Die Datensätze des

Economics & Business Data Center

2015

Economics & Business Data Center
Poschingerstr. 5
81679 München

Zusammenfassung

Das Economics & Business Data Center (EBDC) wurde im Jahr 2008 als Kooperation der LMU München und des Ifo Instituts für Wirtschaftsforschung gegründet und macht sich zum Ziel, neue Felder für die wirtschaftswissenschaftliche Forschung zu erschließen. In diesem Sinne stellt es innovative Datensätze deutscher Unternehmen bereit. Neben den **Mikrodaten der regelmäßigen ifo Befragungen**, bietet es verknüpfte Datensätze, die sowohl Umfragedaten des ifo Instituts als auch externe Bilanzdaten der Firmendatenbanken Amadeus und Hoppenstedt enthalten. Derzeit steht das **EBDC Business Expectations Panel** (Fokus: konjunkturelle Faktoren/Bilanzdaten), das **EBDC Business Investment Panel** (Fokus: Investitionen/Bilanzdaten) und das **EBDC Business Innovation Panel** (Fokus: Innovationen/Bilanzdaten) zur Verfügung, wobei beide Datensätze sowohl in zeitlicher als auch inhaltlicher Dimension kontinuierlich erweitert werden sollen.

Auf Grund der hohen Vertraulichkeit und der Datenschutzmaßnahmen, die den Unternehmen des ifo-Panels zugesichert werden, können die EBDC-Daten nur in den Räumlichkeiten des ifo Instituts und nur unter strengen Sicherheitsvorkehrungen genutzt werden. Die Daten werden außerdem anonymisiert und nur mit einjähriger Zeitverzögerung zur Verfügung gestellt. Nachfolgende Ausführungen geben einen Überblick über Datenbasis, Umfang und Zugang zu den EBDC Unternehmenspanels und bieten darüber hinaus Informationen zur Verknüpfungstechnik des Probabilistischen Record Linkage.

Einleitung

Das LMU Economics & Business Data Center (EBDC) wurde Anfang 2008 im Rahmen einer Kooperation der BWL- und VWL-Fakultäten der LMU sowie des ifo Instituts für Wirtschaftsforschung gegründet. Unterstützt wurde es von der Exzellenzinitiative des Bundes und der Länder zur Förderung von Wissenschaft und Forschung an deutschen Hochschulen im Rahmen von LMUexcellent.

Wichtigstes Ziel des Economics & Business Data Center ist die Erstellung innovativer Unternehmenspanels, die die ifo Umfragedaten mit verschiedenen externen Firmen- und Bilanzdatenbanken verknüpfen. Idee ist, die in den ifo Mikrodatensätzen enthaltenen Aspekte wie unternehmensspezifische Erwartungen, Einschätzungen und Pläne um Bilanz- und Strukturdaten zu erweitern, um so neue, einzigartige Ansätze für die empirische volks- und betriebswirtschaftliche Forschung zu ermöglichen. Zu den Aufgaben des EBDC gehören daher auch die Beschaffung und Verwaltung wichtiger Datenquellen für Forschung und Lehre, die zentrale Bereitstellung, Aktualisierung und Dokumentation von Datenbanken externer Anbieter, sowie der Erwerb von entsprechenden Support-Tools. Das EBDC hält in dieser Hinsicht eine geeignete hard- und softwaretechnische Infrastruktur vor und leistet Unterstützung im Hinblick auf die softwaretechnische Wissensvermittlung. Weiterhin bietet es interessierten Forschern und Nachwuchswissenschaftlern im Rahmen des Ifo Datapools[1] Zugang zu den, seit 1949 regelmäßig deutschlandweit erhobenen, Umfragedaten des ifo Instituts für Wirtschaftsforschung, sodass sich für empirisch arbeitende Wissenschaftler hohe Synergieeffekte ergeben.

Im Hinblick auf die neuen EBDC Business Panels wurden die Mikrodaten des ifo Konjunkturtest, des ifo Investitionstest sowie des ifo Innovationstest mit den Bilanzdaten der Unternehmensdatenbanken Amadeus und Hoppenstedt verknüpft. Erstmals liegen somit drei umfangreiche Datensätze deutscher Unternehmen vor, welche durch die Integration verschiedener Datenbestände die gleichzeitige Untersuchung von erwarteten bzw. geplanten und realisierten Größen ermöglichen. Die regelmäßig aktualisierten Panels umfassen sowohl historische als auch aktuelle Daten, wobei das EBDC Business

[1] Neben den ifo Mikrodaten der Unternehmensbefragungen verfügt das ifo Institut für Wirtschaftsforschung bzw. das EBDC auch über externe Mikro- und Makrodaten.

Expectations Panel 1980, das EBDC Business Investment 1964 und das EBDC Innovation Panel 1982 beginnt.

Professoren, Gastforschern und Nachwuchswissenschaftlern soll mit der vorliegenden Dokumentation ein Überblick über Datenbasis, Umfang und Verfügbarkeit der neuen EBDC-Datensätze sowie ein Einblick in die Verknüpfungstechnik des Probabilistischen Record Linkage gegeben werden. In Abschnitt 2 werden zunächst die zugrundeliegenden Datenquellen, d.h. der ifo Konjunkturtest bzw. der ifo Investitionstest sowie die Firmendatenbanken Amadeus und Hoppenstedt erläutert. Eine Übersicht über das Record Linkage gibt Kapitel 3. In Kapitel 4 werden Datenumfang und Bestandteile der EBDC Business Panels dargestellt und geplante Erweiterungen diskutiert. Der Zugang zu den Daten wird abschließend in Kapitel 5 beschrieben.

1. Datenbasis

Die EBDC-Unternehmenspanels entstehen aus der Verknüpfung des ifo Konjunkturtest (KT), des ifo Investitionstest (IT) sowie des ifo Innovationstests (INNO) mit den externen Bilanzdatenbanken Amadeus und Hoppenstedt. In diesem Abschnitt werden die jeweiligen Datenquellen kurz erläutert, anschließend wird auf die Verknüpfungstechnik des Probabilistischen Record Linkage eingegangen.

