

Schätzung der Codierung gemäß NACE Rev. 2 / WZ 2008 für den Mikrozensus (MZ) / Labour Force Survey (LFS) 2008

<u>1.</u>	<u>Hintergrund</u>	2
<u>2.</u>	<u>Vorgaben Eurostats für den Übergang von der NACE Rev. 1.1 auf die NACE Rev. 2 im LFS</u> ...	2
<u>3.</u>	<u>Doppelcodierung im Mikrozensus 2009</u>	2
<u>4.</u>	<u>Schätzverfahren</u>	3

1. Hintergrund

Die Verordnung (EG) Nr. 1893/2006 des Europäischen Parlamentes und des Rates vom 20.12.2006¹ fordert für den LFS, dass die Mitgliedstaaten ab dem 1. Quartal 2008 die neue Wirtschaftszweigsystematik NACE Rev. 2 anwenden. Eine Reihe von Staaten – darunter auch Deutschland – konnte diese Anforderung nicht erfüllen. In Deutschland wurde im Rahmen des MZ/LFS 2008 noch die alte NACE Rev. 1 bzw. WZ 2003 erhoben².

Mit dem Europäischen Amt für Statistik (Eurostat) wurde vereinbart, dass die Mitgliedstaaten, die noch nicht ab dem 1. Quartal 2008 NACE Rev. 2-Einzeldaten liefern können, dies bis Ende Juni 2009 nachholen können. Somit ist eine nachträgliche Schätzung der Codierung gemäß NACE Rev. 2 bzw. WZ 2008 auf Ebene der Einzeldaten des MZ/LFS 2008 erforderlich.

2. Vorgaben Eurostats für den Übergang von der NACE Rev. 1.1 auf die NACE Rev. 2 im LFS

Ein eindeutiger Weg für den Übergang von der NACE Rev. 1.1 auf die neue NACE Rev. 2 im LFS ist von Eurostat nicht vorgegeben worden. Allerdings haben Eurostat und die nationalen Statistischen Institute der EU-Mitgliedstaaten ein Handbuch für alle EU-Statistiken zur Rückrechnung von Zeitreihen in Bezug auf die NACE Rev. 2 verfasst („Back Casting Handbook“³). Die dort beschriebenen Verfahren werden in Mikro- und Makroansätze unterteilt und setzen voraus, dass eine Doppelcodierung nach alter und neuer NACE für zumindest eine Referenzperiode⁴ durchgeführt wird. Mikroansätze zeichnen sich dadurch aus, dass mit Hilfe der Doppelcodierung für jede Erhebungseinheit ein Umstieg von der NACE Rev. 1.1 auf die NACE Rev. 2 erfolgt. Die WZ 2008-Schlüssel aus dem Jahr der Doppelcodierung werden den jeweiligen Erhebungseinheiten in allen Rückrechnungsjahren zugewiesen, in denen sie in der Stichprobe vorkommen. Bei einem Makroansatz erfolgt der Umstieg mit Hilfe eines aus der Doppelcodierung gewonnenen Umsteigerschlüssels für Aggregate bzw. Schätzwerte. Auf eine Verbindung zwischen historischen Mikro- und Makrodaten wird bei diesem Verfahren verzichtet. Dem Abschnitt 4 ist zu entnehmen, dass als Schätzverfahren für den MZ/LFS 2008 ein Mix aus Mikro- und Makroansatz verwendet wurde.

3. Doppelcodierung im Mikrozensus 2009

Im Mikrozensus 2009 wurde bei der gegenwärtig ausgeübten Haupterwerbstätigkeit sowohl die die WZ 2008 als auch die WZ 2003 (jeweils 3-Stellerebene) abgefragt. Dabei wurde in der computerunterstützten Erhebung durch Interviewer/-innen – wie sonst auch üblich – zunächst der vom Befragten genannte Klartext eingetragen. Anschließend erfolgte über das Lookup⁵ oder

¹ <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:393:0001:0039:DE:PDF>

² In Deutschland lag die Zuordnung des Stichwortverzeichnisses zur Codierung der WZ 2008 erst im Februar 2008 vor. Das Stichwortverzeichnis (über 30.000 Schlagwörter mit Vercodung) ist aber für die Erhebung von Wirtschaftszweigen in einer Bevölkerungsbefragung unverzichtbar. Allein die Gliederung der Systematik reicht nicht aus, um die Begrifflichkeiten der Befragten signieren zu können.

