-Abteilungen 10 bis 37 und 65 bis 67 mit mindestens 10 Beschäftigten.[3] Auf der ersten Stufe wurden unter anderem der Standort, der Wirtschaftszweig, die Anzahl der Beschäftigten und Informationen über die Tarifbindung eines Betriebs erhoben. Für jeden Betrieb, der in der ersten Stufe ausgewählt wurde, wurde auf der zweiten Stufe eine Stichprobe der Beschäftigten gezogen. Bei Betrieben mit weniger als 50 Beschäftigten wurden alle Beschäftigten ausgewählt. Zum Berichtskreis auf der zweiten Stufe gehörten alle sozialversicherungspflichtigen Beschäftigten mit Ausnahme von Auszubildenden, Vertreter□innen juristischer Personen, Heimarbeiter□innen und Beschäftigte in Altersteilzeit

[3] Der Auswahlzeitpunkt und der Zeitpunkt der tatsächlichen Erhebung waren jedoch nicht identisch. Dadurch weisen 9012 Betriebe im Datenmaterial weniger als 10 Beschäftigte aus. 4040 Betriebe weisen sogar gar keine Beschäftigten aus.

[2]. Auf der zweiten Stufe wurden unter anderem Geschlecht, Alter, Bruttoverdienst oder die Stellung der Beschäftigten im Betrieb erhoben. Eine vollständige Liste aller erhobenen Merkmale auf der ersten und zweite Stufe findet sich im zugehörigen Metadatenreport [3].

Die GLS 1995 wurde 1996 rückwirkend für Oktober 1995 durchgeführt [2]. Für die ausgewählten Betriebe und Beschäftigten bestand Auskunftspflicht. Das im FDZ für das Projekt bereitgestellte Datenmaterial enthält Daten zu insgesamt 26 331 Betrieben und ca. 875 000 Beschäftigten. Davon sind 16 107 Betriebe mit fast 615 000 Beschäftigten in den WZ93-Abteilungen 10 bis 37 tätig. Fehlende oder fehlerhafte Werte (bezogen auf die Überschneidungsmerkmale) gibt es in ≤ 3 Fällen. Aufgrund der geringen Anzahl sind diese Fälle unproblematisch und die entsprechenden Betriebe werden im Hinblick auf die Verknüpfung nicht berücksichtigt. Bei 1647 Betrieben wird für die Klassifikation der Wirtschaftszweige noch die ältere Systematik der Wirtschaftszweige 1979 in der Fassung für das Produzierende Gewerbe (SYPRO) verwendet. Aufgrund großer Unterschiede zwischen SYPRO und WZ93 werden diese Betriebe ebenfalls nicht weiter berücksichtigt.

Die Umsteigeschlüssel zwischen der originalen Betriebsnummer des URS und der in der GLS 1995 verwendeten systemfreien Betriebsmummer liegen für (fast alle) Betriebe in den Bundesländern Bayern, Brandenburg, Rheinland-Pfalz, Saarland, Sachsen, Sachsen-Anhalt und Thüringen vor. Die Umsteigeschlüssel sind für insgesamt 9642 Betriebe vorhanden.[4] Davon sind 5667 Betriebe in den WZ93-Abteilungen 10 bis 37 tätig.

2.2 MBB und KBE 1995

Der MBB 1995 war eine monatlich stattfindende Vollerhebung aller 1995 in den WZ93-Abteilungen 10 bis 37 tätigen Betriebe mit mindestens 20 Beschäftigten.[5] Aufgrund von Löschfristen des FDZ liegt für das Jahr 1995 nur noch das kumulierte Jahresergebnis vor. Um einzelne Monatsmeldungen zu approximieren, werden quantitative und qualitative Merkmale daher mittels Jahresmittel bzw. -modus interpoliert. Da nicht jeder Betrieb ganzjährig dem Berichtskreis angehörte, meldeten manche Betriebe weniger als 12-mal pro Jahr. Einige Betriebe meldeten aber auch aus nicht näher bekannten Gründen mehr als 12-mal pro Jahr. Im Mittel (±Standardabweichung) meldete jeder Betrieb ca. 11.58 (±1.80)-mal pro Jahr. Erhoben wurden unter anderem der Standort, der Wirtschaftszweig, die Anzahl der tätigen Personen oder der erwirtschaftete Umsatz. Eine vollständige Liste aller erhobenen Merkmale findet sich im zugehörigen Metadatenreport [4].

Die KBE 1995 ergänzte den MBB 1995 um die in den WZ93-Abteilungen 10 bis 37 tätigen Betriebe mit weniger als 20 Beschäftigten.[6] Zur Entlastung der Kleinbetriebe wurde die KBE 1995 nur im September 1995 durchgeführt und es wurden weniger Merkmale erhoben. Eine vollständige Liste aller erhobenen Merkmale findet sich im zugehörigen Metadatenreport [4].

[4] Die Umsteigeschlüssel sind aus Datenschutzgründen in einem separaten Datensatz (den sogenannten Leitbändern) enthalten. Diese enthalten die Umsteigeschlüssel für insgesamt 9839 Betriebe. Für 197 Betriebe ist der Umsteigeschlüssel zwar vorhanden, jedoch finden sich diese Betriebe nicht im Datenmaterial der GLS 1995.

[5] Für einige WZ93-Abteilungen galten Ausnahmen. Details finden sich im Metadatenreport [4]. 6022 Betriebe (12 %) weisen im Jahresmittel weniger als 20 tätige Personen aus.

[6] Für einige WZ93-Abteilungen galten ebenfalls Ausnahmen. 2010 Betriebe (4 %) weisen 20 oder mehr tätige Personen aus.

MBB und KBE 1995 bilden zusammen eine Vollerhebung aller 1995 in den WZ93-Abteilungen 10 bis 37 tätigen Betrieben. Das im FDZ für das Projekt bereitgestellte Datenmaterial umfasst Informationen über 49 325 bzw. 54 286 Betriebe. Für die befragten Betriebe bestand Auskunftspflicht. Theoretisch könnte ein Betrieb im Laufe eines Jahres zum Berichtskreis beider Erhebungen gehören. Um dies auszuschließen, wurde anhand der Betriebsnummer überprüft, ob Betriebe im Datenmaterial doppelt vorkommen. Da dies bei keinem Betrieb der Fall war, umfasst das Datenmaterial insgesamt 103 611 Betriebe. Fehlende oder fehlerhafte Werte gibt es in ≤ 3 Fällen im MBB 1995. Wie bei der GLS werden die entsprechenden Betriebe nicht weiter berücksichtigt. In der KBE 1995 wurden keine problematischen Werte identifiziert.

3 Vorhandene Überschneidungsmerkmale

Der Standort, der Wirtschaftszweig, der Handwerksrolleneintrag und die Anzahl der Beschäftigten/tätigen Personen eines Betriebs werden sowohl in der GLS 1995 als auch im MBB und in der KBE 1995 erhoben.[7] Im Folgenden werden Trennschärfe und Konsistenz dieser Merkmale untersucht und heuristisch bewertet.

3.1 Trennschärfe der Merkmale

Unter der Trennschärfe eines Merkmals wird im Folgenden die mittlere Anzahl von potenziellen Zuordnungen der Betriebe in der GLS 1995 zu Betrieben im MBB und in der KBE 1995 verstanden, die anhand der Ausprägungen des Merkmals gebildet werden können. Je höher die mittlere Anzahl von potenziellen Zuordnungen, desto geringer ist die Trennschärfe des Merkmals.

Standort. Der Standort eines Betriebs wird durch den Amtlichen Gemeindeschlüssel (AGS) codiert. Der AGS ist ein aus acht Ziffern bestehender Code, der zur einheitlichen Identifikation geografischer Regionen in Deutschland verwendet wird. Der Code besitzt eine hierarchische Struktur. Die ersten beiden Ziffern des Codes identifizieren das Bundesland, die dritte Ziffer den Regierungsbezirk, die vierte und fünfte Ziffer den Kreis und die letzten drei Ziffern die Gemeinde, in der ein Betrieb angesiedelt ist. Im Mittel (±Standardabweichung) können mit dem Standort 157 (±257) potenzielle Zuordnungen pro Betrieb in der GLS 1995 gebildet werden. In dichter besiedelten Regionen ist die Trennschärfe natürlich schlechter. In München ist die Anzahl von potenziellen Zuordnungen mit 1592 am höchsten.