2.1. ifo Umfragedaten

Das Unternehmenspanel des ifo Instituts für Wirtschaftsforschung besteht aus vier regelmäßig durchgeführten Standard Unternehmensbefragungen: dem ifo Konjunkturtest (KT), dem ifo Investitionstest (IT), dem ifo Innovationstest (INNO) und dem World Economic Survey (WES). Der monatliche erhobene ifo KT richtet sich auf unternehmensspezifische Einschätzungen und Erwartungen zu Geschäftslage, Markt- und Wettbewerbssituation und ist daher auch Basis für den monatlich veröffentlichten ifo Geschäftsklimaindex. Demgegenüber fragen der ifo Investitionstest (halbjährlich erhoben) das Investitionsvolumen, der Innovationstest (jährlich) die unternehmerische Innovationstätigkeit bzw. der WES (vierteljährlich) die internationalen Konjunkturaussichten ab. In den Unternehmensbefragungen des ifo Instituts werden meist ordinale oder prozentuale und weniger absolute oder monetäre Größen verlangt, da sich gezeigt hat, dass die Erfragung

exakter Kaufpläne die Realität nicht genau genug abbilden kann. Dies ist darauf zurückzuführen, dass die Hälfte der Kaufentscheidungen privater Haushalte aber auch kleiner und mittelständischer Unternehmen spontan getätigt werden.[2] In diesem Sinne haben sich vor allem im KT monatliche „Einstellungsfragen“ wie die Frage nach der aktuellen Geschäftslage, etc. bewährt, welche implizit die tatsächliche bzw. erwartete Gewinnentwicklung der Unternehmen wiederspiegeln und somit einen Blick auf die voraussichtliche Konjunkturentwicklung gestatten. Der ifo-Ansatz wird dabei nicht nur auf nationaler Ebene registriert, sondern auch von der Europäischen Kommission und der OECD unterstützt. Aufgrund der angestrebten europaweiten Harmonisierung der Konjunkturumfragen gab es daher im Januar 2002 auch im ifo Panel eine Umstellung. Es findet seit diesem Zeitpunkt keine Unterscheidung mehr zwischen Berichts- und Erhebungsmonat statt. Ein Überblick über die Themen der gestellten Fragen in den einzelnen ifo Unternehmensbefragungen sowie über zahlreiche, darauf basierende, wissenschaftliche Studien findet sich bei Seiler (2012).

Beispiel: ifo Konjunkturtest für das Verarbeitende Gewerbe

Von der Struktur her gliedert sich der ifo Konjunkturtest in die Bereiche Verarbeitendes Gewerbe (KT VG), Handel (KT HAN), Bau (KT BAU) und Dienstleister (KT DL), wobei jeweils monatliche Standard- bzw. periodisch wiederkehrende Sonderfragen gestellt werden. Die an den Umfragen teilnehmenden Unternehmen erhalten hierzu i.A. einen Fragebogen, der sich jeweils auf ein Produkt/ eine Produktgruppe (KT VG, KT DL) bzw. auf eine Sparte/Geschäftsbereich (KT HAN, KT BAU) bezieht. Große Unternehmen im KT VG, deren Hauptumsatz sich aus der Produktion verschiedener Produkte zusammensetzt beantworten daher auch mehrere Fragebögen pro Monat. Beispielhaft lässt sich für den Konjunkturtest für das Verarbeitende Gewerbe folgende inhaltliche Ausrichtung feststellen:

[2] Vgl. Goldrian (2004).

Standardfragen:	*Sonderfragen:*
Auftragsbestand, Kapazitätsauslastung, Behinderung der Produktionstätigkeit, Wettbewerbsposition (Inland, Ausland), Beschäftigung, Lagerhaltung	*Ertragslage, Kreditvergabe Innovationen, besondere Anlässe*

Abb. 1: Inhalte des ifo Konjunkturtest für das Verarbeitende Gewerbe

Die befragten Unternehmen im KT VG finden meist binäre bzw. ordinal skalierte Antwortkategorien vor,[3] d.h. pro Standardfrage stehen jeweils nur zwei („1" ja, „2" nein) bzw. drei verschiedene Antwortmöglichkeiten („1" besser, „2" gleich, „3" schlechter) zur Auswahl. Es werden hier also Tendenzaussagen gemacht, zudem hängt die gegebene Antwort natürlich auch von der unternehmenseigenen Interpretation ab. So bezieht sich die Frage nach der „Geschäftslage" zwar auf die konjunkturelle Lage eines Unternehmens (jeweiliger Produktbereich) – woran ein Unternehmen diese abliest kann aber individuell verschieden sein.

Insgesamt umfasst der Konjunkturtest für das Verarbeitende Gewerbe ca. 300 Produktgruppen wobei diese so bemessen sind, dass sie in sich möglichst homogen sind. Um „authentische Umfragedaten"[4] zu erhalten wurde sowohl der fachlichen Repräsentativität (Produktvielfalt) als auch der Firmenrepräsentativität (Größe, Rechtsform, etc.) Rechnung getragen. Seit 1991 weißt der KT VG darüber hinaus keinerlei strukturelle Brüche auf, sodass sich pro Monat durchschnittlich 3.000 Meldungen bei einer Antwortquote von 80-85% ergeben. Intensive Firmenkontakte halten die Panelgröße und -zusammensetzung dabei auf einem repräsentativen Niveau.

ifo Investitionstest

Im Rahmen des Investitionstest werden Unternehmen des Verarbeitenden Gewerbes zweimal jährlich zu ihrer Investitionstätigkeit befragt. Die Umfrage findet im Frühjahr und im Herbst statt, wobei der IT im Gegensatz zum KT ein Unternehmen und dessen Investitionen

[3] Vereinzelt werden auch Prozentangaben gefordert.
[4] Vgl. Goldrian (2004).

als Ganzes anspricht. Jedes Unternehmen wird daher der Branche zugeordnet, in der der Produktionsschwerpunkt liegt. Im Allgemeinen wird nicht zwischen Standard- und Sonderfragen unterschieden, es gibt jedoch über das Jahr, über die Jahre und je nach West-/ Ost Erhebung verschiedene Themenschwerpunkte. Diese lassen sich grob wie folgt zusammenfassen:

Regelmäßige Fragen:[5]	*Fragen bis 2001:*
Investitionen/Investitionspläne laufendes Jahr/letztes Jahr/kommendes Jahr, Investitionsziele und –struktur, Einflüsse auf die Investitionstätigkeit	*Entwicklung der Produktionskapazität, Höhe der gemieteten Investitionsgüter, Finanzierung der Investitionen*

