http://www.statistik.gv.at/

Methodenbericht

Modellbasierte Datenergänzung der Konjunkturstatistik im Produzierenden Bereich

Gerlinde Dinges
Fachbereich Methodik
Statistik Austria

Martin Haitzmann
Unternehmensstatistik
Statistik Austria

Abstract

Die im Jahr 1996 eingeführte EU-harmonisierte monatliche Konjunkturstatistik im Produzierenden Bereich stellt eine der zentralen Informationsquellen zur Beurteilung der konjunkturellen Entwicklung Österreichs und des gesamten Europäischen Wirtschafts- und Währungsraums dar. Ihre nationalen und internationalen Verwendungszwecke sind vielfältig. Dem Anliegen der Wirtschaft entsprechend wird diese Statistik als Vollerhebung mit variablen Abschneidegrenzen (Schwellenwerte) unter Berücksichtigung eines standardisierten Repräsentanzkriteriums geführt, wodurch Klein- und Kleinstunternehmen (rund 80% der Grundgesamtheit) von statistischen Verpflichtungen weitestgehend befreit werden. Um darüber hinaus auch den Bedürfnissen der Datennutzer zu entsprechen und EU-Vorgaben hinsichtlich vorgegebener Repräsentanzkriterien vollständig erfüllen zu können, erfolgt ab 2009 zusätzlich zur bisherigen Veröffentlichung primärstatistisch erfasster Ergebnisse der größeren meldepflichtigen Einheiten auch eine entsprechende Ergebnisdarstellung der Wirtschaftsleistung aller Unternehmen und Betriebe, die im Unternehmensregister der STATISTIK AUSTRIA als aktive Einheiten der Grundgesamtheit des Produzierenden Bereichs erfasst sind. Der folgende Beitrag ist vor allem der methodischen Konzeption des Modells zur Datenergänzung und dessen Umsetzung in der Praxis gewidmet, wobei Ergebnisse und Datenqualität auch aus Sicht der Unternehmensstatistikexperten beleuchtet werden. Zur praktischen Veranschaulichung dienen ausgewählte Wirtschaftszweige und Merkmale der Konjunkturstatistik 2009.

1. Einleitung

Die Konjunkturstatistik im Produzierenden Bereich (KJP) besteht in dieser Form seit 1996 und wird nach einem Konzept durchgeführt, das den EU-Vorgaben entspricht. Sie dient dazu, Informationen über kurzfristig beobachtbare und messbare Phänomene im Bereich der Sachgüterproduktion und im Bauwesen zu sammeln und zu verarbeiten. Die Ergebnisse der KJP bilden - bei Anwendung geeigneter statistischer Methoden - eine der wesentlichen Grundlagen für die Beobachtung des Wirtschaftszyklus und liefern politischen Entscheidungsträgern auf nationaler wie auch europäischer Ebene sowie für die Europäische Zentralbank wichtige empirische Evidenz. Darüber hinaus bietet die KJP den Marktproduzenten selbst harmonisierte Informationen zum Verständnis ihrer Märkte und zum Vergleich ihrer Tätigkeit und Leistung mit Wettbewerbern desselben Wirtschaftszweigs auf nationaler und internationaler Ebene.

Dem Anliegen der Wirtschaft entsprechend wird die KJP als **Vollerhebung mit variablen Abschneidegrenzen** unter Berücksichtigung eines standardisierten Repräsentanzkriteriums geführt. Diese Erhebungsform wird auch als **„Konzentrationsstichprobe"** bezeichnet. Auskunftspflicht besteht nur bei Überschreitung normierter Schwellenwerte (grundsätzlich Beschäftigtenschwelle, bei Nichterreichung eines bestimmten Repräsentanzgrades zusätzlich Auskunftspflicht unterhalb der Beschäftigtenschwelle bei Überschreitung einer vorgegebenen Umsatzschwelle), wodurch Klein- und Kleinstunternehmen von statistischen Verpflichtungen weitestgehend ausgenommen sind.
Um den Bedürfnissen der Datennutzer zu entsprechen und EU-Vorgaben hinsichtlich vorgegebener **Repräsentanzkriterien** vollständig zu erfüllen, erfolgt ab 2009 neben der bisherigen Veröffentlichung der primärstatistisch erfassten Ergebnisse für die größeren Unternehmen und Betriebe auch eine entsprechende Ergebnisdarstellung für die statistische Grundgesamtheit des Produzierenden Bereichs.
Da die vorgegebenen rechtlichen und daraus resultierenden konzeptiven Rahmenbedingungen keine konventionelle Hochrechnung zulassen,[1] wurde ein **modellbasierter Ansatz zur Datenergänzung** entwickelt, welcher den Informationsvorrat der Primärerhebung der KJP ausschöpft und zusätzlich Informationen aus Verwaltungsquellen effizient nutzt.
Ein ähnlicher Weg zur Berechnung der Grundgesamtheit wird seit mehreren Jahren bereits in den jährlichen Leistungs- und Strukturstatistiken[2] durch Einbindung von einerseits **primärstatistischen Ergebnissen**[3] und andererseits **sekundärstatistischen Datenquellen**[4] beschritten. Die ungleich höhere Komplexität der KJP gestaltet die Erstellung eines plausiblen Datenkörpers für die statistische Grundgesamtheit jedoch besonders schwierig, denn im Gegensatz zu den jährlichen Leistungs- und Strukturstatistiken müssen in der KJP kurzfristige monatliche Ergebnisse vorliegen, und darüber hinaus ist zusätzlich zum Aktivitätsansatz (Ergebnisdarstellung nach der schwerpunktmäßigen wirtschaftlichen Aktivität der Unternehmen und Betriebe) eine Darstellung nach dem Güteransatz (Ergebnisdarstellung aller gleichartig klassifizierten Güter) erforderlich.

