FDZ Methodenreport

Methodische Aspekte zu Arbeitsmarktdaten

Nr. 5/2007

KombiFiD – Kombinierte Firmendaten für Deutschland

Konzeption der Machbarkeitsstudie für eine Zusammenführung von Unternehmensdaten der Statistischen Ämter, des Instituts für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit und weiterer Datenproduzenten

Stefan Bender(Forschungsdatenzentrum der BA im IAB)
Joachim Wagner (Institut für Volkswirtschaftslehre der Leuphana Universität Lüneburg) und
Markus Zwick (Statistisches Bundesamt)

Bundesagentur für Arbeit

Vorbemerkung

Das Projekt KombiFiD wird vom Bundesministerium für Bildung und Forschung (BMBF) finanziell gefördert. Für hilfreiche Anmerkungen zum Projektantrag danken wir Reinhard Hujer und Hans-Jürgen Krupp.

Zusammenfassung

Bei den Statistischen Ämtern des Bundes und der Länder und im Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit werden zahlreiche Daten für Unternehmen bzw. Betriebe gesammelt und aufbereitet. Durch das Unternehmensregister ist bei den Statistischen Ämtern eine „Masterdatei“ entstanden, die es technisch möglich macht, Unternehmensdaten aus den Beständen der genannten und weiteren Institutionen zusammenzuführen. Das Projekt KombiFiD – Kombinierte Firmendaten für Deutschland soll zeigen, dass ausgewählte Datenbestände tatsächlich über die Grenzen der jeweiligen Datenproduzenten zusammengeführt und der Wissenschaft zur Verfügung gestellt werden können, und es soll gleichzeitig demonstrieren, dass das Analysepotenzial dieser kombinierten Datensätze sehr viel höher ist als das der einzelnen Bestandteile. Der vorliegende Beitrag stellt die Konzeption dieses Projekts und das geplante Vorgehen vor.

Inhaltsverzeichnis

1. Grundidee des Projekts

Bei den Statistischen Ämtern des Bundes und der Länder und im Institut für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit (BA) werden eine Reihe von Unternehmens- bzw. Betriebsdaten gesammelt und aufbereitet, von denen viele über die Forschungsdatenzentren (FDZ) für Wissenschaftler zugänglich sind. Durch das Unternehmensregister (URS) ist zudem eine „Masterdatei" entstanden, die u.a. Identifikatoren für das Zusammenführen von Informationen aus unterschiedlichen Datenquellen enthält. Damit ist es grundsätzlich technisch möglich, Unternehmensdaten aus den Beständen der genannten Institutionen zusammenzuführen. Das Projekt *KombiFiD – Kombinierte Firmendaten für Deutschland* setzt genau hier an. Es soll zeigen, dass ausgewählte Datenbestände über die Grenzen der jeweiligen Datenproduzenten zusammengeführt und der Wissenschaft zur Verfügung gestellt werden können, und es soll gleichzeitig demonstrieren, dass das Analysepotenzial dieser kombinierten Datensätze sehr viel höher ist als das der einzelnen Bestandteile.

Die betrachtete Einheit ist in diesem Projekt das Unternehmen im Sinne einer rechtlich selbständigen Einheit, wobei ein Unternehmen mehrere Betriebsstätten an unterschiedlichen Orten besitzen kann. Eine Verknüpfung der Datenbestände aus den unterschiedlichen Erhebungen auf der Betriebsebene ist allerdings nicht möglich, da viele Daten nur für Unternehmen erhoben werden, das URS nur eine Verknüpfung auf der Unternehmensebene technisch ermöglicht, und sich Betriebsdaten in vielen Fällen[1] zu Unternehmensdaten aggregieren lassen, umgekehrt eine Aufspaltung von Unternehmensdaten auf einzelne Betriebe aber in der Regel nicht möglich ist.