Abb. 2: Inhalte des ifo Investitionstest

Der Investitionstest gibt folglich Aufschluss über die getätigten und beabsichtigten Investitionen und macht Ziele sowie Einflüsse auf die Investitionstätigkeit deutlich. Hinsichtlich der Erhebungsjahre und der einzelnen Fragen ist zwischen West und Ost zu unterscheiden: im Westen wurde der Investitionstest ab 1987 erhoben, im Osten dagegen erst ab 1992, wobei sich die Fragen im Osten geringfügig unterscheiden. Generell werden im Investitionstest aber nicht nur Tendenzantworten erbeten, es werden vielmehr auch absolute Werte abgefragt (Vergangenheits- und Plandaten über die Höhe der Investitionen). Für Details sei auf den Investitionstest und die entsprechende Variablenliste verwiesen.

2.2. Firmendatenbank Amadeus

Die Amadeus-Firmendatenbank ist neben den ifo Umfragedaten und der Hoppenstedt-Datenbank zentrale Quelle für die EBDC Business Panel. Sie ist ein Produkt der Bureau van Dijk Electronic Publishing GmbH (BvDEP), einem der führenden europäischen Anbieter globaler Unternehmensinformationen, und enthält Geschäfts- und Finanzinformationen zu

[5] Bis Herbst 2002 wurden als Einflüsse auf die Investitionstätigkeit unter anderem die Faktoren „Technische Entwicklung“ sowie „Akzeptanz neuer Techniken“ geführt. Diese beiden Variablen wurden ab Herbst 2002 unter dem Faktor „Technische Faktoren“ zusammengefasst. Weitere Änderungen: siehe Variablenliste.

mehr als 11 Millionen, hauptsächlich nicht-börsennotierten Unternehmen aus 41 Ländern Europas, wobei derzeit ~ 1,5 Mio. deutsche Unternehmen erfasst sind.

Für die Firmendatenbestände werden Informationen marktführender lokaler Institutionen und renommierter Unternehmen der jeweiligen Länder herangezogen. Die Abschlussdaten für deutsche Unternehmen stammen von Creditreform bzw. der Creditreform Rating AG, die zur Creditreform Gruppe gehört. Diese stellt seit mehr als 125 Jahren Bonitätsauskünfte in Kunden-Lieferanten-Beziehungen bereit und ist heute europaweit Marktführer für Bonitätsinformationen. Zentrale Quellen der Amadeus-Datenbank sind die MARKUS-Datenbank, welche i.A. Gesellschaftsinformationen deutscher Handelsregisterunternehmen mit einem Bonitätsindex von maximal 499 enthält (Verband Creditreform) und die DAFNE-Datenbank, die Jahresabschlüsse, Beteiligungsdaten, etc. aller publizierenden deutschen Unternehmen beinhaltet (Creditreform Rating AG). Im Unterschied zur DAFNE-Datenbank (Rohdatenformat) liegen die Daten in Amadeus jedoch in einem einheitlichen, standardisierten Bilanzformat vor, welches von nationalen bzw. internationalen Rechnungslegungsvorschriften abstrahiert. So besteht jeder Unternehmensbericht aus insgesamt 23 Bilanzpositionen, 25 Positionen der Gewinn- und Verlustrechnung, 20 Finanzkennzahlen und zahlreichen deskriptiven Informationen wie bspw. Branchencodes, Gesellschafterstrukturen, Aktien-, oder Kursinformationen. Für die Unternehmen der EBDC-Unternehmenspanels wurden hieraus über 50 Positionen ausgewählt – Beteiligungs-, Aktien- und Kursinformationen wurden jedoch zunächst vernachlässigt.[6] Bzgl. der Aktualität garantiert Amadeus generell, dass Abschlussinformationen für nicht börsennotierte Unternehmen spätestens nach einem Zeitraum von 15 Monaten in der Datenbank verfügbar sind.

2.3. Bilanzdatenbank Hoppenstedt

Die Hoppenstedt-Bilanzdatenbank ist ein Produkt der Hoppenstedt Firmeninformationen GmbH, welche als Teil der Hoppenstedt-Gruppe einer der führenden Anbieter von Wirtschafts- und Brancheninformationen in Deutschland ist. Wesentliche Geschäftsfelder sind neben der Bereitstellung von Firmeninformationen der Adressverkauf, Kredit- und Risikoanalysen sowie die Publikation von Fachzeitschriften. Je nach Kundenstatus und Nutzungsrechten kann daher aus einer Vielzahl an Datenbanken gewählt werden. Für die

[6] Diese können jedoch über die Historischen Datenbanken im EBDC exportiert werden.

Firmendatenbestände werden Informationen aus externen Quellen, wie z.B. Bundesanzeiger, Handelsregister, Wirtschaftspresse oder Geschäftsberichte, herangezogen bzw. bei Bedarf auch im direkten Dialog ermittelt. Laut Hoppenstedt werden alle bekannten Änderungen tagesaktuell ausgewertet, aufbereitet und in die jeweilige Datenbank übernommen, weshalb sich die angebotenen Unternehmensinformationen vor allem durch Aktualität, Qualität und Datentiefe auszeichnen.

Die Informationen für die EBDC Business Panel wurden der Hoppenstedt-Bilanzdatenbank entnommen,[7] welche aktuell mehr als 3,5 Mio. Abschlüsse von über 1 Mio. deutschen Unternehmen aus den Bereichen Industrie, Handel, Dienstleistungen, Versicherungen und Banken beinhaltet. Erfasst sind hier fast alle seit 2005 publizierten Abschlüsse, die Informationen für große Unternehmen gehen außerdem teilweise bis ins Jahr 1987 zurück. Die erhobenen Daten zu Bilanzen, Gewinn- und Verlustrechnungen einzelner Firmen sind jeweils in unterschiedlichen Detailtiefen abrufbar (Normbilanz: maximal verfügbare Positionen nach der jeweiligen Rechnungslegungsvorschrift; Verkürzte Bilanz: ca. 90; Kurzbilanz: ca. 30 Positionen), zudem ermöglichen separate, eng am jeweiligen Original orientierte Bilanzschemata die Berücksichtigung der einzelnen Abschlussarten nach HGB, IAS und US-GAAP.

