Verknüpfung prozessgenerierter Firmendaten das Projekt KombiFiD

Tanja Hethey-Maier

Ausgangslage

•Das Institut für Arbeitsmarkt und Berufsforschung (IAB) bietet internen und externen Wissenschaftlern Zugang zu Betriebsdaten

•Diese Betriebsdaten kommen entweder aus IAB eigenen Befragungen oder aus den Meldungen zur Sozialversicherung der Bundesagentur für Arbeit (BA)

•Datenbestände:
- Betriebs-Historik-Panel (BHP)
- IAB Betriebspanel

Das Betriebs-Historik-Panel

German RLC

- Alle Betriebe des gesamtdeutschen Raumes mit mindestens einem sozialversicherungspflichtig Beschäftigten bzw. ab 1999 mindestens einem geringfügig Beschäftigten

- Stichtag: 30.6

- Zeitraum: 1975-2008

- Quelle: Arbeitgebermeldungen zur Sozialversicherung

Variablenspektrum BHP

- Betriebsmerkmale
  ID, Bundesland, Wirtschaftszweig, Gründung, Schließung

- Beschäftigtenstruktur
  Anzahl Beschäftigte nach Alter, Geschlecht, Bildung, Qualifikation

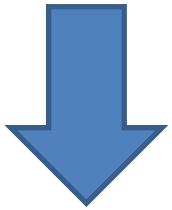
- Gehaltsstruktur
  Quartile Bruttotagesentgelt Vollzeitbeschäftigte nach Geschlecht, Bildung

- Beschäftigtenzugänge /abgänge
  gesamt, nach Alter, Geschlecht, Wiedereintritte, Betriebszugehörigkeitsdauer

Problematik

Verstärkter Wunsch nach Anreicherung der Beschäftigten- und Entgeltinformationen aus dem BHP um Performance Größen der Betriebe (z.B. Umsatz)

Viele Datensätze mit diesen Informationen liegen auf der Unternehmensebene vor (z.B. Dafne, Markus)



Verknüpfung von Betriebs- mit Unternehmensdaten

Das Projekt KombiFiD

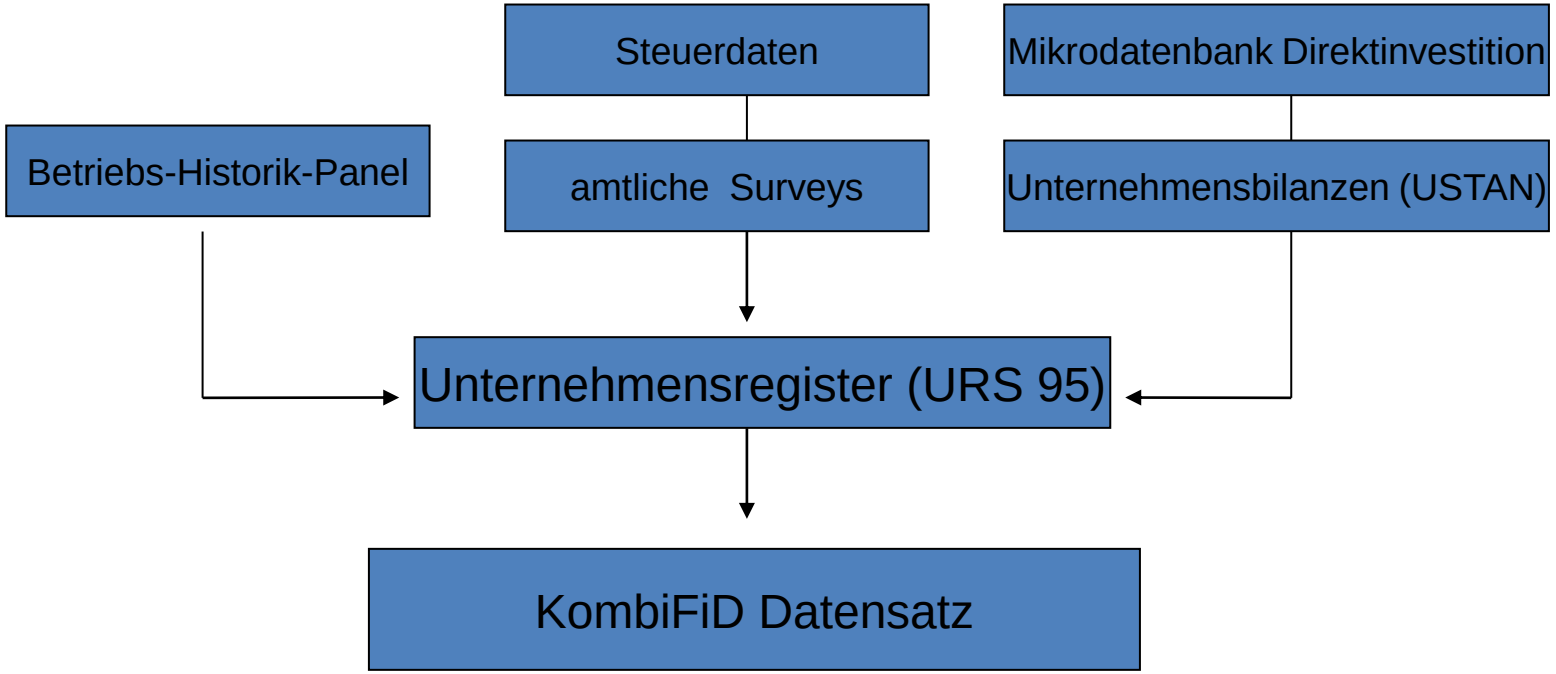
Projekttitel:	Kombinierte Firmendaten für Deutschland (KombiFiD)
Projektlaufzeit:	2007-2012
Finanzierung: (BMBF)	Bundesministerium für Bildung und Forschung
Koordination:	Statistisches Bundesamt