Wirtschaftszweig. Der Wirtschaftszweig, in dem ein Betrieb tätig ist, wird durch die WZ93 klassifiziert. Die WZ93 löste 1995 die bis dahin in den Erhebungen verwendete SYPRO ab. Der aus fünf Ziffern bestehende Code besitzt eine hierarchische Struktur. Die ersten beiden Ziffern identifizieren die WZ93-Abteilung, die dritte Ziffer die WZ93-Gruppe und die vierte und fünfte Ziffer die WZ93-Klasse bzw. die WZ93-Unterklasse. Die ersten vier Ziffern sind äquivalent zur Systematik der Wirtschaftszweige in der Europäischen Gemeinschaft (NACE). Während in der GLS 1995 alle fünf Ziffern des Codes verwendet

[7] In der GLS 1995 und im MBB 1995 wird die Anzahl der Beschäftigten/tätigen Personen nach Angestellten und Arbeiter□innen unterschieden. Da dies für die KBE 1995 nicht der Fall ist, wird nur die Gesamtanzahl der Beschäftigten/tätigen Personen als Überschneidungsmerkmal aufgefasst.

werden, werden im MBB und in der KBE 1995 nur die ersten vier Ziffern verwendet. Damit die Codes vergleichbar sind, werden für die GLS 1995 auch nur die ersten vier Ziffern berücksichtigt.[8] Im Mittel (±Standardabweichung) können mit dem Wirtschaftszweig 1102 (±1364) potenzielle Zuordnungen pro Betrieb in der GLS 1995 gebildet werden. Im Wirtschaftszweig „Druckerei (ohne Zeitungsdruckerei)“ gibt es mit fast 8000 die meisten potenziellen Zuordnungen. Die Trennschärfe des Merkmals „Wirtschaftszweigs“ ist deutlich schlechter als die des Merkmals „Standorts“.

Handwerksrolleneintrag. Ein Betrieb, der in Deutschland ein genehmigungspflichtiges Handwerk ausübt, muss in die von den Handwerkskammern geführte Handwerksrolle eingetragen werden. Der Handwerksrolleneintrag wird in vier Kategorien eingeteilt. Aufgrund der geringen Anzahl von Merkmalsausprägungen ist die mittlere Anzahl von potenziellen Zuordnungen pro Betrieb in der GLS 1995 rechnerisch nicht darstellbar. Die Trennschärfe des Merkmals „Handwerksrolleneintrags“ ist daher gering.

Anzahl der Beschäftigten/tätigen Personen. In der GLS 1995 werden alle sozialversicherungspflichtigen Beschäftigten, mit Ausnahme von Auszubildenden, Vertreter□innen juristischer Personen, Heimarbeiter□innen und Beschäftigten in Altersteilzeit, erhoben. Im MBB und in der KBE 1995 werden dagegen alle im Betrieb tätigen Personen erhoben. Die Anzahl der tätigen Personen eines Betriebs ist damit definitionsbedingt höher als die Anzahl der Beschäftigten in der GLS 1995. Hinzu kommen Abweichungen aufgrund der abweichenden Erhebungszeiträume. Die Trennschärfe des Merkmals „Anzahl der Beschäftigten/tätigen Personen“ ist aufgrund dieser Unterschiede als gering zu bewerten.

3.2 Inkonsistenz in den Merkmalsausprägungen

Im nächsten Schritt wird untersucht, inwieweit sich die abweichenden Erhebungszeiträume und die abweichende Merkmalsdefinition auf die Konsistenz der Überschneidungsmerkmale auswirken. Unter der Konsistenz eines Merkmals wird im Folgenden das Ausmaß an Abweichungen in den Merkmalsausprägungen bei Betrieben verstanden, für die die Umsteigeschlüssel zwischen der originalen Betriebsnummer des URS und der in der GLS 1995 verwendeten systemfreien Betriebsnummer vorhanden sind. Von den 5667 Betrieben in der GLS 1995, für die die originale Betriebsnummer vorhanden ist, können 4367 Betriebe im Datenmaterial des MBB und 651 Betriebe im Datenmaterial der KBE 1995 identifiziert werden.[9] Die echte Zuordnung ist also für insgesamt 5018 Betriebe bekannt. Da sich diese Zuordnungen jeweils auf denselben Betrieb beziehen, können Abweichungen in den Ausprägungen der Überschneidungsmerkmale auf die strukturellen Unterschiede zwischen den Erhebungen zurückgeführt werden. Je geringer das Ausmaß der Abweichungen, desto höher ist die Konsistenz eines Merkmals.

Standort. Der Standort unterscheidet sich bei 145 bekannten Zuordnungen (3 %). Es beziehen sich also 145 Zuordnungen jeweils auf denselben Betrieb, obwohl sich die AGS der Zuordnungen in mindes-

[8] In der GLS 1995 gibt es Betriebe mit nur vierstelligen Codes. Diese Codes beziehen sich dann auf Codes der SYPRO, die im Datenmaterial der GLS 1995 teilweise noch Verwendung findet. Da Betriebe mit SYPRO-Codes nicht berücksichtigt werden, sind Verwechselungen ausgeschlossen.

[9] Es wurden 649 Betriebe in der GLS 1995 in den WZ93-Abteilungen 10 bis 37 identifiziert, die weder im Datenmaterial des MBB noch im Datenmaterial der KBE 1995 zu finden sind.

tens einer Ziffer unterscheiden. Die ersten beiden Ziffern (Bundesland) sind in keinem der Fälle betroffen. In 134 Fällen unterscheiden sich die AGS in der dritten Ziffer. In 6 Fällen sind die vierte oder fünfte Ziffer betroffen und in 14 Fällen mindestens eine der letzten drei Ziffern. Bei genauerer Betrachtung der 134 bekannten Zuordnungen, die sich im Regierungsbezirk unterscheiden, fällt auf, dass sich diese Zuordnungen ausschließlich auf Betriebe in Sachsen beziehen. Das Statistische Landesamt des Freistaats Sachsen verwendet seit Januar 1996 die Ebene des Regierungsbezirks im AGS. Davor wurde diese Verwaltungseinheit nicht genutzt. Da die GLS 1995 im Jahr 1996 rückwirkend für Oktober 1995 durchgeführt wurde, werden in der GLS 1995 teilweise die neuen Schlüssel verwendet. Aber auch im MBB und in der KBE 1995 werden teilweise die neuen Schlüssel verwendet. Alle 134 Abweichungen können so erklärt werden. Die verbleibenden Abweichungen im Kreis und in der Gemeinde lassen sich vermutlich auf Interpolationsfehler infolge von Betriebsverlagerungen oder kleineren Kreis- und Gemeindereformen zurückführen. Da bis auf wenige Abweichungen alle Inkonsistenzen erklärt werden können, ist die Konsistenz des Merkmals „Standorts“ sehr hoch.

Wirtschaftszweig. Der Wirtschaftszweig unterscheidet sich bei 417 bekannten Zuordnungen (8 %). In 196 Fällen ist eine der ersten beiden Ziffern des WZ93-Codes betroffen. In 289 Fällen ist die dritte Ziffer betroffen und in 309 die vierte Ziffer. Die Anzahl der Abweichungen nimmt zu, je ausdifferenzierter der Code ist. Ein wirkliches Muster ist in den Abweichungen jedoch nicht zu erkennen. Da die Umschlüsselung zwischen WZ93 und SYPRO nicht immer eindeutig möglich ist, könnten die Inkonsistenzen die Folge einer nicht eindeutigen Umschlüsselung zwischen den beiden Klassifikationssystemen sein. Es könnte möglich sein, dass eine Zuordnung mit abweichenden WZ93-Codes einen identischen SYPRO-Code besitzt, dieser aber je nach Erhebung in unterschiedliche WZ93-Codes umgeschlüsselt wurde. Für diese Vermutung spricht, dass vier Betriebe in der GLS 1995 mit vorhandener originaler Betriebsnummer nicht in den WZ93-Abteilungen 10 bis 37 tätig sind, während dieselben Betriebe im Datenmaterial des MBB und der KBE 1995 enthalten sind – demnach also in den WZ93-Abteilungen 10 bis 37 tätig sind. Aufgrund dieser Probleme ist die Konsistenz des Merkmals „Wirtschaftszweig“ deutlich schlechter als die des Merkmals „Standort“.