Im vorliegenden Methodenbericht wird die Konzeption des Modells zur Datenergänzung und dessen Umsetzung in der Praxis beschrieben.

2. Monatliche Primärerhebung der KJP

Bei der KJP[5] handelt es sich um ein Instrument mit zwei unterschiedlichen Zielsetzungen.

[1] Meldepflicht besteht nur bei Überschreitung normierter Schwellenwerte, folglich liefert dieses Auswahlverfahren kein repräsentatives Sample für eine konventionelle Hochrechnung auf die Grundgesamtheit.

[2] Siehe auch Standarddokumentation zur Leistungs- und Strukturstatistik im Produzierenden Bereich auf der Homepage der STATISTIK AUSTRIA unter www.statistik.at > Dokumentationen > Produktion und Bauwesen.

[3] Primärdaten werden eigens für die betreffende Statistik direkt beim auskunftspflichtigen Unternehmen bzw. Betrieb erhoben.

[4] Sekundärdaten können aus Register- oder Verwaltungsquellen gewonnen werden oder stehen aus anderen statistischen Erhebungen zur Verfügung.

[5] Eine ausführliche Methodenbeschreibung findet sich in der Standarddokumentation zur KJP auf der Homepage der STATISTIK AUSTRIA unter www.statistik.at > Dokumentationen > Produktion und Bauwesen.

Es sollen einerseits die empirischen Voraussetzungen für Konjunkturanalysen für den Produzierenden Bereich geschaffen und andererseits das heimische Güteraufkommen dieses Wirtschaftsbereichs abgebildet werden. Somit werden mit der monatlichen KJP die Vorgaben der **EU-Verordnung über Konjunkturstatistiken**, Anhang A (Industrie) und Anhang B (Baugewerbe),[6] und jene zur Erfassung der **nationalen Güterproduktion im Sinne der EU-Verordnung zur Einführung einer Gemeinschaftserhebung über die Produktion von Gütern** (PRODCOM)[7] erfüllt. Grundsätzlich erstreckt sich die KJP auf alle Unternehmen, Betriebe (fachliche Einheiten), Arbeitsgemeinschaften (ARGEN) und Betriebe gewerblicher Art von Körperschaften des öffentlichen Rechts gemäß § 2 KStG 1988,[8] die den **Abschnitten der ÖNACE 2008**[9]

- Bergbau und Gewinnung von Steinen und Erden (Abschnitt B),
- Herstellung von Waren (Abschnitt C),
- Energieversorgung (Abschnitt D),
- Wasserversorgung; Abwasser- und Abfallentsorgung und Beseitigung von Umweltverschmutzungen (Abschnitt E) sowie
- Bauwesen (Abschnitt F)

zuzuordnen waren und die Tätigkeit selbständig, regelmäßig und in der Absicht zur Erzielung eines Ertrags oder sonstigen wirtschaftlichen Vorteils ausüben.

Dabei besteht **Auskunftspflicht** für

- alle **Unternehmen** (Ein- und Mehrbetriebsunternehmen) und **Betriebe** (fachliche Einheiten) sowie Betriebe gewerblicher Art von Körperschaften des öffentlichen Rechts gemäß § 2 KStG 1988, die am 30. September des der Berichtsperiode vorangegangenen Kalenderjahres **20 und mehr Beschäftigte** hatten,
- alle **Arbeitsgemeinschaften** (ARGEN), unabhängig von der Beschäftigtenzahl, ab ihrer Gründung bis zu ihrer Auflösung sowie
- alle im Kalenderjahr der Berichtsperiode **neu gegründeten** bzw. durch Umstrukturierung **entstandenen statistischen Einheiten**, in Abhängigkeit von der Beschäftigtenzahl zum Zeitpunkt der Neugründung bzw. Umstrukturierung.