[7] Zum Zeitpunkt der Datenerhebung für das EBDC waren bei Hoppenstedt ca. 120.000 Unternehmen verfügbar.

3. Verknüpfungsmethode – Probabilistisches Record Linkage

Zur Verknüpfung der ifo Umfragen auf der einen und der Bilanzdatenbanken Amadeus und Hoppenstedt auf der anderen Seite wird auf die Adressinformationen der Unternehmen in den einzelnen Datenbeständen zurückgegriffen. Auf diese Weise lassen sich jeweils zwei Zuordnungstabellen (ifo-Amadeus und ifo-Hoppenstedt) bilden, welche anschließend im entsprechenden EBDC Business Panel zusammengeführt werden können.

Im Folgenden soll beispielhaft das Record Linkage der Adressdaten des ifo Konjunkturtest (ifo KT) mit jenen aus der Firmendatenbank Amadeus beschrieben werden, wobei hier zusätzlich zum normalen Verknüpfungsprozess - der beim Record Linkage mit Hoppenstedt bzw. im Falle des ifo Investitionstest analog verläuft- ein sogenannter „Goldstandard“ erzeugt wurde. Dieser Goldstandard wird benötigt, um die Match- bzw. Nonmatch-Gewichte für jede Adressvariable zu ermitteln. Die Verknüpfung der Adressdatensätze des Ifo Datapools (Basis: gesamter KT inkl. Bau, Handel, Dienstleister, Industrie, mit den deutschen Firmen in der Amadeus-Datenbank erfolgte dabei mit Hilfe der Matching-Software MTB (Merge Toolbox), welche am „Center for quantitative Methods and Survey Research“ der Universität Konstanz entwickelt wurde.

MTB macht eine Zuordnung unterschiedlicher Datensätze bei großem Datenumfang und unterschiedlichen Adressaufbereitungen möglich und wird eingesetzt, wenn kein eindeutiger Schlüssel wie z.B. eine Firmenkennziffer vorliegt. Auf Basis des Probabilistischen Record Linkage, das auf der Theorie von Newcombe et. al (1959), aufbaut und von Fellegi und Sunter (1969) formalisiert wurde, werden gleiche Namens-/Adressdaten zusammengefügt. Der Grad der Übereinstimmung - die „Ähnlichkeit“ - der Variablen wird hierbei mit Hilfe von Wahrscheinlichkeiten ermittelt.
Diese „Ähnlichkeit“ errechnet sich aus dem Quotienten der Wahrscheinlichkeit, dass für gleiche Firmen die Variable x aus beiden Datenmengen als übereinstimmend erkannt wird (M-Wahrscheinlichkeit) und der Wahrscheinlichkeit, dass für ungleiche Firmen die Variable x aus beiden Datenmengen als übereinstimmend gewertet wird (u-Wahrscheinlichkeit). Im Idealfall beträgt der Quotient 1/0, wobei Abweichungen zwischen den verschiedenen Variablen unterschiedlich bewertet werden müssen.
Aus diesem Grund wird für jede Variable der Logarithmus der M-Wahrscheinlichkeit (= Match-Gewicht) bzw. u-Wahrscheinlichkeit (= Nonmatch-Gewicht) berechnet, um im Falle eines Matches bzw. Nonmatches die Qualitätsvariable („Quality“) der Verknüpfung zu

erhöhen bzw. zu reduzieren. Die Quality-Variable bildet dabei den Grad der Ähnlichkeit einer Datensatzverknüpfung ab und ergibt sich aus der Summe der Gewichte der einzelnen Adressvariablen.

Da die Parameter (Gewichte) des Probabilistischen Record Linkage empirisch gewonnen werden mussten, wurde aus der Menge der Datensätze zunächst eine gesichert zuzuordnende Teilmenge gebildet – ein sogenannter „Goldstandard“. Als Goldstandard wird generell eine Verknüpfung bezeichnet, die eine eindeutige Zuordnung der Datensätze zweier Datenbanken ermöglicht. Das erste Record Linkage erfolgte deshalb über Telefonnummer, Faxnummern und Email-Adresse. Diese Variablen waren jedoch nicht in allen Datensätzen befüllt und da es auch erhebliche Unterschiede bei der Systematik der Telefon-/Faxnummerneingabe gab, war eine Verknüpfung der gesamten Datenmenge auf diese Weise nicht möglich. Es ließen sich so 40% der ifo-Einträge direkt einem Unternehmen aus Amadeus zuordnen. Es zeigte sich allerdings, dass hierbei auch Datensätze als Matches identifiziert wurden, die eigentlich nicht zusammengehören. Dies ist häufig dann der Fall, wenn innerhalb einer Firma mehrere Untereinheiten existieren – z.B. Verwaltungsgesellschaft, Holding und Geschäftsführung – die sich alle am selben Standort befinden und so möglicherweise über dieselbe zentrale Telefon-, Faxnummer oder Email-Adresse verfügen. Einige der verknüpften Goldstandardpaare waren somit also keine True Matches im eigentlichen Sinne, sie waren nur True Matches im Sinne der Zugehörigkeit zu einem größeren Unternehmensverbund.

Innerhalb dieser verknüpften Datenmenge wurden nun die Gewichte für die einzelnen Adressvariablen berechnet. Dafür wurde die Zahl der richtigen bzw. falschen Übereinstimmungen zur Grundgesamtheit der Vergleiche ins Verhältnis gesetzt. Um das Ergebnis nicht zu verzerren, wurden die Gewichte zusätzlich für verschiedene Aufbereitungsvarianten der Variablen ermittelt. Ziel hierbei war, das Ergebnis aufgrund häufig vorkommender Namenssequenzen nicht zu positiv bzw. aufgrund unterschiedlicher Schreibweise des gleichen Namens nicht zu negativ zu bewerten:
die am besten differenzierende Variante der Matchvariablen ist jene mit der größten Anzahl an wahren (M-Wahrscheinlichkeit) und einer möglichst geringen Anzahl an falschen Übereinstimmungen (u-Wahrscheinlichkeit). Da der Goldstandard auch Matches enthielt die im eigentlichen Sinne nicht als solche bezeichnet werden konnten, wurde die Richtigkeit der ermittelten Gewichte nochmals anhand einer, 2000 Datensätze umfassenden, Zufallsstichprobe aus dem Goldstandard per Hand kontrolliert. Aus den daraus