Handwerksrolleneintrag. Der Handwerksrolleneintrag unterscheidet sich bei 5 bekannten Zuordnungen (<1 %). Aufgrund der geringen Fallzahl lässt sich kein Muster bei den Abweichungen erkennen. Die Konsistenz des Merkmals „Handwerksrolleneintrag“ ist sehr hoch.

Anzahl der Beschäftigten/tätigen Personen. Differenzen in der Anzahl der Beschäftigten und der Anzahl der tätigen Personen eines Betriebs sind angesichts der abweichenden Merkmalsdefinition und abweichender Erhebungszeiträume nicht überraschend. Da im MBB und der KBE 1995 alle tätigen Personen eines Betriebs erhoben wurden, sollte bei den bekannten Zuordnungen die Anzahl der tätigen Personen höher sein als die Anzahl der Beschäftigten. Bei 3757 bekannten Zuordnungen (75 %) ist das auch tatsächlich der Fall. Die absoluten Differenzen aus der Anzahl der Beschäftigten und der Anzahl der tätigen Personen (im Folgenden Beschäftigtendifferenz) sind teils enorm. Andererseits sind die Beschäftigtendifferenzen bei einigen bekannten Zuordnungen auch unerwartet gering (siehe Tabelle 1).

Tabelle 1: Deskriptive Statistiken für die Verteilung der absoluten Differenz aus der Anzahl der tätigen Personen und der Anzahl der Beschäftigten für die bekannten Zuordnungen.

Zuordnungen zu	n	Mittelwert	Std. Abw.	Min.$^{\varnothing}$	Max.$^{\varnothing}$	Q25	Q50	Q75
MBB 1995	4367	47.92	240.15	0.00	6223.75	3.08	10.08	31.50
KBE 1995	651	2.98	4.77	0.00	44.33	1.00	2.00	3.50
Insgesamt	5018	42.09	224.54	0.00	6223.75	2.00	7.83	26.00

Tabelle 2: Deskriptive Statistiken für die Verteilung der relativen Differenz aus der Anzahl der tätigen Personen und der Anzahl der Beschäftigten für die bekannten Zuordnungen.

Zuordnungen zu	n	Mittelwert	Std. Abw.	Min.$^{\varnothing}$	Max.$^{\varnothing}$	Q25	Q50	Q75
MBB 1995	4367	0.16	0.22	0.00	1.00	0.03	0.08	0.17
KBE 1995	651	0.21	0.27	0.00	1.00	0.05	0.11	0.23
Insgesamt	5018	0.16	0.23	0.00	1.00	0.03	0.08	0.18

Da Betriebe in der KBE 1995 in der Regel weniger als 20 tätige Personen ausweisen, ist die Größenordnung der Beschäftigtendifferenz bei Zuordnungen zu Betrieben in der KBE 1995 – verglichen mit Zuordnungen zu Betrieben im MBB 1995 – deutlich geringer. Interessant ist, dass es 175 Zuordnungen gibt, bei denen die Anzahl der Beschäftigten trotz abweichender Merkmalsdefinition und abweichender Erhebungszeiträume exakt übereinstimmt. Da die Größenordnung der Abweichungen auch in Relation zur Betriebsgröße betrachtet werden sollte, sind in Tabelle 2 die gleichen deskriptiven Statistiken für die Verteilung der relativen Beschäftigtendifferenz aufgeführt.[10] Bei Verwendung der relativen Differenzen verschwinden die Unterschiede im Hinblick auf die Größenordnung zwischen Zuordnungen zu Betrieben im MBB und zu Betrieben in der KBE 1995. Bei 225 bekannten Zuordnungen ist die relative Beschäftigtendifferenz maximal. Dies liegt in allen Fällen daran, dass der zugehörige Betrieb in der GLS 1995 keine Beschäftigten ausweist. Ein Muster lässt sich weder in den absoluten noch in den relativen Beschäftigtendifferenzen erkennen. Es ist weder möglich, den Einfluss der abweichenden Merkmalsdefinition noch den Einfluss der abweichenden Erhebungszeiträume zu isolieren. Es gibt Zuordnungen, bei denen die Beschäftigtendifferenzen enorm hoch sind. Gleichzeitig gibt es Zuordnungen, bei denen sie äußerst gering sind. Möglicherweise heben sich die Auswirkungen der abweichenden Merkmalsdefinition mit den Auswirkungen der abweichenden Erhebungszeiträume bei manchen Zuordnungen gegenseitig auf, während sie sich bei anderen gegenseitig verstärken. Die Konsistenz des Merkmals „Anzahl der Beschäftigten/tätigen Personen“ ist deshalb nur gering.

4 Verknüpfungsvorschlag

Sei $\mathcal{Z}$ die Menge aller potenziellen Zuordnungen der Betriebe in der GLS 1995 mit Betrieben im MBB und in der KBE 1995. Formal entspricht die Verknüpfung der GLS 1995 mit dem MBB und der KBE 1995

[10] Die relative Beschäftigtendifferenz einer Zuordnung ist definiert als die absolute Beschäftigtendifferenz relativ zum Maximum aus der Anzahl der tätigen Personen und der Anzahl der Beschäftigten.

auf Betriebsebene einer Partitionierung von Z in zwei disjunkte Mengen M und U. Dabei soll M alle Zuordnungen enthalten, die sich auf denselben Betrieb beziehen, und U das Komplement von M (in Z) sein. Gesucht ist eine Entscheidungsregel, die Z geeignet in M und U zerlegt. Im Folgenden wird eine deterministische Entscheidungsregel gewählt.[11]

4.1 Blocking-Schritt

Da Z fast 1.5 Milliarden potenzielle Zuordnungen enthält, ist die rechnerische Darstellung von Z nicht möglich. Die meisten potenziellen Zuordnungen in Z beziehen sich jedoch definitionsbedingt nicht auf denselben Betrieb. Im sogenannten Blocking-Schritt wird Z daher darstellbar gemacht, indem Z in zwei disjunkte Mengen Z_M und Z_U zerlegt wird. Dabei soll Z_U eine möglichst große Teilmenge von U sein, sodass Z_M rechnerisch darstellbar ist. Die eigentliche Verknüpfung erfolgt dann nur für die geblockten Zuordnungen in Z_M. Die Darstellung von Z_U ist nicht nötig.

Für den Blocking-Schritt werden Überschneidungsmerkmale benötigt. Um die Anzahl von potenziellen Zuordnungen effektiv zu reduzieren – also eine möglichst große Teilmenge von U zu wählen –, werden Merkmale mit hoher Trennschärfe benötigt. Um zu vermeiden, dass Zuordnungen, die sich auf denselben Betrieb beziehen, fälschlicherweise Z_U zugewiesen werden, sollten die verwendeten Merkmale eine hohe Konsistenz besitzen. Die Merkmale „Standort“ und „Handwerksrolleneintrag“ besitzen beide eine hohe Konsistenz. Allerdings ist die Trennschärfe des Merkmals „Handwerksrolleneintrag“ nur gering, sodass weitere Merkmale benötigt werden. Da nur noch die Merkmale „Wirtschaftszweig“ und „Anzahl der Beschäftigten/tätigen Personen“ als Überschneidungsmerkmale vorhanden sind, wird das Merkmal „Wirtschaftszweig“ gewählt.

Es werden drei Varianten getestet. Potenzielle Zuordnungen werden Z_M zugewiesen, falls der Standort und der Handwerksrolleneintrag übereinstimmen.[12] In Variante V1 wird gefordert, dass zusätzlich auch der WZ93-Code in allen vier Ziffern übereinstimmen muss. Z_M enthät bei dieser Variante 51 240 potenzielle Zuordnungen. In Variante V2 wird gefordert, dass die ersten drei Ziffern des Codes übereinstimmen müssen und in Variante V3, dass nur die ersten beiden Ziffern übereinstimmen müssen. Z_M enthält bei diesen beiden Varianten 77 839 bzw. 172 282 potenzielle Zuordnungen.

4.2 Linkage-Schritt

Im Linkage-Schritt wird eine potenzielle Zuordnung in Z_M genau dann M zugewiesen, wenn die Beschäftigtendifferenz unter allen verbleibenden Zuordnungen desselben Betriebs in der GLS 1995 minimal ist.