Darüber hinaus muss die Erhebungsmasse mindestens 90% des Gesamtumsatzes in jedem der Wirtschaftszweige gemäß den Abteilungen 05 bis 43 der ÖNACE 2008 enthalten (Repräsentanz bzw. Deckungsgrad). Wird dieses **Repräsentanzkriterium** mit Hilfe der voll erhobenen

[6] Verordnung (EG) Nr. 1165/98 über Konjunkturstatistiken, ABl. Nr. L 162 vom 05.06.1998 S. 1, in der Fassung der Verordnung (EG) Nr. 1178/2008.

[7] Verordnung (EWG) Nr. 3924/91 zur Einführung einer Gemeinschaftserhebung über die Produktion von Gütern, ABl. Nr. L 374 vom 31.12.1991, S. 1, zuletzt geändert durch Verordnung (EG) Nr. 1893/2006 zur Aufstellung der statistischen Systematik der Wirtschaftszweige NACE Revision 2 und zur Änderung der Verordnung (EWG) Nr. 3037/90 des Rates sowie einiger Verordnungen der EG über bestimmte Bereiche der Statistik, ABl. Nr. L 393 vom 30.12.2006 S. 1. bzw. vergleiche auch www.statistik.at > Klassifikationen > Klassifikationsdatenbank > Güter > CPA 2008.

[8] Bundesgesetz vom 7. Juli 1988 über die Besteuerung des Einkommens von Körperschaften (Körperschaftsteuergesetz 1988 - KStG 1988), BGBl. Nr.401/1988, idgF.

[9] Vergl. Publikation der STATISTIK AUSTRIA „Systematik der Wirtschaftstätigkeiten - ÖNACE 2008; Band 1 und 2“ bzw. unter www.statistik. at > Klassifikationen > Klassifikationsdatenbank > Wirtschaftszweige > ÖNACE 2008.

Schicht nicht erreicht, so besteht Auskunftspflicht auch über statistische Einheiten mit weniger als 20 Beschäftigten, die in den zwölf Monaten, die dem Stichtag[10] vorangegangen sind, in Summe einen Umsatz (exklusive Umsatzsteuer) von mindestens einer Million Euro erzielten (eingeschränkte Auskunftspflicht unterhalb der gesetzlichen Beschäftigtenschwelle).[11]

3. Weshalb die KJP-Primärerhebung modellbasiert ergänzen?

3.1. Nichterfüllung der Repräsentanz

Das Bundesstatistikgesetz 2000 enthält die Verpflichtung, **Klein- und Kleinstunternehmen in höchstmöglicher Weise von der Auskunftspflicht auszunehmen**. Den Bestrebungen, den geforderten Abdeckungsgrad im Sinne der europäischen Vorgaben durch eine primärstatistische Erhebung zu erreichen, wird damit eine natürliche Grenze gesetzt.
Die nationale Durchführungsverordnung[12] sieht als **Repräsentanzkriterium die Erfassung von 90% des Gesamtumsatzes aller in diesem Zweig tätigen Einheiten je ÖNACE-2008-Abteilung** vor. Die EU-Verordnung über die Güterproduktion[13] normiert den Deckungsgrad von 90% bezogen auf die Inlandsproduktion auf noch tieferer Gliederungsebene, nämlich je NACE-Rev.2-Klasse (entspricht der ÖNACE-2008-Klasse). Nach der nationalen Gesetzeslage dürfen Klein- und Kleinstunternehmen mit weniger als 20 Beschäftigten und einem Jahresumsatz von weniger als einer Million Euro selbst dann nicht zur Erfüllung des Repräsentanzkriteriums in die Erhebung einbezogen werden, wenn der branchenspezifische Gesamtumsatz der in die Erhebung einbezogenen statistischen Einheiten zur Erreichung des Deckungsgrades nicht ausreicht.

Wie aus den Abbildungen 1 und 2 ersichtlich, unterscheidet sich die Primärabdeckung der verschiedenen Wirtschaftsbereiche teils deutlich, was letztendlich auch die Vergleichbarkeit bzw. einheitliche Interpretation von Ergebnissen erschwert. Zusätzlich wird in einigen Wirtschaftsbereichen durch die primärstatistische Erhebung weder das Repräsentanzkriterium der nationalen Durchführungsverordnung (Abbildung 1) noch jenes der PRODCOM-Verordnung (Abbildung 2) erreicht.
Gemäß den europäischen Bestimmungen können jedoch die Mitgliedstaaten die erforderlichen Daten nach dem Grundsatz der verwaltungstechnischen Vereinfachung durch eine Kombination verschiedener Quellen (verbindliche Erhebungen, andere Quellen, die in Bezug auf Genauigkeit und Qualität zumindest gleichwertig sind, oder statistische Schätzverfahren) beschaffen. Dieser Weg wurde nun mit der Modellbasierten Datenergänzung im Produzierenden Bereich beschritten, mit dem Ziel, ein Maximum an Repräsentanz und damit einhergehend, ein Optimum an Qualität des Datenkörpers zu gewährleisten.