Das in KombiFiD erstellte Datenmaterial soll von Beginn an für die Analyse von dynamischen Prozessen geeignet sein. Hierfür werden Paneldaten – Informationen über die Unternehmen zu unterschiedlichen Zeitpunkten – benötigt. Der Zeitraum, für den diese Daten aufbereitet werden sollen, wird einerseits durch die Jahre eingeschränkt, in denen die jeweils herangezogenen Erhebungen durchgeführt wurden bzw. für die das Ausgangsdatenmaterial in geeigneter Form verfügbar ist. Andererseits ist – worauf

[1] Eine Aggregation von Betriebs- zu Unternehmensdaten ist dann nicht möglich, wenn es sich um eine Erhebung handelt, in der nicht alle Betriebe eines Unternehmens erfasst sind, wie dies z.B. in Stichproben von Betrieben der Fall sein kann. Die Entscheidung für das Unternehmen als Einheit schränkt die Verwendbarkeit der kombinierten Daten für Analysen mit regionalem Bezug stark ein, da die Betriebe von Mehrbetriebsunternehmen ihren Sitz in unterschiedlichen Regionen haben können.

noch näher einzugehen sein wird – eine Zusammenführung von Unternehmensdaten über die Grenzen der Datenproduzenten hinweg nur mit Einverständnis der betroffenen Unternehmen möglich, und es ist davon auszugehen, dass die Zustimmung der Unternehmen um so schwieriger zu erhalten sein wird, je aktueller (und damit schützenswerter) diese Daten sind. Als Zeitraum für das zu erstellende KombiFiD-Unternehmenspanel haben wir daher die Jahre von 1995 bis 2003 gewählt.[2]

Das im Projekt KombiFiD neu verfügbar gemachte Datenmaterial wird in erheblichem Maße die empirische Wirtschaftsforschung beeinflussen. Sowohl für die ökonometrische Überprüfung modelltheoretisch hergeleiteter Hypothesen als auch für die wissenschaftliche Politikberatung stehen mit diesen Daten neue Wege offen. Derzeit stehen Forscherinnen und Forschern in den Forschungsdatenzentren der Statistischen Ämter und des IAB einzelne Betriebsdatensätze zu Analysezwecken zur Verfügung. Als Beispiel sei hier auf das IAB-Betriebspanel verwiesen, welches momentan als die am häufigsten genutzte und umfassenste Datenquelle für Betriebsanalysen gilt. Trotzdem hat sie den Nachteil, dass zentrale Merkmale für die Betriebe nicht zur Verfügung stehen (da sie nicht erhoben werden können oder hohe Missing Values-Anteile im Betriebspanel aufweisen), die Fallzahl der untersuchten Betriebe oft sehr gering ist oder die Unternehmensebene für die Analyse fehlt. Mit dem Datenmaterial aus dem KombiFiD-Projekt werden dann zahlreiche Fragestellungen z. B. zur Arbeitskräftenachfrage und Unternehmensbesteuerung erstmals simultan untersucht werden können. Zusammenhänge zwischen Kostenstrukturen, Unternehmenseffizienz und Marktaustritten oder zwischen internationaler Firmentätigkeit und Änderungen der Qualifikationsstruktur der Belegschaft sind weitere beispielhaft zu nennende Themen für empirische Analysen.

Ein weiteres Ziel von KombiFiD ist es, den Unternehmen eine bessere Informationsgrundlage für eigene Entscheidungen zur Verfügung zu stellen. Aus den Veröffentlichungen der wissenschaftlichen Arbeiten können die Unternehmen z.B. ihre Position verglichen mit dem Branchendurchschnitt besser einschätzen. Dabei ist aller-

[2] Das KombiFiD-Unternehmenspanel wird keine Einzelinformationen zu den im Unternehmen tätigen Personen enthalten; die Erstellung eines so genannten Linked-Employer-Employee – Datensatzes ist momentan nicht vorgesehen. Ob dies in einem Folgeprojekt möglich sein wird bleibt zu klären.

dings zu beachten, dass aus Gründen des Datenschutzes keine Einzeldaten für Unternehmen veröffentlicht oder Unternehmen zugänglich gemacht werden.