[11] Neben der deterministischen Entscheidungsregel wurde auch die von Fellegi und Sunter [5] entwickelte probabilistische Entscheidungsregel getestet. Bei dieser Entscheidungsregel wird eine Verteilung für die Abweichungen in den Ausprägungen der Überschneidungsmerkmale unterstellt. Auf Basis der Verteilung wird für jede potenzielle Zuordnung ein Score ermittelt. Übersteigt der Score einen Schwellenwert, wird die Zuordnung M zugewiesen. Ansonsten wird die Zuordnung U zugewiesen. Aufgrund der unsystematisch auftretenden Abweichungen konnte eine Verteilung nicht verlässlich spezifiziert werden und die Entscheidungsregel keine überzeugenden Ergebnisse liefern. Dieser Ansatz wurde daher nicht weiter verfolgt.

[12] Für Zuordnungen, die sich auf Betriebe in Sachsen beziehen, wird die dritte Ziffer des AGS aufgrund der zuvor erläuterten Problematik nicht berücksichtigt.

Eine geblockte Zuordnung wird also genau dann M zugewiesen, wenn gilt:

$$\Delta(i,j) = \min_k \Delta(i,k),$$

wobei $\Delta(i,j)$ die Beschäftigtendifferenz der Zuordnung des $i$ten Betriebs in der GLS 1995 mit dem $j$ten Betrieb im MBB oder in der KBE 1995 bezeichnet. Die Nummerierung bezieht sich dabei immer auf die geblockten Zuordnungen. Existieren mehrere Betriebe in der GLS 1995, die die Bedingung erfüllen, werden sie alle M zugewiesen.

Diese Entscheidungsregel ist angelehnt an ein Vorgängerprojekt von Hafner [1]. In diesem Projekt wurde jedoch nur eine Verknüpfung der GLS 1995 mit dem MBB 1995 angestrebt. Betriebe in der GLS 1995 mit weniger als 20 Beschäftigten wurden bei diesem Projekt nicht berücksichtigt. Aufgrund der verhältnismäßig guten Resultate, die Hafner [1] mit dieser Entscheidungsregel erzielen konnte, wird die Entscheidungsregel auch in diesem Projekt verwendet.

4.3 Ergebnisse

Der vorgeschlagene Verknüpfungsansatz wird im Folgenden anhand der 5018 Betriebe, für die die echte Zuordnung bekannt ist, evaluiert. Als Performance-Maße werden Precision, Recall und F1 verwendet. Der Recall-Wert ist definiert als:

$$\text{Recall} = \frac{\text{RP}}{\text{RP} + \text{FN}},$$

wobei RP die Anzahl der richtig-positiven Zuordnungen bezeichnet und FN die Anzahl der falsch-negativen Zuordnungen. Die richtig-positiven Zuordnungen sind die Zuweisungen zu M, für die bekannt ist, dass sie sich auf denselben Betrieb beziehen. Analog dazu beschreiben die falsch-negativen Zuordnungen die Zuweisungen zu U, für die ebenfalls bekannt ist, dass sie sich auf denselben Betrieb beziehen. Der Recall-Wert misst also, wie gut richtige Zuordnungen von der Entscheidungsregel identifiziert werden können. Der Precision-Wert dagegen misst, wie verlässlich richtige Zuordnungen von der Entscheidungsregel tatsächlich identifiziert werden können. Der Precision-Wert ist definiert als:

$$\text{Precision} = \frac{\text{RP}}{\text{RP} + \text{FP}},$$

wobei FP die Anzahl der falsch-positiven Zuordnungen bezeichnet. Die falsch-positiven Zuordnungen sind die Zuweisungen zu M, für die bekannt ist, dass sie sich nicht auf denselben Betrieb beziehen. Während die richtig-positiven und die falsch-negativen Zuordnungen direkt aus den 5018 bekannten Zuordnungen ermittelt werden können, müssen zur Ermittlung der falsch-positiven Zuordnungen erst (Pseudo-)Fehlzuordnungen konstruiert werden. Wird ein Betrieb in der GLS 1995, für den eine originale Betriebsnummer vorhanden ist, Betrieben mit einer davon abweichenden Betriebsnummer im MBB und in der KBE 1995 zugeordnet, so handelt es sich dabei per Konstruktion um Zuordnungen, die sich nicht auf denselben Betrieb beziehen. Analog dazu kann ein Betrieb im MBB oder in der KBE 1995, für den die systemfreie Betriebsnummer der GLS 1995 vorhanden ist, Betrieben mit einer davon abweichenden Nummer in der GLS 1995 zugeordnet werden. Bei diesen Zuordnungen handelt es sich dann eben-

falls per Konstruktion um Fehlzuordnungen. Die so konstruierten Zuordnungen können zur Ermittlung der falsch-positiven Zuordnungen verwendet werden. Da sich Precision und Recall gegenseitig bedingen, wird außerdem F1 für die Bewertung herangezogen. Der F1-Wert ist das Harmonische Mittel aus Precision- und Recall-Wert, das wie folgt definiert ist:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Es konnten 12 905 Betriebe in der GLS 1995 in V1 zugeordnet werden. In V2 und V3 sind es mit 13 342 bzw. 14 121 etwas mehr. Die beschriebenen Performance-Maße sind für jede der Varianten in Tabelle 3 aufgeführt. Die Anzahl der richtig-positiven Zuordnungen ist bei V2 knapp vor V1 am höchsten. Dem-

Tabelle 3: Performance-Maße für alle drei Varianten des Verknüpfungsvorschlags.

Variante	Zuordnungen zu	n	EP	FN	FP	Precision	Recall	F1
V1	MBB 1995	11 438	3837	530	118	0.97	0.78	0.92
	KBE 1995	1467	505	146	14	0.97	0.88	0.86
	Insgesamt	12 905	4342	676	132	0.97	0.87	0.91
V2	MBB 1995	11 625	3826	541	146	0.96	0.88	0.92
	KBE 1995	1717	526	125	19	0.97	0.81	0.88
	Insgesamt	13 342	4352	666	165	0.96	0.86	0.91
V3	MBB 1995	11 957	3720	647	220	0.94	0.85	0.89
	KBE 1995	2164	517	134	28	0.95	0.79	0.86
	Insgesamt	14 121	4237	781	248	0.94	0.84	0.89

entsprechend ist auch die Anzahl der falsch-negativen Zuordnungen bei V2 leicht geringer. Die Anzahl der falsch-positiven Zuordnungen ist bei V2 etwas höher als bei V1. Insgesamt sind die Unterschiede zwischen V1 und V2 in Bezug auf die Peformance-Maße aber nur gering. V3 sticht als schlechteste Va-riante klar hervor. Die Anzahl der richtig-positiven Zuordnungen ist geringer als bei V1 und die Anzahl der falsch-positiven Zuordnungen ist bei V3 von allen drei Varianten am höchsten. Es fällt außerdem auf, dass der Recall-Wert für Zuordnungen zu Betrieben in der KBE 1995 bei allen drei Varianten deutlich unter dem Recall-Wert für Zuordnungen zu Betrieben im MBB 1995 liegt. Die Precision-Werte unter-scheiden sich zwischen MBB und KBE 1995 dagegen kaum.

Ein Blick auf die falsch-positiven Zuordnungen zeigt, dass sie alle die Folge geringer Beschäftigtendifferenzen sind. Sie entstehen, wenn eine Fehlzuordnung – verglichen mit der richtigen Zuordnung – eine geringere Beschäftigtendifferenz aufweist. Die falsch-negativen Zuordnungen sind überwiegend die Folge einer hohen Beschäftigtendifferenz bei der richtigen Zuordnung. Die übrigen falsch-negativen Zuordnungen entstehen bereits im Blocking-Schritt aufgrund von Abweichungen im Standort, dem Wirtschaftszweig oder dem Handwerksrolleneintrag. Bei V1 entstehen 431 falsch-negative Zuordnungen bereits im Blocking-Schritt. Bei V2 und V3 sind es 303 bzw. 213.

Ein Problem, das von den Performance-Maßen nicht berücksichtigt wird, sind Mehrfachzuordnungen.