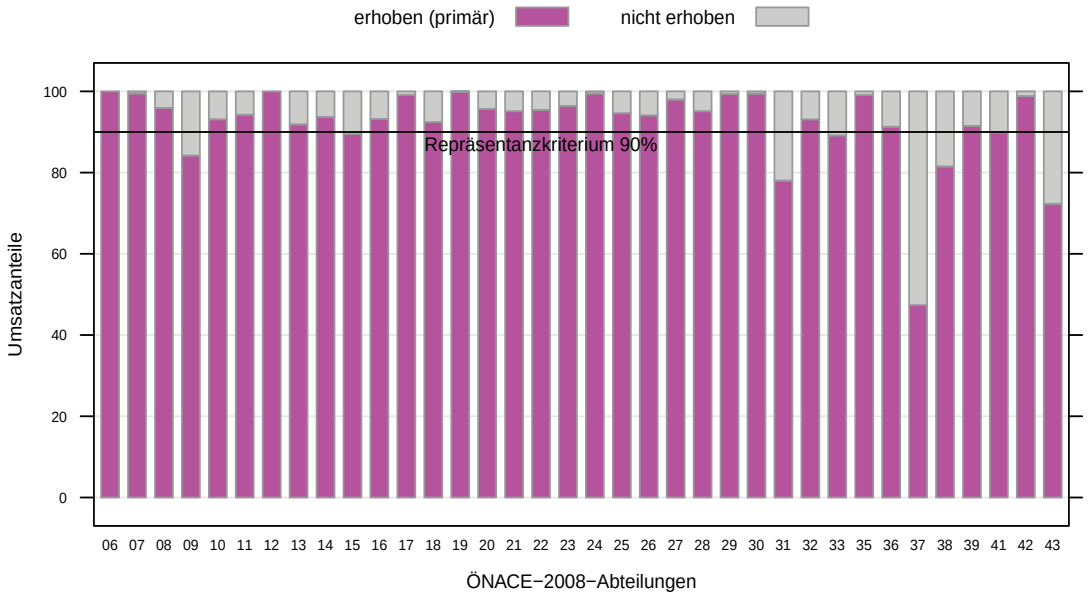
[10] Stichtag ist der 30. September des der Berichtsperiode vorangegangenen Kalenderjahres.

[11] Mehr auf: www.statistik.at > Fragebögen > Unternehmen > Konjunkturerhebung im Produzierenden Bereich.

[12] BGBI. II Nr. 210/2003, zuletzt geändert durch BGBI. II Nr. 315/2007.

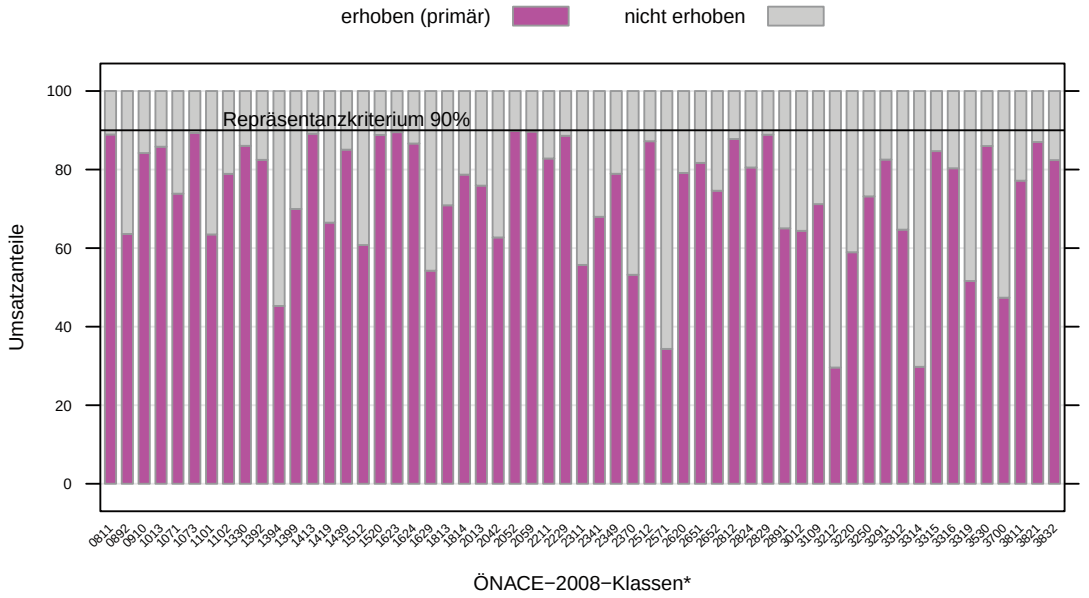
[13] Verordnung (EWG) Nr. 3924/91 des Rates vom 19. Dezember 1991 zur Einführung einer Gemeinschaftserhebung über die Produktion von Gütern, (ABl. Nr. L 196 vom 31. Dezember 1991, S.1), zuletzt geändert durch die VO (EG) Nr. 1893/2006.