³ <http://circa.europa.eu/irc/dsis/nacecpacon/info/data/en/Back%20casting.doc>

⁴ In der NACE-Verordnung ist keine Referenzperiode für eine Doppelcodierung festgelegt.

⁵ Hierbei erfolgt die Signierung anhand der Eingabe eines Klartextes, in dessen Anschluss eine alphabetische Suche durchgeführt wird.

stufenweise über die Hierarchie die Signierung der WZ 2008. Mit der Signierung der WZ 2003 (ebenfalls über das Lookup oder stufenweise über die Hierarchie) wurde die Erfassung des Wirtschaftszweiges abgeschlossen. Der Klartext wurde also nur einmal eingegeben, aber nach zwei verschiedenen Systematiken signiert. Plausibilitätskontrollen im Falle unplausibler Doppelcodierungen waren nicht geschaltet, um die Dauer des Interviews nicht unzumutbar zu verlängern bzw. keine Interviewabbrüche zu provozieren. Dies führte zu 8,4 % fehlerhaften Zuordnungen. Als fehlerhaft wurde eine Zuordnung dann definiert, wenn die Zuordnung nicht mit dem Umsteigerschlüssel des Statistischen Bundesamtes entsprach.

Bei den Merkmalen „Wirtschaftszweig der früheren Erwerbstätigkeit“, „Wirtschaftszweig der Erwerbstätigkeit ein Jahr vor der Erhebung“ und „Wirtschaftszweig der Nebenerwerbstätigkeit“ wurde wegen des großen Aufwands für die Interviewer/-innen auf eine Doppelcodierung verzichtet. Unter Ertragsgesichtspunkten – eine vergleichsweise geringe Bedeutung dieser Merkmale in Veröffentlichungen Eurostats sowie ein möglicherweise nur geringer Gewinn an zusätzlichen Erkenntnissen – scheint diese Vorgehensweise auch gerechtfertigt.

4. Schätzverfahren

Für den MZ/LFS 2008 muss ein Umstieg von der NACE Rev. 1.1 (bzw. WZ 2003) auf die NACE Rev. 2 (bzw. WZ 2008) auf Mikrodatenebene erfolgen. Da für die Erhebungseinheiten in 2008 – insbesondere bis Juni 2009 – nur teilweise WZ 2008-Angaben aus der Doppelcodierung in 2009 übertragbar sind, ist der reine Mikroansatz im Sinne des „Back Casting Handbook“ nicht geeignet⁶. Stattdessen wird ein Verfahren gewählt, das ein Mix aus Mikro- und Makroansatz darstellt: Wie bei einem Makroansatz werden Umsteigerschlüssel ermittelt, die allerdings nicht auf Aggregate angewandt, sondern für eine wahrscheinlichkeitsbasierte Imputation von 3-Stellern der WZ 2008 auf Ebene der Einzeldaten herangezogen werden.

Als Schätzgrundlage für den Umstieg von der WZ 2003 auf die WZ 2008 im MZ/LFS 2008 wurde in erster Linie auf die Doppelcodierung im MZ/LFS 2009, daneben auch auf die Beschäftigtenstatistik (BS) der Bundesagentur für Arbeit (BA) zurückgegriffen. Die hauptsächliche Verwendung des MZ hat den Vorteil, dass im MZ/LFS 2008 Angaben gemäß WZ 2008 generiert werden, die eine möglichst große Zeitreihenkonsistenz gewährleisten. Die Vermeidung eines Zeitreihenbruchs wurde somit als wichtiger eingestuft als die Verwendung eines möglichst fehlerfreien Umsteigerschlüssels aus Registerstatistiken. Die BS der BA wurde nur für selten besetzte WZ 2008/WZ 2003-Kombinationen verwendet, die im bislang vorliegenden Einzelmaterial des MZ/LFS nicht enthalten sind (und möglicherweise auch keinen Eingang in die noch folgenden Quartalsstichproben oder die Jahresstichprobe finden werden).