Wenn im Projekt KombiFiD die technische Realisierbarkeit einer Datenzusammenführung über die Grenzen von Datenproduzenten hinweg nachgewiesen wird, dann soll dies zum Anlass genommen werden, eine Initiative für eine rechtliche Regelung zu starten, die diese Form der Kombination von Unternehmensdaten dauerhaft und ohne hohe bürokratische Hürden ermöglicht. Dies kann die Belastung der Unternehmen durch Befragungen mit Auskunftspflicht spürbar senken. Vielfältige Angaben müssen seitens der Unternehmen heute mehrfach genannt und übermittelt werden. Die Möglichkeit einer Kombination der Angaben aus unterschiedlichen Befragungen schafft hier ein erhebliches Einsparpotential.

2. Datensätze für den kombinierten Unternehmenspanel-Datensatz

Der erste Schritt auf dem Weg zu dem im Projekt KombiFiD zu erstellenden kombinierten Unternehmenspanel-Datensatz ist die Auswahl der Statistiken, die miteinander kombiniert werden sollen. Seitens der Statistischen Ämter des Bundes und der Länder werden in das Projekt folgende Datenbestände eingebracht[3]:

- Unternehmensregister (als Masterdatei)
- Kostenstrukturerhebungen
- Gehalts- und Lohnstrukturerhebungen
- Steuerstatistiken (Körperschafts-, Gewerbe- und Umsatzsteuer)
- Erhebungen in Industriebetrieben (Monatsbericht, Kleinbetriebserhebung, Investitionserhebung)

[3] Wir versuchen hier einen weitestgehend vollständigen Überblick der für das Projekt sinnvollen Datenquellen zu geben. Die generelle Verfügbarkeit der Datensätze ist geklärt, allerdings kann es zu Beschränkungen der einzelnen Datensätze z.B. aus datenschutzrechtlicher Sicht kommen. Die Einkommensteuerstatistik, die ein Gutachter des Projektantrages gerne in der Liste der Datensätze gesehen hätte, fehlt hier, da es sich bei dieser Statistik um eine personenbezogene Statistik handelt. Diese kann aus datenschutzrechtlichen Gründen nicht berücksichtigt werden. Angaben zu den Berichtskreisen der Erhebungen und zu den erhobenen Merkmalen sowie Hinweise auf Publikationen über bzw. mit den Daten aus den Erhebungen finden sich in den jeweiligen Qualitätsberichten, die auf der Seite des Statistischen Bundesamtes unter www.forschungsdatenzentrum.de kostenlos bereit gestellt werden.

Im Projekt KombiFiD kann hier auf Ergebisse eines übergreifenden Projekts zur Integration wirtschaftsstatistischer Daten, das in der amtlichen Statistik unter Federführung des FDZ der Statistischen Landesämter durchgeführt wird, zurückgegriffen werden. In diesem Projekt werden das Unternehmensregister sowie ausgewählte wirtschaftsstatistische Erhebungen der amtlichen Statistik auf Ebene der Unternehmen bzw. Betriebe zu einem integrierten Datenbestand zusammengeführt. Um Doppelarbeiten zu vermeiden, werden die wirtschaftsstatistischen Einzelangaben der amtlichen Statistik, die für KombiFiD benötigt werden, vom FDZ der Statistischen Landesämter im Projekt „Amtliche Firmendaten Deutschlands“ (AFiD) aufbereitet und zur Verfügung gestellt.

Seitens des IAB der BA werden in das Projekt einerseits Datenbestände eingebracht, die aus den zu Betriebsangaben aggregierten Individualinformationen aus der Statistik der sozialversicherungspflichtig Beschäftigten gewonnen wurden, und die sich im Betriebs-Historik-Panel (BHP) finden. Diese Betriebsinformationen werden für das KombiFiD-Projekt erstmals zu Unternehmensinformationen zusammengefasst. Darüber hinaus können Daten aus dem IAB-Betriebspanel einfließen, wobei jedoch zu beachten ist, dass hierbei nur Angaben für Einbetriebsunternehmen verwendet werden können, denn bei Mehrbetriebsunternehmen ist nicht davon auszugehen, dass alle Betriebe eines Unternehmens auch in der Stichprobe des IAB-Betriebspanels enthalten sind.[4]

Das Unternehmensregister dient hierbei als „Masterdatei“. In diesem Register sind alle notwendigen Identifikatoren (wie etwa die Steuernummer oder die BA-Betriebsnummer) enthalten, die ein Zusammenführen der Daten aus verschiedenen Quellen ermöglichen.