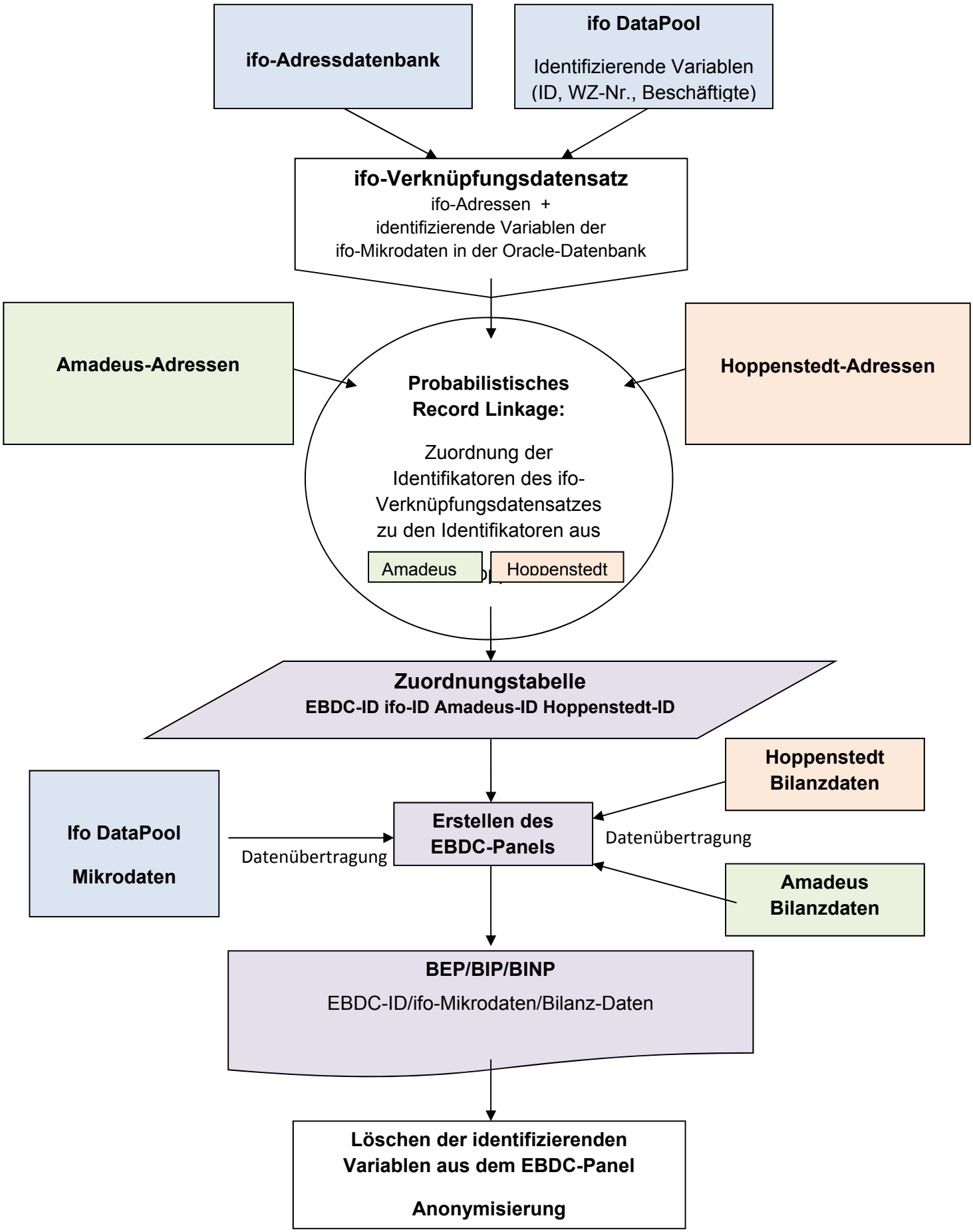
resultierenden Ähnlichkeitsgewichten, die vor dem eigentlichen MTB-Lauf den Variablen zugeordnet wurden, errechnete das Programm die Quality. Wurden für die Verknüpfung eines Unternehmens aus beiden Datenbeständen für alle Variablen positive Matches ermittelt, so ergab sich die Quality als maximales Gesamtgewicht aus der Summe der einzelnen Übereinstimmungsgewichte: z.B. 10,6 + 8,6 + 4,9 + 12,9 = 37,2. Wurde dagegen beispielsweise eine der Variablen von MTB als Nonmatch identifiziert, und damit das (negative) Nonmatch-Gewicht in die Gleichung eingesetzt, verringerte sich die Quality für die jeweilige Verknüpfung entsprechend.

Der beschriebene Abgleich der Variablen aus dem ifo- und Amadeus-Panel wurde mittels einer String-Ähnlichkeitsfunktion durchgeführt, welche N-Gramme der Länge 2 (= Bigramm mit Leerzeichen vor und hinter allen Strings) aus den jeweiligen Variablen miteinander vergleicht, indem ein Raster der Länge 2 über den String gelegt wird. Unterschiede in den Variablenausprägungen wurden linear nach Bigramm-Ähnlichkeit gewichtet, wobei MTB eine hohe Übereinstimmung bei einem kürzeren Namen niedriger bewertet als eine hohe Übereinstimmung bei einem längeren Namen. Die Festlegung eines für alle Variablen gültigen Jaro-Faktors (= Gewichtsanpassung) auf den Wert „2“ führte weiterhin zu einer schnelleren Vergabe des vollen Übereinstimmungsgewichts bei hoher Übereinstimmung und damit zu einer besseren Differenzierung der Matches von den Nonmatches. Für die Auswertung des MTB-Laufs wurde schließlich ein Schwellenwert hinsichtlich der Quality-Variable definiert, wobei allerdings nicht auf die Bewertung des Programms zurückgegriffen wurde welches eine Verknüpfung ab einer gewissen Quality als Match identifiziert. In unserem Fall war es zweckmäßiger, von vorn herein höhere Werte anzusetzen und für einen großen Quality-Bereich eine erneute Handkontrolle durchzuführen. Ziel hierbei war es wieder, den Fehler falsch positiver Treffer zu vermeiden – Pärchen können nicht als True Match qualifiziert werden, nur weil sie über einem Schwellenwert liegen (Ort im Namen kann Ähnlichkeit erhöhen, ebenso falsche Rechtsform bei gleichem Namen). Zum anderen können sich unterhalb des definierten Schwellenwerts in bestimmten Blöcken noch einige True Matches befinden die wegen fehlender Informationen zu stark abgewertet wurden (gleicher Name, Straße leer).