Abbildung 1: Erfassung des Umsatzes im Produzierenden Bereich durch die KJP-Erhebung im 1. Quartal 2009 (Gemäß nationaler Durchführungsverordnung)



Quelle: KJP 2009 (1. Quartal)

Abbildung 2: Erfassung des Umsatzes im Produzierenden Bereich durch die KJP-Erhebung im 1. Quartal 2009 (Gemäß Europäischer Verordnung über die Güterproduktion)

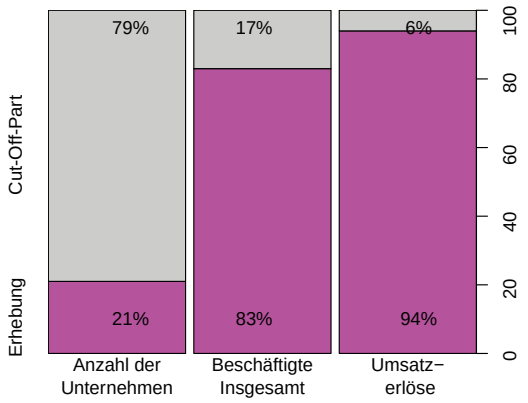


Quelle: KJP 2009 (1. Quartal). - *) Nur ÖNACE-2008-Klassen, die die Mindestrepräsentanz nicht erfüllen.

3.2. Konzeptionelle Vorgaben der Primärerhebung der KJP

Das Erhebungskonzept der KJP wird auch als **Cut-Off-Census**[14] oder „Konzentrationsstichprobe“ bezeichnet, da weniger (ge-)wichtige Information nicht erfasst wird. Der „bewusste Daten-Cut-Off“ liefert Informationen, die keinem repräsentativen Sample im Sinne der Stichprobentheorie entsprechen. In der Wirtschaftstatistik sind Konzentrationserhebungen gängige Praxis. Grundidee dieser Methode ist es, die hohe Merkmalskonzentration wirtschaftsstatistischer Daten zu nutzen und anhand möglichst weniger erhobener Einheiten eine möglichst hohe primärstatistische Abdeckung jener Merkmale zu erreichen, die im Mittelpunkt des Interesses stehen. Abbildung 3 zeigt, dass im 1. Berichtsquartal 2009 nur Daten für 21% der rund 58.000 Unternehmen erhoben werden mussten, um bereits 83% der Gesamtbeschäftigten und 94% der Umsatzerlöse primärstatistisch zu erfassen. Dabei ist jedoch zu bedenken, dass, auch wenn ein bestimmtes vorgegebenes Repräsentanzkriterium erreicht werden kann, diese bewusste Beschränkung auf die „wesentlichen Elemente“ die Annahme eines „vernachlässigbaren Beitrags“ des Cut-Off-Parts impliziert. Für die Berechnung der Konjunkturindikatoren zum Beispiel ist diese Beschränkung auf die „wesentlichen Elemente“ per definitionem ausreichend, manche Analysen, für detaillierte Konjunkturprognosen etwa, erfordern aber vollständige Daten über die statistische Grundgesamtheit.

Abbildung 3: Merkmalskonzentration in der KJP-Grundgesamtheit
Primärabdeckung ausgewählter Unternehmensmerkmale



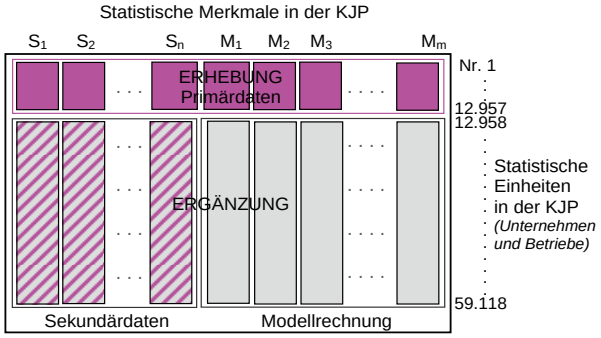
Quelle: KJP 2009 (1. Quartal).