Ein vornehmlich auf Basis der Doppelcodierung in 2009 ermittelter Umsteigerschlüssel wird umso genauer sein, je größer die zugrunde gelegte Stichprobenmasse des MZ/LFS ist. Da der LFS 2008-Datensatz mit NACE Rev. 2-Codierung bis Ende Juni 2009 an Eurostat zu liefern war, konnte nur ein Teil der maximalen Stichprobenmasse des MZ/LFS 2009 (= Jahresmaterial) als Schätzbasis verwendet werden.

⁶ Prinzipiell könnte aufgrund des Rotationsdesigns des Mikrozensus für maximal drei Viertel der Erhebungseinheiten in 2008 die WZ 2008-Angabe aus 2009 zugewiesen werden. Dies darf aber nur für Erhebungseinheiten erfolgen, die in 2008 dieselbe Tätigkeit ausgeübt haben wie in 2009. Dieser Personenkreis ließe sich über den Beginn der derzeit ausgeübten Tätigkeit identifizieren.

Bevor der Umsteiger WZ 2003 → WZ 2008 konstruiert wurde, erfolgte eine nachträgliche Plausibilisierung der WZ-Angaben im MZ/LFS 2009. Dabei wurde angenommen, dass Kombinationen jenseits des amtlichen Umsteigerschlüssels nicht auftreten dürfen, die WZ 2008-Angaben korrekt sind und unplausible Kombinationen gemäß amtlichem Umsteigerschlüssel allein auf Falschzuordnungen zur WZ 2003 beruhen. Letztere Annahme ist insofern kritisch, als die Interviewer/-innen im Erhebungsjahr 2009 an die WZ 2003, nicht aber an die WZ 2008 gewöhnt sind. Andererseits wurden die Wirtschaftszweige im Zuge der Doppelcodierung zunächst nach der WZ 2008 und erst danach nach der WZ 2003 signiert. Überdies wurde vonseiten der MZ-Länderreferenten berichtet, dass sich die stufenweise Signierung über die Hierarchie mit Einführung der WZ 2008 vereinfacht hat. Beides spricht für eine bessere Qualität der WZ 2008-Daten. Ausschlaggebend für die obige Annahme ist aber, dass eine Korrektur der WZ 2003-Daten zu einer größeren Zeitreihenkonsistenz beiträgt.

Bei der Korrektur der WZ 2003-Positionen (3-Steller) wurde wie folgt vorgegangen:

- (1) Gegenüberstellung der Angaben gemäß WZ 2008 und WZ 2003 (jeweils 3-Steller) aus der Doppelcodierung. Kombinationen, die dem amtlichen Umsteigerschlüssel entsprechen und empirisch im MZ/LFS 2009 vorkommen, wurden als „wahr“ eingestuft und gingen unverändert in die Berechnung des Umsteigers ein.
- (2) Ein Teil der nicht dem amtlichen Umsteigerschlüssel entsprechenden Kombinationen wurde „manuell“ durch Veränderung der WZ 2003-Codierung korrigiert.
 - a. Bei 1:1-Zuordnungen im amtlichen Umsteigerschlüssel wurden unzulässige WZ 2003-Positionen in Abhängigkeit von der WZ 2008-Position auf die einzig zulässige WZ 2003-Position korrigiert.
 - b. Im Falle mehrerer WZ 2003-Ex-Positionen wurden unzulässige WZ 2003-Gruppen in Abhängigkeit von der WZ 2008-Gruppe „intuitiv“ auf eine zulässige WZ 2003-Gruppe gesetzt. Intuitiv bedeutet hier, dass die Zuordnung auf Basis der „räumlichen“ Nähe in den Hierarchieebenen erfolgte. So ist bspw. laut dem Umsteiger von der WZ 2008 auf die WZ 2003 nur die Zuordnung 252 (WZ 2008) → 282 (WZ 2003) zulässig. Bei Zuordnungen 252 (WZ 2008) → 28x (WZ 2003), mit $x \neq 2$ wurde unterstellt, dass die Einordnung der WZ über die Hierarchieebenen erfolgte und eine falsche Stufe eingegeben wurde. Die Signierung 28x wurde dann auf 282 gesetzt.
- (3) Die übrigen unplausiblen WZ 2003-Gruppen wurden je WZ 2008-Gruppe in dem Code „999“ zusammengefasst und unter Heranziehung des Umsteigerschlüssels der BS der BA und der aus der Doppelcodierung stammenden empirischen Verteilung der WZ 2003-Gruppen innerhalb des Umsteigerbereichs einer WZ 2008-Gruppe zufallsbasiert korrigiert. Dabei wurden folgende Arbeitsschritte durchgeführt:
 - a. Ermittlung der Verteilung der WZ 2003-Positionen (nach Schritt (2)) im MZ/LFS 2009 pro WZ 2008-Gruppe.
 - b. Prüfung, ob laut amtlichem Umsteigerschlüssel existierende WZ 2008/WZ 2003-Kombinationen, die im MZ/LFS empirisch nicht vorkommen, im BS-Umsteiger auftreten.
 - Wenn ja: Die BS-Umsteigerschlüssel bzw. -gewichte der im MZ/LFS 2009 nicht auftretenden WZ 2003-Gruppen werden weiterverarbeitet.
 - Wenn nein: Die weder in der BS noch im MZ/LFS auftretenden WZ 2003-Gruppen werden als in Deutschland nicht vorkommend betrachtet.
 - c. In jeder WZ 2008-Gruppe mit WZ 2003-Positionen, die in der BS und nicht im MZ/LFS 2009 vorkommen, wird die Summe der BS-Gewichte dieser Kombinationen