3.1. Ergebnisse

Auf diese Weise wurden die Unternehmensadressen des ifo Konjunkturtest mit den Unternehmensadressen aus der **Amadeus/Hoppenstedt Firmendatenbank** verglichen. Da die Adresszuordnung des BEP auf Fragebogenebene (questionnaire_id) stattfindet, können einem Unternehmen aus Amadeus durchaus auch zwei oder mehr ifo-Einträge zugeordnet sein. Abbildung 3 verdeutlicht nochmals den Entstehungsprozess des EBDC Business Expectations Panel. Da im ifo Investitionstest nicht zwischen Produkten unterschieden wird, ist jeder ifo-Adresseintrag genau einem Eintrag bei Amadeus bzw. Hoppenstedt zugeordnet und wiederum über die BIP-ID identifizierbar. Allgemein gilt: waren für ein Unternehmen sowohl Informationen aus Amadeus als auch aus Hoppenstedt verfügbar, wurde Letzteren aufgrund der umfangreicheren Variablenauswahl und -befüllung der Vorzug gegeben.[8]

[8] Die Bilanzinformationen aus der Amadeus-Datenbank liegen ursprünglich in tsd. Euro und gerundet vor, zudem sind einzelne Positionen oftmals aggregiert. Hoppenstedt weist dagegen Rohdaten und eine eng am HGB orientierte Bilanz-Struktur aus.


ifo-Adressdatenbank
ifo DataPool
Identifizierende Variablen
(ID, WZ-Nr., Beschäftigte)
ifo-Verknüpfungsdatensatz
ifo-Adressen +
identifizierende Variablen der
ifo-Mikrodaten in der Oracle-Datenbank
Amadeus-Adressen
Hoppenstedt-Adressen
Probabilistisches
Record Linkage:
Zuordnung der
Identifikatoren des ifo-
Verknüpfungsdatensatzes
zu den Identifikatoren aus
Amadeus
Hoppenstedt
Zuordnungstabelle
EBDC-ID ifo-ID Amadeus-ID Hoppenstedt-ID
Hoppenstedt
Bilanzdaten
Ifo DataPool
Mikrodaten
Datenübertragung
Erstellen des
EBDC-Panels
Datenübertragung
Amadeus
Bilanzdaten
BEP/BIP/BINP
EBDC-ID/ifo-Mikrodaten/Bilanz-Daten
Löschen der identifizierenden
Variablen aus dem EBDC-Panel
Anonymisierung

4. Aufbau der EBDC-Unternehmenspanels

Generell sind alle EBDC Business Panel ähnlich aufgebaut. Die Identifikation erfolgt über eine unternehmenseigene EBDC-ID, das jeweilige Jahr der Beobachtung sowie weitere Datensatz-spezifische Zeitvariablen (s. unten).

Den Abschlussinformationen in den EBDC-Unternehmenspanels liegen, wenn verfügbar, Einzel- statt Konzernabschlüsse zugrunde,[9] wobei die Bilanzdaten der beiden Unternehmensdatenbanken nicht einfach übernommen wurden. Es wurde vielmehr ein neues EBDC-Bilanzschema entwickelt, welches sowohl Amadeus- als auch Hoppenstedt-Variablen integriert und von den bestehenden Unterschieden der ursprünglichen Datenbestände abstrahiert.[10] Das EBDC-Bilanzschema orientiert sich dabei an Bilanz- und GuV-Struktur des Handelsgesetzbuchs (HGB) und weist teilweise auch Variablen nach Gesamt- oder Umsatzkostenverfahren aus.[11] Für eine detaillierte und entsprechend gegliederte Übersicht zu den verfügbaren Bilanz- und GuV-Variablen sei auf die jeweilige Variablenliste des entsprechenden EBDC Business Panels verwiesen.

[9] Die jeweilige Abschlussart wird durch die Variable „reporting_basis“ angezeigt. Limited financial data meint hierbei, dass die Bilanzinformationen nicht veröffentlicht sondern meist individuell erfragt wurden.
[10] Das Umrechnungsschema, welches zur Überführung der ursprünglichen Variablen in die neu generierten EBDC-Bilanzvariablen verwendet wurde, kann am EBDC eingesehen werden. Auch ist eine detaillierte Standard-Bilanz verfügbar, welche zu Orientierungszwecken herangezogen werden kann.
[11] Auf Wunsch und in Ausnahmefällen können die EBDC-Panels auch mit den ursprünglichen Bilanzvariablen aus Amadeus und Hoppenstedt zur Verfügung gestellt werden.

4.1. EBDC Business Expectations Panel

Für das EBDC Business Expectations Panel ergibt sich durch die Verknüpfung eine Kombination aus Monats- (ifo) und Jahresdaten (Bilanz). Die BEP-ID setzt sich im EBDC Business Expectations Panel aus drei Bestandteilen zusammen: aus der Variable „company_id“, einer durchlaufenden Unternehmensnummer, der Variable „questionnaire_id“, die die Fragebögen pro Unternehmen durchnummeriert und der Variable „sector_id“, die die Sektoren enthält, die pro Fragebogen abgefragt werden. Die „questionnaire_id“ steht in den verschieden Erhebungen für unterschiedliche Befragungsebenen. So wird im KT VG ein Fragebogen pro Produkt verschickt, im KT DL pro Dienstleistungssparte, im KT Han pro Produktgruppe und im KT Bau pro Bausparte. Auf einem Frageborgen können jeweils mehrere Sektoren abgefragt werden.
Der Datensatz ist nach BEP-ID, Jahr (year), Monat (month) sortiert und liegt im long-Format vor, d.h. jede Meldung ist über diese drei Variablen identifizierbar. In den einzelnen Monaten sind jeweils die Umfrageergebnisse aus dem ifo KT enthalten, sodass für jedes Jahr bis zu 12 Monatsmeldungen pro BEP-ID vorliegen können. Daran anschließend folgt die Bilanzinformation in einem dafür konstruierten Monat „99“. Vorteil dieser Datensatz-Struktur ist vor allem die übersichtliche Handhabbarkeit, die eine individuelle Zuordnung von Monats- und Jahresinformationen möglich macht.