4. Modellbasierte Datenergänzung (MDE)

Durch die **Modellbasierte Datenergänzung (MDE)** wird zu Beginn eines Berichtsjahres (Berichtsmonat Jänner) der primärstatistisch erhobene Datenkörper - wie in Abbildung 4 skizziert - um den ursprünglichen „Cut-Off-Part“ ergänzt. Die Struktur der Grundgesamtheit kann durch die MDE als Kombination aus Primärdaten, Sekundärdaten und Modellrechnung ohne grundsätzlichen Informationsverlust dargestellt werden.

[14] Ein *Cut-Off-Census* ist die Vollerhebung aller Einheiten, die eine bestimmte Schwelle überschreiten. Ein *Cut-Off-Sample* ist eine Stichprobenziehung nur aus den Einheiten, die eine bestimmte Schwelle überschreiten.

Abbildung 4: Datenkörper der KJP-Grundgesamtheit



Quelle: KJP 2009 (Berichtsmonat Jänner).

Die Datenerstellung erfolgt grundsätzlich auf **Mikroebene**. Für jede statistische Einheit der KJP-Grundgesamtheit steht somit ein vollständiger Datensatz für weitere Ergebnisdarstellungen zur Verfügung. Bei den zu ergänzenden statistischen Einheiten handelt es sich in der Regel um **Einbetriebsunternehmen**, vereinzelt können aber auch Mehrbetriebsunternehmen (und deren Betriebe) in die Schätzbasis fallen (im Jänner 2009 scheinen fünf Fälle auf). Diese Ausnahmefälle werden bei der MDE jedoch als „vernachlässigbar“[15] betrachtet, weshalb bei der weiteren Schätzmodellbeschreibung nicht ausdrücklich auf diese Unterscheidung hingewiesen wird. Grundsätzlich werden in weiterer Folge also Unternehmen und Betriebe gleichgesetzt.

4.1. Schritte zur Mikrodatenerstellung

Zu Beginn eines Berichtsjahres (Berichtsmonat Jänner) wird die Grundgesamtheit N (59.118 Unternehmen und Betriebe im Berichtsjahr 2009) ermittelt. Eine statistische Einheit ist genau dann der Grundgesamtheit der KJP zugeordnet, wenn sie gemäß Unternehmensregister der STATISTIK AUSTRIA (UR)[16] im Berichtszeitraum wirtschaftlich aktiv ist und eine Haupttätigkeit im Sinne einer Unterklasse der ÖNACE 2008 ausübt, welche dem Produzierenden Bereich zuzurechnen ist. Die Schätzbasis umfasst im Berichtsjahr 2009 rund 46.000 nicht meldepflichtige Einheiten. Auch einige Meldeausfälle ohne historische Struktur fallen in die Schätzbasis.[17] Für die zu ergänzenden statistischen Einheiten $n^{(erg)}$ können sogenannte **Eckdaten** (wirtschaftliche Tätigkeit, Anzahl der unselbständig Beschäftigten, monatliche Umsatzerlöse) unter Einbindung von Verwaltungsquellen erstellt werden. Andere Merkmale (wie

[15] Das bedeutet, dass beim selten auftretenden Fall von Mehrbetriebsunternehmen in der Schätzbasis, der Unternehmensumsatz und die Beschäftigten, die aus Verwaltungsquellen für das Berichtsmonat verfügbar sind, über Registerinformationen anteilsmäßig auf das Unternehmen und seine Betriebe aufgeteilt werden. Die so erzeugten Dummy-Einheiten werden im Schätzmodell wie Einbetriebsunternehmen behandelt. Nach dem Schätzvorgang werden die Unternehmensmerkmale über die Betriebe ermittelt und die relevanten Merkmalsblöcke der unterschiedlichen statistischen Einheiten erstellt.

[16] Das UR erfasst alle Einheiten mit mindestens einem unselbständig Beschäftigten oder mehr als 10.000 Euro Jahresumsatz sowie Einheiten des Staates und Non-Profit Organisationen.

[17] Fälle von Unit-Non-Response werden in der KJP über Vorperiodenmeldungen anhand der „historischen Struktur“ des Meldeausfalls imputiert. Für Unit-Non-Response ohne Meldung aus Vorperioden fehlt diese Substitutionsgrundlage, weshalb diese Einheiten durch die MDE in der Schätzbasis mitberücksichtigt werden.