3. Die KombiFiD - Stichprobe

Nach geltender Gesetzeslage ist das Zusammenführen wirtschaftsstatistischer Einzeldaten über die Grenzen der einzelnen Datenproduzenten hinweg nur bei einer schriftlichen Zustimmung der Auskunftgebenden und für ein zeitlich befristetes inhaltliches Projekt möglich. Da eine entsprechende Bitte um Zustimmung im begrenzten Rahmen des KombiFiD - Projekts nicht an alle Unternehmen gerichtet werden kann, ist

[4] Ausführliche Informationen zu in KombiFiD verwendeten Datenbeständen des IAB der BA und den Zugangsmöglichkeiten zu diesen Daten finden sich auf der Webseite des FDZ des IAB unter http://fdz.iab.de.

die Ziehung einer Stichprobe und die schriftliche Befragung der darin enthaltenen Unternehmen erforderlich.

Für diese Befragung wird – nach Abstimmung mit den Datenschützern des Statistischen Bundesamtes und des IAB[5] – eine Stichprobe von 50.000 Unternehmen die Frage nach der Erlaubnis zum Zusammenspielen ihrer Daten und gegebenenfalls weitere Fragen zu ihrem Informationsverhalten gestellt.[6] Die Frage zum Zusammenspiel der Daten lautet sinngemäß:

„Sie informieren in einer Reihe von Verfahren/Befragungen z.B. die Statistischen Ämter über Ihr Unternehmen. Hierbei müssen Sie oftmals die gleichen Daten für verschiedene Sachverhalte angeben. Daher streben die Statistischen Ämter des Bundes und der Länder sowie das Institut für Arbeitsmarkt- und Berufsforschung an, Sie bei Befragungen und Informationsbeschaffungen zu entlasten. Hierzu wollen wir versuchen, Ihre Informationen, die in verschiedenen Datenquellen verfügbar sind, zusammenzuführen. Dieses Zusammenführen erfolgt einmalig für Ihre Angaben aus dem Jahre 2003 bis zurück zum Jahre 1995. Das Zusammenführen dient der besseren Informationsgewinnung und Ihre Angaben werden ausschließlich für forschungsrelevante Fragestellungen verwendet. **Ihre Daten verlassen dabei nicht den abgeschotteten Bereich der beteiligten Institutionen. Ergebnisse werden nur so weitergegeben, dass das einzelne Unternehmen nicht erkennbar ist.**

Wir gehen davon aus, dass – bei einem erfolgreichen Abschluss des Projektes – eine Entlastung des Befragungsprogramms an Sie möglich ist. Gleichzeitig hoffen wir, dass durch die Kombination dieser Quellen, neue Erkenntnisse – beispielsweise für die Arbeitsmarktpolitik – entstehen, die dann politisch umgesetzt werden könnten. Zudem ist ein weiteres Ziel von KombiFiD, den Unternehmen eine bessere Informationsgrundlage für eigene Entscheidungen zur Verfügung zu stellen. Aus den Veröffentlichungen der wissenschaftlichen Arbeiten können die Unternehmen z.B. ihre Position verglichen mit dem Branchendurchschnitt besser einschätzen.

Wir bitten Sie, uns bei dieser Aufgabe zu unterstützen.

---

[5] Eine Sonderrolle nimmt hierbei das IAB-Betriebspanel ein, da dieses von einem Forschungsbereich des IAB in Zusammenarbeit mit TNS infratest durchgeführt wird. Hier steht eine Diskussion über das Procedere einer möglichen Datenbereitstellung an.

[6] Der hier benannte Wortlaut der Frage stellt nur einen ersten Arbeitsentwurf dar. Der genaue Wortlaut soll gemeinsam mit späteren Datennutzern, den jeweiligen Hausjuristen und dem Bundesdatenschutzbeauftragten erarbeitet werden. Hierzu ist eine eigene Arbeitsgruppe vorgesehen.