Es werden zwei Arten von Mehrfachzuordnungen unterschieden: i) ein Betrieb in der GLS 1995, der mehreren Betrieben im MBB und in der KBE 1995 zugeordnet wurde und ii)) ein Betrieb im MBB oder in der KBE 1995, dem mehrere Betriebe in der GLS 1995 zugeordnet wurden. Der Anteil an Mehrfachzuordnungen von Betrieben in der GLS 1995 steigt mit der Variante. Der Anteil liegt bei V1 bei 2.7 % und steigt bis 6.7 % bei V3. Zuordnungen zu Betrieben in der KBE 1995 sind deutlich häufiger betroffen. Hier reicht der Anteil von 18.7 % in V1 bis 35.7 % in V3. Die Anzahl der Zuordnungen pro Mehrfachzuordnung liegt je nach Variante im Mittel zwischen 3 und 4. Im Extremfall sind es etwas unter 100 bei V1 bis etwas unter 200 Zuordnungen bei V3. Die Ursachen für diese Art der Mehrfachzuordnungen sind überwiegend identische Beschäftigtendifferenzen und Beschäftigtendifferenzen von 0. Die andere Art von Mehrfachzuordnungen bezieht sich auf Betriebe im MBB und in der KBE 1995, denen mehrere Betriebe in der GLS 1995 zugeordnet wurden. Auch hier steigt der Anteil der Mehrfachzuordnungen mit der Variante. Der Anteil bei V1 liegt bei 7.9 % und reicht bis 13.6 % bei V3. Unterschiede zwischen Zuordnungen zu Betrieben im MBB und Zuordnungen zu Betrieben in der KBE 1995 gibt es kaum. Die Anzahl der Zuordnungen pro Mehrfachzuordnung liegt je nach Variante im Mittel zwischen 2 und 3. Im Extremfall sind es etwa 10 bei V1 und etwas mehr als 20 bei V3. Die Ursache für diese Art von Mehrfachzuordnungen ist in der Regel, dass es für mehrere Betriebe in der GLS 1995 nur einen einzigen passenden Betrieb im MBB oder in der KBE 1995 gibt.

Zusammenfassend lässt sich sagen, dass V1 die beste Variante im Hinblick auf die Performance-Maße und die Variante mit dem geringsten Anteil an Mehrfachzuordnungen ist. Da vermutet wurde, dass die Ergebnisse stark von der Betriebsgröße abhängen, wurden die Betriebe in neun Beschäftigtengrößenklassen eingeteilt und Precision-, Recall- und F1-Werte für jede der Klassen ermittelt. Die Klassen wurden dabei wie folgt gebildet:

Anzahl der Beschäftigten	0	in Klasse	0
	mehr als 0, aber weniger als 10		1
	mehr als 10, aber weniger als 20		2
	mindestens 20, aber weniger als 50		3
	mindestens 50, aber weniger als 100		4
	mindestens 100, aber weniger als 250		5
	mindestens 250, aber weniger als 500		6
	mindestens 500, aber weniger als 1000		7
	mindestens 1000		8

Maßgeblich war immer die Anzahl der Beschäftigten in der GLS 1995. Abbildung 1 stellt Precision-, Recall- und F1-Werte in Abhängigkeit der Beschäftigtengrößenklassen für jede der drei Varianten dar. Die Precision-Werte sind bei allen drei Varianten konstant sehr hoch bei deutlich über 0.90. Bei V3 schwankt der Wert etwas mehr. Die Recall-Werte nehmen bei allen drei Varianten mit steigender Beschäftigtengrößenklasse zu. Je höher die Anzahl der Beschäftigten in der GLS 1995, desto höhere Recall-Werte werden erzielt. Während der Recall-Wert insgesamt zwischen 0.84 und 0.87 liegt, gibt es Klassen, bei denen der Wert bei über 0.90 liegt. In kleineren Klassen liegt der Wert allerdings auch deutlich unter 0.80. Ab Klasse 2 liegt der F1-Wert bei allen Varianten bei über 0.90. Auch unter diesem Gesichtspunkt

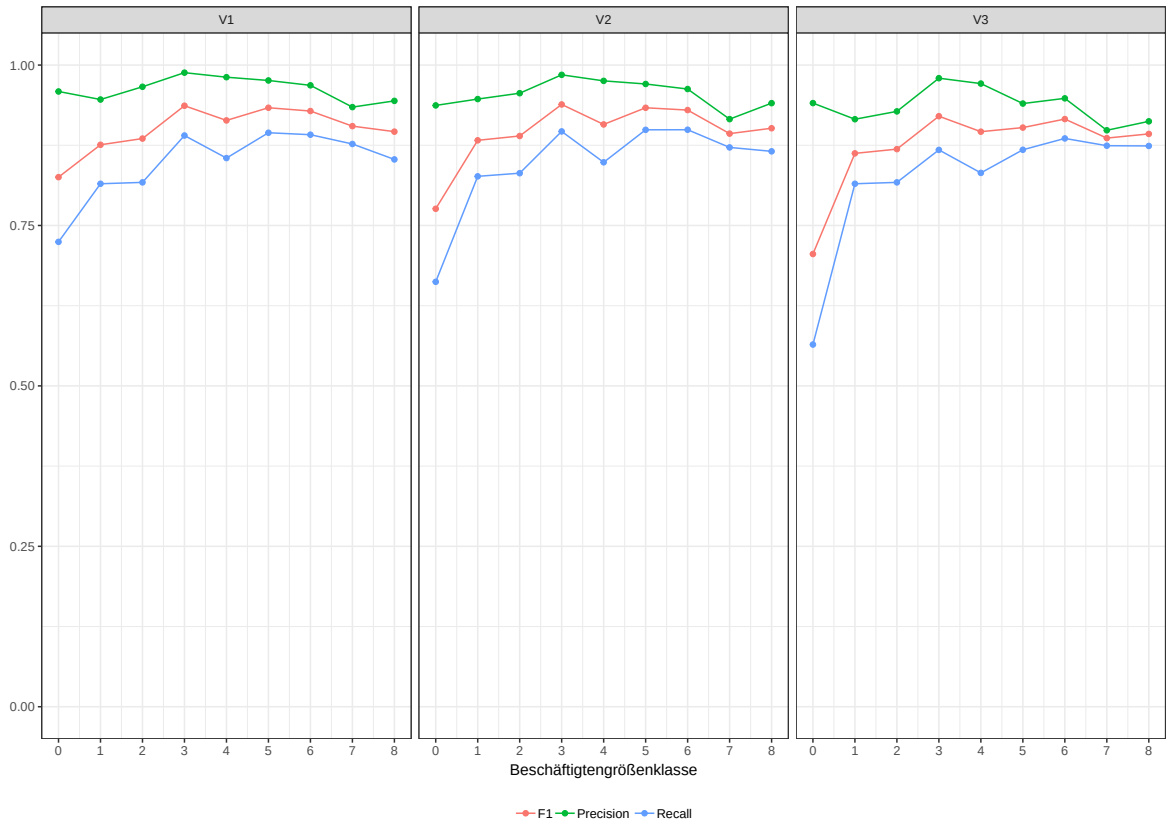


Abbildung 1: Precision-, Recall- und F1-Werte des Verknüpfungsvorschlags in Abhängigkeit der neun Beschäftigtengrößenklassen.

erscheint V1 – also die Zuordnung über Standort, Handwerksrolleneintrag sowie den vierstelligen WZ93-Code – als die beste Variante.

5 Abschließende Bemerkungen

Trotz der geringen Anzahl von Überschneidungsmerkmalen und der strukturellen Unterschiede zwischen den Erhebungen, die zum Teil zu erheblichen unsystematischen Abweichungen in den Merkmalsausprägungen führen, konnten mit dem vorgeschlagenen Record-Linkage-Ansatz gute Ergebnisse erzielt werden. Für einige Beschäftigtengrößenklassen wurden Precision- und Recall-Werte von über 90 % erzielt. Ungelöst blieb die Frage nach dem Umgang mit Mehrfachzuordnungen.

Aufgrund der geringen Anzahl von Überschneidungsmerkmalen, die teils nur eine geringe Trennschärfe besitzen, und der strukturellen Unterschieden zwischen den Erhebungen ist es fraglich, ob wesentlich bessere Verknüpfungen überhaupt möglich sind. Für einige Anwendungen ist eine Verknüpfung auf Basis des Standorts und des Wirtschaftszweigs möglicherweise ausreichend, weshalb neben dem Verknüpfungsvorschlag auch die geblockten Zuordnungen bereitgestellt werden.