Arbeitsvolumen, Arbeitskosten, Produktion eines Unternehmens) werden über diese Eckdaten, basierend auf aktuellen primärstatistischen Zusammenhängen, mittels geeigneter Schätzmodelle erzeugt. **Beschäftigtenbezogene Merkmale** (*Löhne, Arbeitsstunden* usw.) werden dabei über ein robustes lineares Regressionsmodell berechnet, während für **umsatzbezogene Merkmale** (*abgesetzte Produktion, technische Gesamtproduktion*, ...) ein kombinierter Ansatz angewendet wird, dessen Ziel es ist, für die statistischen Einheiten der verschiedenen Wirtschaftszweige eine möglichst plausible Verteilung der Produktionsarten und deren Darstellung auf Güterebene (ÖPRODCOM-Gliederung) zu erhalten. Beim Güteransatz erfolgt die Darstellung der Merkmalsgruppen nach der statistischen Einheit *Betrieb*. Beim Aktivitätsansatz werden Merkmalsgruppen nach den statistischen Einheiten *Unternehmen und Betrieb* in Verbindung mit der schwerpunktmäßigen wirtschaftlichen Aktivität derselben dargestellt. Vorrangiges Ziel der MDE ist es, Mikrodaten zu erzeugen, die eine beliebige Ergebnisdarstellung für die **Grundgesamtheit der KJP** auf **Aktivitätsebene** und auf **Güterebene** erlauben.

4.2. Schätzmodell auf Aktivitätsebene

Die Schätzung erfolgt auf Aktivitätsebene über die schwerpunktmäßige wirtschaftliche Aktivität einer statistischen Einheit. Konzeptionell wurde hier das für die Leistungs- und Strukturstatistik entwickelte LS-Schätzmodell[18] implementiert, wobei jedoch aufgrund abweichender Merkmalskataloge und der zum Zeitpunkt der Schätzung unterschiedlichen Verfügbarkeit von Verwaltungsdaten entsprechende Modelladaptierungen erforderlich waren. Im Wesentlichen basieren die Überlegungen zum KJP-Grundmodell jedoch auf Erfahrungen und Erkenntnissen, die im Zuge der Entwicklung des LS-Schätzmodells gewonnen wurden. Grundidee und Vorgehensweise der MDE auf Aktivitätsebene werden in diesem Abschnitt erläutert.

Erstellung der Eckdaten

Über die **Verknüpfung der Unternehmen des UR** mit den vorliegenden Umsatzsteuervoranmeldungen (UVA) der **Finanzbehörde** sowie den Beschäftigtenmeldungen des **Hauptverbands der österreichischen Sozialversicherungsträger (HV)** erfolgt für jede einzelne der $n^{(erg)}$ Einheiten der Schätzbasis zuerst die Erstellung der zugehörigen Eckdaten. Beginnend mit der Zuweisung der jeweiligen unternehmens- bzw. betriebsspezifischen Merkmale, wie etwa Wirtschaftsaktivität oder Rechtsform, die aus dem UR für jede statistische Einheit zur Verfügung stehen, erfolgt anschließend die Erweiterung dieser Eckdaten durch die Einbindung der unselbständig Beschäftigten des HV, gegliedert nach Geschlecht und Qualifikationen (Arbeiter, Angestellte, Lehrlinge), und durch die Übernahme der monatlichen UVA-Meldungen eines Unternehmens. Bei fehlender Beschäftigtenmeldung vom HV wird aufgrund des hohen Verknüpfungsgrades im UR davon ausgegangen, dass das Unternehmen keine unselbständig Beschäftigten hat. Fehlende UVA-Meldungen[19] werden unter Berücksichtigung

[18] Ein ausführlicher *Methodenbericht zur Modellbasierten Datenergänzung in der Leistungs- und Strukturstatistik* erscheint im 1. Quartal 2010. Eine kurze Methodenbeschreibung findet sich auch in der Standarddokumentation zur Leistungs- und Strukturstatistik im Produzierenden Bereich auf der Homepage der STATISTIK AUSTRIA unter www.statistik.at > Dokumentationen > Produktion und Bauwesen.

[19] Keine fristgerechte Erstattung der UVA-Meldung bei den Finanzbehörden; keine UVA-Meldepflicht für Einheiten mit Jahresumsatz <100.000 €, etc.

der individuellen Unternehmensentwicklung und der zugehörigen Branchenentwicklung bzw. auch anhand historischer Informationen imputiert. Für die weiteren Berechnungen gilt:

Die KJP-Grundgesamtheit ist unterteilt in L Branchen $B_1, ..., B_L$ mit $B = \bigcup_{j=1}^{L} B_j$, $B_j \cap B_r = \emptyset$.