Sind Sie einverstanden, wenn wir Ihre vorhanden Daten ($Nennung der Dateien$ für ein zeitlich befristetes Forschungsprojekt zusammenführen?

Ja, ich bin einverstanden:

Datum, Unterschrift

Nein, ich bin nicht einverstanden, weil (*nach Möglichkeit eine Begründung*):

Datum, Unterschrift

Wir werden Sie gerne regelmäßig über den Stand der Arbeiten informieren.

Ja, ich habe Interesse an Informationen über den laufenden Stand der Arbeiten.

Nein, ich habe kein Interesse."

Für alle Unternehmen, die schriftlich diesem Zusammenführen zugestimmt haben, werden dann die Datensätze aus den genannten Statistiken zusammengeführt.

Die hierbei zu befragende Stichprobe kann keine einfache Zufallsstichprobe aus dem URS sein, da sie dann nur wenige größere Unternehmen enthalten würde. Außerdem liegen einige der oben genannten in das Projekt einzubeziehenden Statistiken selbst nur für Stichproben von Unternehmen vor, und es wäre bei einer Zufallsstichprobe nicht sicher gestellt – sondern eben Zufall – dass gerade die Unternehmen, die z.B. in den Kostenstrukturerhebungen erfasst wurden, auch in der KombiFiD-Stichprobe enthalten sind.

Ausgangspunkt für die Auswahl der Unternehmen, die schriftlich um eine Zustimmung zur Zusammenführung ihrer Daten gebeten werden sollen, sind daher die Kostenstrukturerhebungen und die Gehalts- und Lohnstrukturerhebungen, denn hierbei handelt es sich um Stichproben und nicht um Totalerhebungen für den jeweiligen Berichtskreis. Die Unternehmen, für die Angaben aus diesen Statistiken vorliegen, sind Teil der KombiFiD-Stichprobe. Weiterhin einbezogen werden dann Unternehmen aus Bereichen der Wirtschaft, die in diesen Statistiken nicht hinreichend enthalten sind, wobei die Details des Stichproben-Designs noch zu erarbeiten sind. Verfügbare Datensätze der Statistischen Ämter wären hierbei die Erhebung für industrielle Kleinbetriebe im Bergbau und Verarbeitenden Gewerbe, die jährliche Investitionserhebung bei Unternehmen des Bergbaus und des Verarbeitenden Gewerbes, die Monatsberichte für Betriebe des

Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden und die Steuerdateien zur Umsatzsteuer, Körperschaftssteuer und Gewerbesteuer.

4. Projektorganisation und Ablauf

KombiFiD ist ein Gemeinschaftsprojekt des FDZ des Statistischen Bundesamtes (Projektleitung: Dr. Markus Zwick), des FDZ der BA im IAB (Projektleitung: Stefan Bender) und des Instituts für Volkswirtschaftslehre / Empirische Wirtschaftsforschung der Leuphana Universität Lüneburg (Projektleitung: Prof. Dr. Joachim Wagner). Eine enge Kooperation besteht mit dem Projekt „Integration wirtschaftsstatistischer Daten der amtlichen Statistik“ des FDZ der Statistischen Landesämter.

Das Projekt ist in zwei Phasen aufgeteilt. In Phase 1 steht die Datenerschließung im Mittelpunkt. Hier werden ausgewählte Datensätze für die beschriebene Stichprobe von Unternehmen zusammengeführt. Ziel ist die Erstellung, Qualitätsprüfung und Dokumentation eines umfangreichen Unternehmenspanel-Datensatzes mit bisher in dieser Kombination in Deutschland nicht verfügbaren Informationen.