Um die Anzahl der Überschneidungsmerkmale zu erhöhen, wurde versucht, zwei weitere Merkmale aus

dem vorhandenen Datenmaterial zu konstruieren: i) ein Indikator, der angibt, ob es sich bei einem Betrieb um den Betrieb eines Mehrbetriebsunternehmens handelt, und ii) einen Indikator, der überprüft, ob die Lohnkosten einer potenziellen Zuordnung signifikant höher als der erwirtschaftete Umsatz sind. Die Konstruktion misslang jedoch in beiden Fällen, da die vorhandene Information in den Daten nicht ausreichend ist, um die beiden Merkmale verlässlich zu konstruieren.

Es gab darüber hinaus den Ansatz, die Inkonsistenzen im Wirtschaftszweig zu erklären, indem für jeden WZ93-Code die Menge der SYPRO-Codes gebildet wurde, die in den entsprechenden WZ93-Code umgeschlüsselt werden. Neben den WZ93-Codes sollte dann auch immer die Menge der SYPRO-Codes für eine potenzielle Zuordnung verglichen werden. Ist der Schnitt der Mengen nicht leer, so muss es einen SYPRO-Code geben, der in mehrere WZ93-Codes umgeschlüsselt wird. Abweichende WZ93-Codes können dann die Folge einer abweichenden Umschlüsselung der Codes sein. Da die Umsteigeschlüssel zwischen den Klassifikationssystemen nicht (mehr) vorhanden sind, hätten die alten und neuen Schlüssel erst für alle Wirtschaftszweige recherchiert werden müssen. Aufgrund des enorm hohen Arbeitsaufwands und der unbekannten Erfolgswahrscheinlichkeit wurde dieser Ansatz aber letztendlich wieder verworfen.

Grundsätzlich könnte die GLS 1995 durch die in diesem Projekt erfolgte Zuordnung der bisher fehlenden Betriebsnummern für Auswertungen im FDZ verwendet werden. Die nun gegebene Verknüpfbarkeit mit weiteren Wirtschaftsstatistiken erhöht das Analysepotenzial der GLS deutlich. So ist zum Beispiel die Verknüpfung mit dem Datenmaterial der GLS 2001 und der nachfolgenden Verdienststrukturerhebungen zu einem Betriebs-Panel ab 1995 möglich.

Referenzen

[1] Hans-Peter Hafner. “Matching der Daten der Gehalts- und Lohnstrukturerhebung 1995 mit dem Monatsbericht im Verarbeitenden Gewerbe 1995”. (2008). Nur für den internen Gebrauch vorgesehen. (Siehe S. 6, 14).

[2] Peter Kaukewitsch. “Ergebnisse der Gehalts- und Lohnstrukturerhebung 1996 für 1995”. *Wirtschaft und Statistik* 1 (1998) (siehe S. 7, 8).

[3] Forschungsdatenzentrum der Statistischen Landesämter. *Gehalts- und Lohnstrukturerhebung im Produzierenden Gewerbe und im Dienstleistungsbereich. Erhebungen 1990 bis 2001*. EVAS-Nummer 62111. Metadatenreport (siehe S. 8).

[4] Forschungsdatenzentrum der Statistischen Landesämter. *Allgemeine und methodische Informationen zum AFiD-Panel Industriebetriebe 1995 bis 2016*. EVAS-Nummer 42111. Metadatenreport (siehe S. 8).

[5] Ivan Fellegi und Alan Sunter. “A theory for record linkage”. *Journal of the American Statistical Association* 64.328 (1969) (siehe S. 13).

A Übersicht über die bereitgestellten Daten

Für mögliche spätere Nutzungen werden die vorgeschlagene Verknüpfung sowie die geblockten Zuordnungen bereitgestellt. Darüber hinaus enthält das bereitgestellte Datenmaterial die Information aus den Leitbändern.[13] Da mit Variante V1 des Verknüpfungsvorschlags die besten Ergebnisse erzielt wurden, wird nur diese Variante bereitgestellt. Neben der systemfreien Nummer der GLS 1995 und der originalen Betriebsnummer des URS enthält das bereitgestellte Datenmaterial noch weitere Merkmale. In Tabelle 4 sind alle enthaltenen Merkmale mit Beschreibung aufgelistet.

Tabelle 4: Merkmalsliste des bereitgestellten Datenmaterials.

Handle	Datentyp	Merkmal	Bemerkung
`idgls`	<char>	systemfreie Nummer der GLS 1995	Die siebenstellige Nummer wird zur eindeutigen Identifikation der Betriebe in der GLS 1995 verwendet.
`idurs`	<char>	originale Betriebsnummer des URS	Die elfstellige Nummer wird zur eindeutigen Identifikation der Betriebe im URS verwendet.
`V1`	<char>	AGS des Betriebs	Die dritte Ziffer des Codes für Betriebe aus Sachsen ist problematisch.
`V2`	<char>	NACE-Code des Betriebs	Der Code enthält nur die ersten vier Stellen des WZ93-Codes. Die vierstelligen Codes entsprechen den NACE-Codes.
`V4`	<real>	Anzahl der tätigen Personen	Die (interpolierte) Anzahl der tätigen Personen im MBB und in der KBE 1995.
`V4`	<bool>	Indikator für MBB oder KBE 1995	Der Indikator gibt an, ob der Betrieb im Datenmaterial des MBB oder der KBE 1995 enthalten ist.
`Delta`	<real>	Beschäftigtendifferenz	Differenz aus der Anzahl der Beschäftigten in der GLS 1995 und der (interpolierten) Anzahl der tätigen Personen im MBB und in der KBE 1995.
`Match`	<int>	Indikator für Zuordnungstyp	`=1`, wenn es sich um eine echt-positive Zuordnung handelt; `=0`, wenn es sich um eine falsch-positive Zuordnung handelt; `=-1`, wenn es sich um eine falsch-negative Zuordnung handelt. Die Angabe fehlt, wenn aus den Leitbändern keine Information über eine Zuordnung ermittelt werden konnte.

[13] Da das Datenmaterial auch die falsch-negativen Zuordnungen enthält, ist die Anzahl an Zuordnungen natürlich etwas höher als in Abschnitt 4 angegeben.

B Code

Der hier dokumentierte R-Code wurde unter Version 3.5.0 erstellt. Als zusätzliches Paket wurde `tidyverse` in der Version 1.2.0 verwendet.

Einlesen des Datenmaterials

Um das Datenmaterial einzulesen, wurde die globale Variable `PATH_DATA` erstellt. Die `list`-Variable enthält die vollständigen Pfade zu den Rohdaten der GLS 1995, des MBB und der KBE 1995 sowie zu den Leitbändern. Noch während des Einlesens werden die später benötigten Überschneidungsmerkmale angepasst und gegebenenfalls repariert. Für jeden Datensatz wurde eine eigene Funktion geschrieben:

```
readdata = list(
        function(df) mutate(read_csv2(df), # für die GLS 1995
            ID.x = str_c(str_pad(,EF1U1 2, "left", "0"), str_pad(EF1U2, 5, "left", "0")),
            V0   = -1,
            V1   = str_c(EF5U1, EF5U2, EF5U3, EF5U4),
            V2   = str_c(case_when(str_length(EF14) == 5) ~ str_trunc(EF11, 4, "right", ""), TRUE ~
                NA_character_),
            V3   = str_c(EF7),
            V4   = as.numeric(EF14 + EF15 + EF16 + EF17)
        ),
        function(df) mutate(read_csv2(df), # für den MBB 1995
            ID.y = str_c(str_pad(BNR, 11, "left", "0"))
            V0   = 1,
            V1   = str_c(mb_06),
            V2   = str_c(mb_08),
            V3   = str_c(mb_07),
            V4   = as.numeric(mb14 / mb_10)
        ),
        function(df) mutate(read_csv2(df), # für die KBE 1995
            ID.y = str_c(str_pad(BNR, 11, "left", "0"))
            V0   = 0,
            V1   = str_c(K05_1995),
            V2   = str_c(K07_1995),
            V3   = str_c(K06_1995),
            V4   = as.numeric(K08_1995)
        ),
        function(df) mutate(read_csv2(df), # für die Leitbänder
            ID.x = str_c(str_pad(id, 6, "left", "0")),
            ID.y = str_c(str_pad(urs, 11, "left", "0"))
        )
    )
```

Vor allem im Datenmaterial der GLS 1995 müssen Merkmale angepasst werden. Die einzelnen Merkmalsausprägungen des AGS müssen aus den Ausprägungen mehrerer Merkmale zusammengefügt werden. Der WZ93-Code wird auf vier Stellen gekürzt und alle SYPRO-Codes (erkennbar an der Codelänge) werden entfernt. Die systemfreie Nummer in der GLS 1995 sowie die Betriebsnummer im MBB und in der KBE 1995 werden mit führenden Nullen auf sieben bzw. elf Stellen gebracht. Das gilt auch für das Datenmaterial der Leitbänder.