verglichen mit dem Anteilswert der WZ 2003-Positionen, die im MZ/LFS 2009 vorkommen, außerhalb des amtlichen Umsteigerschlüssels liegen und noch nicht „manuell“ korrigiert wurden (Code „999“).

- ca. Ist der Anteilswert der noch unplausiblen WZ 2003-Positionen kleiner als die Summe der vorgenannten BS-Gewichte, werden diese WZ 2003-Positionen auf die im MZ/LFS 2009 nicht vorkommenden BS-Positionen verteilt, und zwar zufällig entsprechend der Häufigkeitsverteilung dieser BS-Positionen. Genauso wäre verfahren worden, wenn der Anteilswert der noch unplausiblen WZ 2003-Positionen der Summe der vorgenannten BS-Gewichte entsprochen hätte. Ein solcher Fall trat jedoch nicht auf.
- cb. Ist der Anteilswert der noch unplausiblen WZ 2003-Positionen größer als die Summe der vorgenannten BS-Gewichte, werden diese WZ 2003-Positionen auf die im MZ/LFS 2009 nicht vorkommenden BS-Positionen und die im MZ/LFS 2009 besetzten plausiblen bzw. plausibilisierten WZ 2003-Positionen verteilt, und zwar zufällig entsprechend der gemeinsamen Häufigkeitsverteilung der BS-Positionen und der im MZ/LFS (ggf. nach „intuitiver“ Korrektur) belegten WZ 2003-Positionen. Bei der Ermittlung dieser gemeinsamen Häufigkeitsverteilung wurde angenommen, dass die BS-Gewichte „wahre“ Werte darstellen. Deshalb wurden die Gewichte der im MZ/LFS (ggf. nach „intuitiver“ Korrektur) belegten WZ 2003-Positionen entsprechend verkleinert.
- d. Für die übrigen im MZ/LFS belegten WZ 2003-Gruppen, die nicht intuitiv oder mit Hilfe der BS korrigiert wurden und bei gegebener WZ 2008-Angabe außerhalb des amtlichen Umsteigerschlüssels liegen, erfolgt innerhalb der jeweiligen WZ 2008-Gruppen eine Zuordnung auf die anderen zulässig belegten WZ 2003-Positionen, und zwar zufällig entsprechend der Häufigkeitsverteilung dieser WZ 2003-Positionen.
- e. Der 3-Steller „023“ der WZ 2008 („Sammeln von wild wachsenden Produkten (ohne Holz)“) dürfte laut Auskunft der Gruppe E 1 in Deutschland nicht vorkommen. Im MZ/LFS 2009 ist er bislang nicht belegt, allerdings in sehr geringem Umfang in der BS (in Kombination mit den laut amtlichem Umsteigerschlüssel zulässigen WZ 2003-Positionen „011“ und „020“). Aufgrund der Auskunft der Gruppe E 1 fanden die BS-Gewichte für die WZ 2003-Positionen „011“ und „020“ keinen Eingang in die Schätzung des Umsteigers für den MZ/LFS 2008.