Der in Phase 1 erstellte Datensatz soll dann in Phase 2, in der der Fokus auf der Datenbereitstellung liegt, der Wissenschaft zur Verfügung gestellt werden. Die Daten sind hierbei während des gesamten Projektzeitraums ausschließlich im FDZ der Statistischen Ämter des Bundes und der Länder und im FDZ der BA im IAB vorhanden. Der Zugang zu diesen Daten ist in der Projektphase nur Forscherinnen und Forschern im Rahmen eines Nutzungsvertrages möglich. Der Zugang erfolgt ausschließlich über Datenfernverarbeitung und Gastaufenthalte. Eine Übermittlung der Unternehmensdaten (Einzeldaten) an Dritte findet nicht statt. Diese strenge Regelung hat z.B. die Konsequenz, dass Mitarbeiterinnen/Mitarbeiter des IAB oder der Statistischen Ämter ebenfalls einen Gastaufenthalt bzw. das Datenfernrechnen beantragen müssen, um mit diesen Daten arbeiten zu können.

In der zweiten Projektphase soll ferner geprüft werden, inwieweit der neue Datenbestand um Angaben ergänzt werden kann, die in der Deutschen Bundesbank vorhanden sind. Hierbei handelt es sich vor allem um Bilanzdaten und um Informationen über die Direktinvestitionsverflechtung der Unternehmen. Erste Gespräche mit Vertretern des Forschungszentrums der Deutschen Bundesbank haben hier Kooperationsmöglichkei-

ten deutlich gemacht. Weiter ist vorgesehen zu prüfen, ob die rechtliche Möglichkeit besteht, die Daten weiterer Datenproduzenten in das Projekt zu integrieren.

Das Projekt ist insgesamt auf drei Jahre angelegt und beginnt im Herbst 2007. Wir hoffen, den kombinierten Unternehmenspanel-Datensatz im Frühsommer 2008 für die Nutzung in den FDZ bereitstellen zu können. Zur frühzeitigen Integration der empirisch forschenden Sozial- und Wirtschaftswissenschaft ist im ersten Projektjahr eine Nutzerkonferenz vorgesehen. Hier soll neben der Datenproduktion insbesondere die spätere Nutzung der Daten thematisiert werden. Ferner wird ein wissenschaftlicher Beraterkreis (WBK) eingerichtet, der die weiteren Projektarbeiten begleitet.

5. Schlussbemerkungen

KombiFiD ist zweifellos ein Projekt mit ehrgeizigen Zielen, das als Machbarkeitsstudie konzipiert ist. Der Wissenschaft, den Datenschützern und der Politik soll in einem Pilotprojekt gezeigt werden, dass die Zusammenführung von Unternehmensdaten über die Grenzen der Datenproduzenten hinweg unter den gegebenen Bedingungen technisch und rechtlich realisierbar ist, und dass die so zusammengeführten Daten eine neue Dimension in der Analyse von amtlichen bzw. prozessproduzierten Daten eröffnen. Die Qualität der Analysen auf Grundlage des im Projekt erstellten bisher nicht verfügbaren Datenmaterials wird dabei ganz entscheidend von dem Rücklauf der Befragung, der Zusammenführungsquote und den möglicherweise bestehenden Inkonsistenzen in den zusammengeführten Datenmaterialien abhängen.

Wenn das Projekt erfolgreich ist, sollte alles unternommen werden, dass das Zusammenführen dieser Daten eine dauerhafte Perspektive hat. Damit soll erreicht werden, dass solche Datenbestände in der Zukunft systematisch und dauerhaft bereitgestellt werden können. Dies erfordert eine Änderung gesetzlicher Regelungen, denn heute muss – wie im hier vorgestellten Projekt - die schriftliche Zustimmung von jedem Auskunft gebenden Unternehmens hierzu eingeholt werden. Das Ziel muss sein, eine Rechtgrundlage zu schaffen, die das Zusammenführen wirtschaftsstatistischer Daten für Wissenschaft und Politikberatung ohne eine solche Einzelerlaubnis gestattet.

Imprint

FDZ Methodenreport
No. 5/2007

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Service
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Stefan Bender, Dagmar Herrlinger

Technical production
Dagmar Herrlinger

Download
http://doku.iab.de/fdz/reporte/2007/MR_05-07.pdf

Internet
http://fdz.iab.de/

Corresponding author
Stefan Bender, Tel.: +49-911/179-3082
E-Mail: stefan.bender@iab.de