Das Einlesen erfolgt anschließend mit:

```
raw = PATH_DATA %>%
    map2(readdata, ~.x %>% y)
```

Pre-Processing des Datenmaterials

Im Pre-Processing-Schritt sollen fehlende und fehlerhafte Werte entfernt werden. Dazu wird die Funktion:

```
cleandata = function(df, id) drop_na(transmute(df,
        !!!id,
        V0 = V0,
        V1 = case_when(str_length(V1) == 8 ~ V1, TRUE ~ NA_character_),
        V2 = case_when(str_length(V2) == 4 ~ V2, TRUE ~ NA_character_),
        V3 = case_when(str_length(V3) %in% seq(0, 3) ~ V3, TRUE ~ NA_character_),
        V4 = case_when(V4 >= 0 ~ V4, TRUE ~ NA_real_)
    ))
```

verwendet. Merkmalsausprägungen, die nicht der vorgegebenen Länge entsprechen, werden als fehlende Werte klassifiziert und entfernt. Das Pre-Processing der Daten erfolgt mit:

```
gls1995_full = raw %>% pluck(1) %>%
    cleandata("ID.x")

mbb1995 = raw %>% pluck(2) %>%
    cleandata("ID.y")

kbe1995 = raw %>% pluck(3) %>%
    cleandata("ID.y")

lbr1995 = raw %>% pluck(4) %>%
    transmute(ID.x, ID.y)
```

Da sich die Verknüpfung auf Betriebe in der GLS 1995 beschränkt, die in den WZ93-Abteilungen 10 bis 37 tätig sind, werden die übrigen Betriebe mit:

```
gls1995 = gls1995_full %>%
    filter(str_sub(V2, 0, 2) %in% seq(10, 37))
```

entfernt. Außerdem wird das Datenmaterial des MBB und der KBE 1995 mit:

```
mbx1995 = bind_rows(mbb1995, kbe1995)
```

zusammengefügt.

Analyse der bekannten Zuordnungen

Neben den bekannten Zuordnungen werden auch gleich die (Pseudo-)Fehlzuordnungen konstruiert. Die Menge aller möglichen Kombinationen der beiden Identifikationsmerkmale in den Leitbändern lässt sich in die bekannten Zuordnungen und in bekannte Fehlzuordnungen zerlegen. Diese Menge ist rechnerisch jedoch nicht darstellbar. Die Menge wird daher nur auf Bundeslandebene gebildet. Dazu werden zuerst zwei temporäre Variablen benötigt:

```
a = lbr1995 %>% transmute(ID.x, t = str_sub(ID.x, 0, 2))
b = lbr1995 %>% transmute(ID.y, t = str_sub(ID.x, 0, 2))
```

Die bekannten Zuordnungen sind aus den Leitbändern bekannt und können somit identifiziert werden:

```
M = inner_join(a, b, "t") %>%
    inner_join(lbr1995, c("ID.x", "ID.y")) %>%
    inner_join(gls1995, "ID.x") %>%
    inner_join(mbx1995, "ID.y")
```

Das Entfernen der bekannten Zuordnungen liefert die Fehlzuordnungen:

```
U = inner_join(a, b, "t") %>%
    anti_join(lbr1995, c("ID.x", "ID.y")) %>%
    inner_join(gls1995, "ID.x") %>%
    inner_join(mbx1995, "ID.y")
```

Für die Analyse wird eine Funktion geschrieben, die auch später zur Analyse der Fehlzuordnungen verwendet wird:

```
features = function(df)
{
    transmute(df,
        n = 1,
        X10 = as.numeric(str_sub(V1.x, 0, 8) != str_sub(V1.y, 0, 8)),
        X11 = as.numeric(str_sub(V1.x, 0, 2) != str_sub(V1.y, 0, 2)),
        X12 = as.numeric(str_sub(V1.x, 3, 3) != str_sub(V1.y, 3, 3)),
        X13 = as.numeric(str_sub(V1.x, 4, 5) != str_sub(V1.y, 4, 5)),
        X14 = as.numeric(str_sub(V1.x, 6, 8) != str_sub(V1.y, 6, 8)),
        X20 = as.numeric(str_sub(V2.x, 0, 4) != str_sub(V2.y, 0, 4)),
        X21 = as.numeric(str_sub(V2.x, 0, 2) != str_sub(V2.y, 0, 2)),
        X22 = as.numeric(str_sub(V2.x, 3, 3) != str_sub(V2.y, 3, 3)),
        X23 = as.numeric(str_sub(V2.x, 4, 4) != str_sub(V2.y, 4, 4)),
        X30 = as.numeric(V3.x != V3.y),
        X40 = abs(V4.x - V4.y),
        X41 = X40 / pmax(V4.x, V4.y)
    )
}
```

Die Funktion zählt die Abweichungen in den Ausprägungen der Überschneidungsmerkmale. Die Analyse erfolgt dann sowohl getrennt nach MBB und KBE 1995 als auch gemeinsam für beide Erhebungen mit:

```
list(KBE1995 = filter(M, V0.y == 0), MBB1995 = filter(M, V0.y == 1), MBX1995 = M) %>%
  map(~.x %>% features() %>% descriptives()) %>%
  bind_rows(.id = "Zuordnungen zu")
```

Die Funktion `descriptives` erstellt gängige deskriptive Statistiken. Die Funktion ist wie folgt definiert:

```
descriptives = function(df)
{
    summarize_all(group_by(gather(df, key, val), key),
        list(
            sum    = ~sum(., na.rm = 1),
            mean   = ~mean(., na.rm = 1),
            sd     = ~sd(., na.rm = 1),
            fdzmin = ~mean(head(sort(.), 3), na.rm = 1),
            fdzmax = ~mean(tail(sort(.), 3), na.rm = 1),
```

```
            q25    = ~quantile(., probs = 0.25, na.rm = 1),
            q50    = ~quantile(., probs = 0.50, na.rm = 1),
            q75    = ~quantile(., probs = 0.75, na.rm = 1)
        )
    )
}
```

Blocking-Schritt

Zunächst werden wieder zwei temporäre Variablen erstellt. Die beiden Variablen erlauben das Bilden aller möglichen Zuordnungen der Betriebe auf Basis des AGS ohne Regierungsbezirk und den ersten beiden Ziffern des WZ93 Codes. Das ist nötig, da die dritte Ziffer des AGS für Betriebe aus Sachsen problematisch ist:

```
a = gls1995 %>% mutate(s = sub("^(.{2})[0-9]", "\\1X", V1), t = str_sub(V2, 0, 2))
b = mbx1995 %>% mutate(s = sub("^(.{2})[0-9]", "\\1X", V1), t = str_sub(V2, 0, 2))
```

Anschließend kann der Blocking-Schritt für alle drei Varianten vollzogen werden:

```
B1 = inner_join(a, b, c("s", "t", "V3")) %>%
    filter(V1.x == V1.y | str_sub(V1.x, 0, 2) == 14) %>%
    filter(str_sub(V2.x, 0, 4) == str_sub(V2.y, 0, 4)) %>%
    mutate(V1 = V1.y, V2 = V2.y, V3 = V4.y, V4 = V0.y, Delta = abs(V4.x - V4.y)) %>%
    select(-s, -t)

B2 = inner_join(a, b, c("s", "t", "V3")) %>%
    filter(V1.x == V1.y | str_sub(V1.x, 0, 2) == 14) %>%
    filter(str_sub(V2.x, 0, 3) == str_sub(V2.y, 0, 3)) %>%
    mutate(V1 = V1.y, V2 = V2.y, V3 = V4.y, V4 = V0.y, Delta = abs(V4.x - V4.y)) %>%
    select(-s, -t)

B3 = inner_join(a, b, c("s", "t", "V3")) %>%
    filter(V1.x == V1.y | str_sub(V1.x, 0, 2) == 14) %>%
    filter(str_sub(V2.x, 0, 2) == str_sub(V2.y, 0, 2)) %>%
    mutate(V1 = V1.y, V2 = V2.y, V3 = V4.y, V4 = V0.y, Delta = abs(V4.x - V4.y)) %>%
    select(-s, -t)
```