Die einzelnen Spalten des EBDC Business Expectations Panel enthalten nacheinander die nach ihrer Funktion geordneten Variablen: Identifikationsvariablen, Bilanz- und GUV- sowie ifo KT-Variablen. Zu den Identifikationsvariablen zählen neben der BEP-ID, dem Jahr und dem Monat auch Angaben zu Branchencodes, Beschäftigtengrößenklassen, Bundesland, Börsennotierung, Rechtsform, etc..[12] Die Variablenliste enthält neben einer Übersicht zu allen Variablen, deren Bedeutung und Besonderheiten auch die Fragen der einzelnen ifo KT-Erhebungen inklusive Erhebungszeitraum und –rhythmus. Das Kriterium ihrer Übernahme aus dem ifo KT in das BEP ist, dass sie in mindestens 2 der 4 Erhebungen entsprechend vorkommen.[13]

Die Auswertung des Datensatzes (deskriptive Statistiken, Befüllung ausgewählter Variablen, etc.) kann am EBDC eingesehen und kopiert werden. Einen ersten Eindruck im Hinblick auf die Repräsentativität der Unternehmen im EBDC Business Expectations Panel

[12] Aus Gründen der Anonymisierung wurde die Bundesland-Angabe bei großen Unternehmen (> 10.000 Beschäftigte) gelöscht.
[13] Daneben kann auch ein BEP-Datensatz zur Verfügung gestellt werden in dem alle Fragen enthalten sind.

liefert jedoch bereits Abbildung 4.[14] Hier sind die Antworten der im EBDC Business Panel enthaltenen Firmen des Bereichs Verarbeitendes Gewerbe auf die Frage nach vorhandenen Kreditbeschränkungen im Vergleich zum Ifo Credit Constraint Indicator (Kredithürde) dargestellt. Dieser bildet den Anteil der Unternehmen im Verarbeitenden Gewerbe ab, die die Kreditvergabepolitik der Banken als restriktiv beurteilen. Deutlich wird hier vor allem in den letzten Jahren die Übereinstimmung der beiden Anteile, weshalb das (weniger umfangreiche) EBDC-Panel im Hinblick auf den ifo Konjunkturtest für das Verarbeitende Gewerbe durchaus als repräsentativ bezeichnet werden kann.

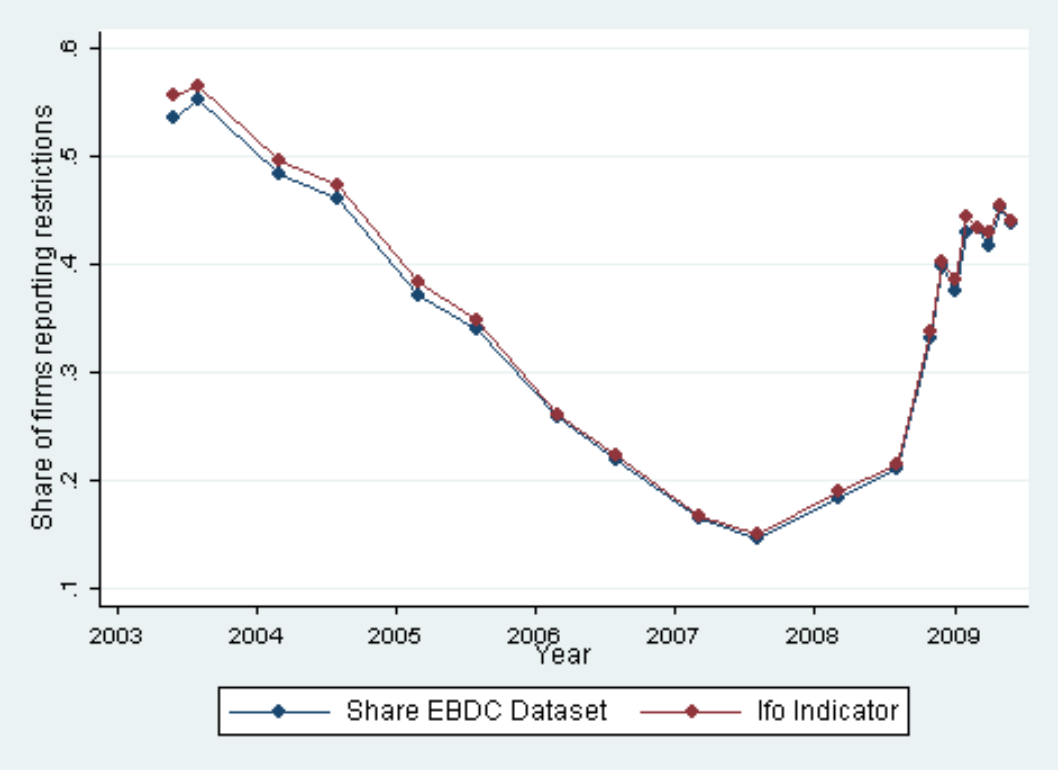


Abb. 4: Vergleich EBDC Expectations Panel und ifo Kredithürde

[14] Quelle: EBDC, ifo Institut

4.2. EBDC Business Investment Panel

Durch die halbjährliche Abfrage des ifo Investitionstest (IT) ergibt sich für das EBDC Business Investment Panel eine leicht veränderte Struktur. Der Datensatz enthält keine Monatsangabe mehr, dafür aber eine Variable „survey_base“ die angibt, ob die Daten aus der Frühjahrs- (1) oder Herbst- (2) Erhebung des IT bzw. aus einer Bilanz (99) stammen.