Institutionen übergreifende Verknüpfung von Firmendaten der

- Statistischen Ämter
- der Bundesagentur für Arbeit /IAB
- Deutsche Bundesbank

im Rahmen einer Machbarkeitsstudie (Stichprobe ca. 65.000 Firmen)

Datenbestände



Unternehmensregistersystem 95 (URS 95)

Gesetzliche Grundlage: Council Regulation No 2186/93 (1993)
Verwaltung/Pflege: Statistischen Ämter

Das URS 95 beinhaltet Angaben zu allen Unternehmen, deren wirtschaftliche Tätigkeit zum BIP zu Marktpreisen beiträgt, zu allen rechtlichen Einheiten, die für sie verantwortlich sind und zu allen örtlichen Einheiten, die von ihnen abhängen.

Aufbau/Pflege erfolgt dauch durch administrative Daten der BA

Liste aller BA Betriebsnummern zu einem Unternehmen

Aufgabe

Integration der Daten der Deutschen Bundesbank in den KombiFiD Datensatz mittels Namens- und Adressabgleich zwischen BHP und MiDi/USTAN

Schritt 1

Manuelle Suche von Unternehmen des HDAX im BHP

- Meldeverhalten von großen Unternehmen
- Struktur der Namens- und Adressinformation von Betrieben im BHP

Ergebnisse

•Unternehmensadresse entspricht IMMER einer Betriebsadresse

•IAB Daten speichern Strassenanschrift Postleitzahl, Unternehmensadresse enthält immer mal wieder Postfach oder Großempfänger Postleitzahlen (identisch auf 2-Steller Ebene)

•Art der Meldung (zentral in einer Betriebsnummer oder aufgeteilt) hängt neben der Unternehmensgröße stark vom Wirtschaftszweig ab

Schritt 2

Verknüpfung des BHP mit MiDi/USTAN über Namens- und Adressinformationen mittels Record-Linkage Techniken

Adressbestände

Deutsche Bundesbank

- MiDi (2006)
- USTAN (2006)
- Dafne /Hoppenstedt Auszug (2006) n= ca. 76.000

IAB

- IAB Betriebedatei (2006) n= ca. 2 Millionen

Datenaufbereitung (pre-processing)

Einheitliche Schreibweise

- Entfernung von Umlauten, Sonderzeichen
- nur Großbuchstaben
- Standardisierung von Namens- und Adressbestandteilen (z.B. Rechtsform)

Einheitliche Merkmale

- Name
- Rechtsform (Füllgrad IAB: ca. 65%)
- Straße
- Hausnummer
- Ort
- Postleitzahl

exakter Abgleich (Modelle)

exakt 1: Name, Rechtsform, Ort, Strasse, Hausnummer

Blocking: 3-Steller PLZ

exakt 2: Name, Rechtsform, 5-Steller PLZ, Strasse, Hausnummer

Blocking: 3-Steller PLZ

exakt 3: Name, Rechtsform, Ort, Strasse, Hausnummer

Blocking: 2-Steller PLZ

exakter Abgleich (Ergebnisse)

Modell	Anzahl übereinstimmende Variablen	Anzahl Unternehmen	Davon im KombiFiD Rücklauf gefunden
exakt 1	5+4	22.036	2.338
exakt 2	5+4	682	85
exakt 3	5+4	19	4
Summe	**5+4**	**22.737**	**2.427**

distanzbasierter Abgleich (Modelle)

link 1: Name,

Blocking: 5-Steller PLZ, Strasse, Hausnummer

link 2: Name

Blocking: 2-Steller PLZ, Ort, Strasse, Hausnummer

link 3: Name

Blocking: 5-Steller PLZ, Strasse

distanzbasierter Abgleich (Ergebnisse)

Modell	Anzahl übereinstimmende Variablen	Anzahl Unternehmen	Davon im KombiFiD Rücklauf gefunden
link 1	>=0,7 (manuelle Auswahl zwischen 0,6 -0,7)	15.435	1.396
link 2	>=0,7	114	22
link 3	>=0,7	2.976	307
Summe		**18.525**	**1.725**

Endstand: ca. 55% der Bundesbankunternehmen wurden gelinkt

Analyse der KombiFiD Unternehmen

gesamt	MiDi (2006)	USTAN (2006)	Dubletten
4.152	771	3.089	292

Dubletten:

- Mehrheit ist auf der Betriebsnummer eine Dublette
- wenige Fälle sind erst nach Aggregation auf Unternehmensebene Dubletten

Ergebnisse

•Pre-processing Programm für IAB Betriebsangaben

•Betriebedatei enthält NICHT einen Großteil der Unternehmenslandschaft (Rechtsformen die keine soz.vers.pfl. Beschäftigten aufweisen wie z.B. Holdings)

•Unternehmensnamen in der Betriebedatei variieren innerhalb des Unternehmens kaum

•Unternehmensnamen werden historisch schlecht gepflegt

•URS Schlüssel bildet historische Unternehmenszusammensetzung ab

weitere Schritte

Für KombiFiD:

•Validierung der gelinkten Einheiten durch Deutsche Bundesbank

Generell:

•Zugang zum gesamten URS !!!
•Umgang mit Namen- und Rechtsformangaben
•weitere linkage Modelle (geänderte Blocking Strategie)
•Alternative Wege der Umschlüsselung auf Unternehmensebene

Vielen Dank für Ihre Aufmerksamkeit!

Kontakt:
tanja.hethey-maier@iab.de