Im Anschluss an die Korrektur der WZ 2003-Angaben standen alle WZ 2008/WZ 2003-Kombinationen des MZ/LFS 2009 im Einklang mit dem amtlichen Umsteigerschlüssel. In nachfolgender Übersicht ist ersichtlich, in welchem Umfang unzulässige Doppelcodierungen durch die oben beschriebenen Maßnahmen korrigiert wurden:

	Unzulässige Doppelcodierungen ¹⁾	Korrekturmaßnahmen			
		2a	2b	3ca, 3cb	3d
Fälle	8 721	4 087	1 233	192	3 209
In %	100	46,9	14,1	2,2	36,8

¹⁾ Anzahl der gemäß amtlichem Umsteigerschlüssel unzulässigen WZ 2003-Gruppen, wenn angenommen wird, dass die Zuordnung zur WZ 2008 korrekt ist.

Die Anteilswerte der einer WZ 2003-Gruppe gegenübergestellten WZ 2008-Positionen stellen den Umsteigerschlüssel WZ 2003 → WZ 2008 für den MZ/LFS 2008 dar. Für die Schätzung der WZ 2008-Gruppen auf Einzeldatenebene wurden diese Anteilswerte als Wahrscheinlichkeiten interpretiert und für die Ermittlung einer Wahrscheinlichkeitsverteilung herangezogen. Für jede erwerbstätige Person (= Personen, die die Fragen zur gegenwärtig ausgeübten (Haupt-)Erwerbstätigkeit beantwortet haben) wurde dann eine Zufallszahl zwischen 0,00001 und 1,00000⁷ generiert, wobei jede Zahl die gleiche Wahrscheinlichkeit aufwies, gezogen zu werden. Auf Basis der Zufallszahl und der Wahrscheinlichkeitsverteilung erfolgte eine Zuordnung zu einer laut amtlichem Umsteigerschlüssel zulässigen WZ 2008-Gruppe.

Wird der ermittelte Umsteiger testweise im MZ/LFS-Material des 1. Quartals sowie des Aprils 2009 angewandt, ergibt sich eine geschätzte Verteilung auf die WZ 2008-Gruppen, deren Güte z. B. mit Hilfe des Dissimilaritätsindex und der tatsächlichen Verteilung überprüft werden kann. Der Dissimilaritätsindex misst den Abstand zweier Verteilungen. Für die mit dem Umsteiger generierte und tatsächliche Verteilung der Erwerbstätigen auf die WZ 2008-Gruppen ergibt sich ein Dissimilaritätsindex von 0,70, d.h. lediglich 0,70% der Erwerbstätigen müssten in der generierten Verteilung anders zugeordnet werden, um Gleichheit mit der tatsächlichen Verteilung zu erreichen. Daraus lässt sich ableiten, dass MZ/LFS-intern mit dem ermittelten Umsteiger WZ 2003 → WZ 2008 eine hohe Genauigkeit erzielt wird.

Für die „Variablen Wirtschaftszweig der gegenwärtigen Tätigkeit“, „Wirtschaftszweig der früheren Erwerbstätigkeit“, „Wirtschaftszweig der Erwerbstätigkeit ein Jahr vor der Erhebung“, „Wirtschaftszweig der Nebenerwerbstätigkeit“ wurden keine gesonderten Umsteiger ermittelt, da keine Informationen aus einer Doppelcodierung vorliegen. Für diese Variablen wurde derselbe Umsteiger verwendet wie für die gegenwärtig ausgeübte Haupterwerbstätigkeit. Daraus resultierende Verzerrungen werden als vernachlässigbar angesehen.

⁷ Fünf Nachkommastellen sind erforderlich, weil sehr kleine Anteilswerte von WZ 2003-Gruppen im Umsteigerbereich einer WZ 2008-Gruppe auftreten.