Der Datensatz ist nach BIP-ID (company_id), Jahr (year), Erhebungsquelle (survey) sortiert. Er liegt ebenfalls im long-Format vor, d.h. jede Meldung ist über diese drei Variablen auch identifizierbar. Zusätzliche Identifikationsvariablen eines Unternehmens sind verschiedene Branchencodes, Beschäftigtengrößenklassen, Bundesland, Börsennotierung, Rechtsform, etc..[15] Die einzelnen Spalten des Panels enthalten wiederum die nach ihrer Funktion geordneten Variablen: Identifikationsvariablen, Bilanz- und GUV- sowie ifo IT-Variablen. Die Variablenliste für das EBDC Business Investment Panel enthält neben einer Übersicht zu allen Variablen, deren Bedeutung und Besonderheiten die Fragen der einzelnen ifo IT-Erhebungen inklusive Erhebungszeitraum und -rhythmus.

Die Auswertung des Datensatzes (deskriptive Statistiken, Befüllung ausgewählter Variablen, etc.) kann am EBDC eingesehen und kopiert werden.

4.3. EBDC Business Innovation Panel

Der bereits erwähnte ifo Innovationstest (INNO) bezieht sich, ebenso wie der ifo KT VG, auf einzelne Produkte eines Unternehmens. Zudem wird für die jährlichen Umfragen des Innovationstest ein Teil der Unternehmen aus dem ifo Konjunkturtest für das Verarbeitende Gewerbe herangezogen, sodass diese beiden Datenquellen zusammen das EBDC Business Innovation Panel bilden. Abgefragt werden im INNO sowohl Innovationsaktivitäten und -ziele, als auch Innovationsimpulse und –hemmnisse, weshalb er für verschiedenste Forschungsarbeiten von Interesse sein dürfte. Generell werden Produkt- und Prozessinnovationen angesprochen, zudem gibt es wechselnde Sonderfragen in einzelnen Jahren. Das Zusatzmodul enthält dieselbe ID/Jahr/Monat–Kombination (Monatsangabe: 98) wie das EBDC Business Expectations Panel. Die Sonderfragen Innovation im Ifo Konjunkturtest wird jedes Jahr im Dezember im Rahmen des KT gestellt und befindet sich unter der Monatsangabe = 97. Inhaltlich fragt sie nach den Perspektiven für ein

[15] Aus Gründen der Anonymisierung wurde die Bundesland-Angabe bei großen Unternehmen (> 10.000 Beschäftigte) gelöscht.

spezifisches Produkt (Markt wachsend, stagnierend oder schrumpfend) sowie nach dem Innovationsstatus bzw. der Entwicklungsphase.

4.4. Erweiterungen

Alle EBDC Business Panel werden regelmäßig aktualisiert, wobei aus Anonymitätsgründen ein einjähriger time-lag zum aktuellen Rand eingehalten wird. Die Aktualisierung bezieht sich generell sowohl auf die Zeitdimension als auch auf die Basis der enthaltenen Unternehmen. So werden auch zusätzliche, in den jeweiligen Adressdatenbanken neu aufgenommene Firmen jeweils per Record Linkage in das bestehende Panel integriert. Eine weitere Möglichkeit stellt die paarweise Verknüpfung der EBDC Business Panel sowie die Verknüpfung der Datensätze mit anderen, externen Daten wie Aufsichtsrat- oder Eigentümerstrukturen dar.

5. Zugang

Das EBDC sieht sich als Dienstleister, welcher Professoren, Gastforscher und Doktoranden bei ihren Forschungsprojekten unterstützt und hierzu unter anderem die EBDC Business Panel bereitstellt. Bei den Forschungsprojekten muss es sich um nicht kommerzielle, wissenschaftliche Themen handeln, die auf Basis der EBDC-Daten empirisch bearbeitet werden.

Auf Grund der hohen Vertraulichkeitsanforderungen bzw. der Verpflichtung zur Geheimhaltung von Befragungsergebnissen und Unternehmensidentität bzgl. der Daten des ifo Unternehmenspanels können die EBDC-Unternehmenspanels nur in den Räumlichkeiten des ifo Instituts genutzt werden. Es wird ein Rechner ohne Zugang zu Internet, Drucker oder anderen externen Datenträgern zur Verfügung gestellt, welcher nur in Anwesenheit eines EBDC-Mitarbeiters genutzt werden kann. Dieser wird nach Beendigung des Aufenthalts auch sicherstellen, dass die zu Forschungszwecken aufbereiteten, anonymisierten Daten keine Rückschlüsse auf einzelne Unternehmen oder die Panelzusammensetzung insgesamt ermöglichen und die Ergebnisdateien im Stata-Format nach erfolgreicher Prüfung versenden.

Zugang zu den EBDC Business Panel kann über ein Formular im Internet beantragt werden,[16] wobei zusätzlich ein einseitiges Konzept einzureichen ist in dem das Forschungsvorhaben sowie wichtige Randinformationen (Termine, etc.) kurz erläutert werden. Auf Anfrage sendet das EBDC auch ein Testpaket per Email, welches einen anonymisierten EBDC-Testdatensatz im Stata-Format sowie die Variablenliste zum jeweiligen Original-Panel enthält. Die Forschungsvorhaben werden vom EBDC ausdrücklich unterstützt und sind daher kostenfrei – der Zugang zu den EBDC-Daten ist allerdings von der Verfügbarkeit freier Arbeitsplätze abhängig.

[16] http://www.cesifo-group.de/portal/page/portal/ifoContent/N/data/EBDC_Container/EBDC_Angebot_Container/EBDC_Vertrag.pdf

7. Literaturverzeichnis

Fellegi, I.P., Sunter, A.B. (1969): A Theory for Record Linkage. Journal of the American Statistical Association, 64, 1183-1210.

Goldrian, G. (2004): Handbuch der umfragebasierten Konjunkturforschung. ifo Beiträge zur Wirtschaftsforschung, 15, ifo Institut für Wirtschaftsforschung, München.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., James, A.P. (1959): Automatic Linkage of Vital Records. Science, 130, 954-959.

Oppenländer, K.H., Poser, G. (1989): Handbuch der Ifo-Umfragen: 40 Jahre Unternehmensbefragungen des Ifo-Instituts für Wirtschaftsforschung. Duncker & Humblot, München

Seiler (2012): The Data Sets of the LMU-ifo Business & Economics Data Center – A Guide for Researchers, Journal of Applied Social Science Studies 132(4), 609-618.