Falls es Zuordnungen gibt, die sich auf Betriebe in Sachsen beziehen und bei denen alles bis auf den Regierungsbezirk übereinstimmt, werden sie einander zugeordnet. Außerdem werden Merkmale gebildet, die später benötigt werden bzw. im bereitgestellten Datenmaterial enthalten sein sollen (AGS, WZ93-Code, Anzahl der tätigen Personen, Indikator für MBB oder KBE 1995 und die Beschäftigtendifferenz). Die Varianten werden abschließend für die weitere Verwendung in einer `list`-Variablen gespeichert:

```
B = list(V1 = B1, V2 = B2, V3 = B3)
```

Linkage-Schritt

Der Linkage-Schritt ist bei allen drei Varianten identisch:

```
RL = B %>%
    map(~.x %>% group_by(ID.x) %>% filter(Delta == min(Delta)) %>% ungroup())
```

Für jeden Betrieb in der GLS 1995 werden Zuordnungen zu Betrieben im MBB und in der KBE 1995 entfernt, wenn es eine Zuordnung für denselben Betrieb mit geringerer Beschäftigtendifferenz gibt.

Ergebnisse

Um die Verknüpfung zu evaluieren, werden die Performance-Maße Precision, Recall und F1 berechnet. Um diese Maße zu berechnen, werden die Anzahl der richtig-positiven, die Anzahl der falsch-negativen und die Anzahl der falsch-positiven Zuordnungen benötigt. Dazu wird die Funktion:

```
performance = function(df, Q)
{
    tibble(
        m  = nrow(df),
        TP = nrow(semi_join(df, M, c("ID.x", "ID.y"))),
        FP = nrow(semi_join(df, U, c("ID.x", "ID.y"))),
        FN = nrow(anti_join(Q, df, c("ID.x", "ID.y"))),
        Precision = TP / (TP + FP),
        Recall    = TP / (TP + FN),
        F1        = 2 * (Precision * Recall) / (Recall + Precision)
    )
}
```

vewendet. Die richtig-positiven Zuordnungen lassen sich direkt aus einem Vergleich der ermittelten Zuordnungen mit den bekannten Zuordnungen bestimmen. Die falsch-positiven Zuordnungen werden durch einen Vergleich der ermittelten Zuordnungen mit den bekannten Fehlzuordnungen bestimmt. Die bekannten Zuordnungen, die in den ermittelten Zuordnungen fehlen, ergeben die falsch-negativen Zuordnungen. Die Performance-Maße berechnen sich wie zuvor beschrieben.

Die Ergebnisse werden mit:

```
RL %>%
    map(~list(KBE1995 = filter(.x, V0.y == 0), MBB1995 = filter(.x, V0.y == 1), MBX1995 = .x)) %>%
    map(~.x %>% map2(list(KBE1995 = filter(M, V0.y == 0), MBB1995 = filter(M, V0.y == 1), MBX1995 = M),
         ~.x %>% performance(.y)))
    map(~.x %>% bind_rows(.id = "Zuordnungen zu")) %>%
    bind_rows(.id = "Variante")
```

in die gewünschte Form gebracht. Sie werden sowohl getrennt nach MBB und KBE 1995 als auch für beide Erhebungen gemeinsam berechnet. Analog dazu erhält man die Ergebnisse unterteilt in die Beschäftigtengrößenklassen:

```
RL %>%
    map(~.x %>% mutate(class = classes(V4.x)) %>% split(.$class)) %>%
    map(~.x %>% map2(M %>% mutate(class = classes(V4.x)) %>% split(.$class), ~.x %>% performance(.y))
        %>% bind_rows(.id = "Klasse")) %>%
    bind_rows(.id = "Variante") %>%
    select(Variante, Klasse, TP, FP, FN, Precision, Recall, F1) %>%
    group_by(Variante) %>%
    complete(Klasse = str_c(seq(0, 8))) %>%
    fill(Precision, Recall, F1)
```

Die Beschäftigtengrößenklassen werden mit:

```
classes = function(x) case_when(x >= 1000 ~ 8, x >= 500 ~ 7, x >= 250 ~ 6, x >= 100 ~ 5, x >= 50 ~ 4, x
    >= 20 ~ 3, x >= 10 ~ 2, x > 0 ~ 1, x == 0 ~ 0)
```

gebildet.

Analyse der Fehl- und Mehrfachzuordnungen

Die falsch-negativen und falsch-positiven Zuordnungen lassen sich analog zu den bekannten Zuordnungen mit:

```
RL %>%
    map(~list(KBE1995 = filter(.x, V0.y == 0), MBB1995 = filter(.x, V0.y == 1), MBX1995 = .x)) %>%
    map(~.x %>% select(ID.x, ID.y)) %>%
    map(~.x %>% ~inner_join(U, ., c("ID.x", "ID.y"))) %>%
    map(~.x %>% features() %>% descriptives()) %>%
    map(~.x %>% bind_rows(.id = "Zuordnugnen_zu")) %>%
    bind_rows(.id = "Variante")
```

bzw.:

```
RL %>%
    map(~list(KBE1995 = filter(.x, V0.y == 0), MBB1995 = filter(.x, V0.y == 1), MBX1995 = .x)) %>%
    map(~.x %>% select(ID.x, ID.y)) %>%
    map(~.x %>% ~anti_join(M, ., c("ID.x", "ID.y"))) %>%
    map(~.x %>% features() %>% descriptives()) %>%
    map(~.x %>% bind_rows(.id = "Zuordnugnen_zu")) %>%
    bind_rows(.id = "Variante")
```

analysieren. Die Mehrfachzuordnungen werden mit:

```
RL %>%
    map(~list(KBE1995 = filter(.x, V0.y == 0), MBB1995 = filter(.x, V0.y == 1), MBX1995 = .x)) %>%
    map(~.x %>% map(~.x %>% count(ID.x) %>% summarize(
       single = sum(n[which(n == 1)]),
       multi  = sum(n[which(n > 1)]),
       total  = sum(n),
       quote  = multi/total,
       mean   = mean(n[which(n > 1)]),
       sd     = sd(n[which(n > 1)]),
       max    = max(n)))
    ) %>%
    map(~.x %>% bind_rows(.id = "Mehrfachzuordnungen von Betrieben in der GLS 1995")) %>%
    bind_rows(.id = "Variante")
```

bzw.:

```
RL %>%
    map(~list(KBE1995 = filter(.x, V0.y == 0), MBB1995 = filter(.x, V0.y == 1), MBX1995 = .x)) %>%
    map(~.x %>% map(~.x %>% count(ID.y) %>% summarize(
       single = sum(n[which(n == 1)]),
       multi  = sum(n[which(n > 1)]),
       total  = sum(n),
       quote  = multi/total,
       mean   = mean(n[which(n > 1)]),
       sd     = sd(n[which(n > 1)]),
       max    = max(n)))
    ) %>%
```

```
    map(~.x %>% bind_rows(.id = "Mehrfachzuordnungen zu Betrieben im MBB/KBE 1995")) %>%
    bind_rows(.id = "Variante")
```

untersucht. Es wird nach Mehrfachzuordnungen von Betrieben in der GLS 1995 und Mehrfachzuordnungen zu Betrieben im MBB und in der KBE 1995 unterschieden. Gezählt werden jeweils die Anzahl der eindeutigen Zuordnungen und die Anzahl der Mehrfachzuordnungen. Für die Mehrfachzuordnungen werden die mittlere Anzahl an Zuordnungen inklusive Standardabweichung und der maximalen Anzahl an Zuordnungen ermittelt. Anschließend werden die Ergebnisse in die gewünschte Form gebracht.

Statistische Ämter des Bundes und der Länder,
Arbeitspapier Nr. 51– Record-Linkage bei fehlenden Betriebsnummern im Produzierenden Gewerbe