$j = 1, ..., L$ ... laufender Branchenindex

$i = 1, ..., N$ ... Index der statistischen Einheiten

n_j ... Anzahl der statistischen Einheiten einer Branche j

$N = \sum_{j=1}^{L} n_j$ ... Anzahl der statistischen Einheiten insgesamt ($N = n^{(erh)} + n^{(erg)}$)

$UVA_{i,m}$ ... UVA-Meldung einer statistischen Einheit i im Berichtsmonat m

$M_i := \{(m-t) : t \in \{0, 1, ..., 11\}, UVA_{i,(m-t)} \text{ liegt vor }\}$

$J_j := \{i : \text{statistische Einheit } i \text{ ist in Branche } j \text{ klassifiziert}\}$

$J_j^{(erh)} := \{i \in J_j : \text{Einbetriebsunternehmen in Erhebung}\}$

$J_j^{(erg)} := \{i \in J_j : \text{statistische Einheiten in Ergänzung}\}$

Nur Unternehmen, für die im gleitenden Jahresverlauf die letzten zwölf UVA-Meldungen vollständig vorliegen, gehen in die Berechnung von Formel (1) ein. Dem branchenspezifischen mittleren Umsatzanteil des Berichtsmonats m entspricht:

$$\bar{u}_{j,m} = \frac{\sum_{i \in J_j^\star} \frac{UVA_{i,m}}{\sum_{t=0}^{11} UVA_{i,(m-t)}}}{\sum_{i \in J_j^\star} 1}, \qquad J_j^\star := J_j^{(erg)} \backslash \{i : |M_i| < 12\}. \tag{1}$$

Einheiten mit fehlender Meldung für den Berichtsmonat, jedoch mindestens sechs vorliegenden UVA-Meldungen im gleitenden Jahresverlauf, werden mittels eines Umsatzsubstituts berücksichtigt. Dem imputierten Umsatz einer Unternehmenseinheit i für den Berichtsmonat m entspricht:

$$\hat{U}_{i^\circ,m} = \frac{\sum_{(m-t) \in M_{i^\circ}} UVA_{i^\circ,(m-t)}}{\sum_{t=1}^{11} \bar{u}_{j^\circ,(m-t)} I_{i^\circ,(m-t)}} \bar{u}_{j^\circ,m}, \qquad \begin{array}{l} i^\circ \in \{i \in J_j^{(erg)} : |M_i| \geq 6,\ UVA_{i,m} \text{ fehlend}\}, \\ j^\circ \in \{j : i^\circ \in J_j^{(erg)}\}. \end{array} \tag{2}$$

$$\text{Mit } I_{i^\circ,(m-t)} = \begin{cases} 1 & \text{wenn } UVA_{i^\circ,(m-t)} \text{ vorliegt,} \\ 0 & \text{sonst.} \end{cases}$$

Für Einheiten ohne verwendbare UVA-Meldungen wird der erforderliche Monatsumsatz anhand des für das Unternehmen letzt verfügbaren Jahreswertes (Ergebnisse der Umsatzsteuererklärung, der Leistungs- und Strukturstatistik bzw. KJP)[20] und der monatlichen UVA-Branchenentwicklung imputiert (8% der ergänzten Unternehmensumsätze).

[20] Fortgeschrieben über branchenspezifische Umsatzentwicklungen im Zeitverlauf.

Nach Erstellung und Vervollständigung der Eckdaten erfolgt die modellbasierte Berechnung der restlichen beschäftigten- und umsatzbezogenen Merkmale (Löhne, Arbeitsstunden, Eigenproduktion usw.).

Auswahl der Modellbasis

Um den zu schätzenden $n^{(erg)}$ nicht erhobenen Einheiten möglichst ähnliche erhobene Einheiten aus $n^{(erh)}$ zugrunde zu legen, erfolgt eine iterative Auswahl der zur Schätzung erforderlichen **Modellbasis** innerhalb von Umsatzklassen und Wirtschaftstätigkeit - „bottom up“, beginnend mit der tiefsten Gliederungsebene.[21]

Für die Klassenauswahl der Modellbasis gilt:

$J_{j\alpha} := \{i \in J_j^{(erh)} : \ U_i < \tilde{U}_{j\alpha}, \ \alpha = min(\{\alpha' \in (1,7] : (\alpha'/10) n_j^{(erh)} \geq 20\})\}.$

U_i ... Umsatz einer statistischen Einheit i

$\tilde{U}_{j\alpha}$ ... α-Dezil der Umsatzverteilung einer Branche j der Erhebung

$n_{j\alpha}$ ... Anzahl der statistischen Einheiten einer Branche j der Modellbasis

Bei den $n_{j\alpha}$ Einheiten der Modellbasis handelt es sich um jene erhobenen Einheiten[22] einer Branche j, deren Monatsumsatz unter dem α-Dezil $\tilde{U}_{j\alpha}$ der primärstatistischen Umsatzverteilung der betreffenden Branche liegt. $\tilde{U}_{j\alpha}$ entspricht dabei dem kleinsten α-Dezil bei dem mindestens 20 Beobachtungen in der Modellbasis aufscheinen. Weisen jedoch auch die kleinsten 70% der Umsatzverteilung keine ausreichende Primärbesetzung auf, erfolgt die Auswahl auf einer übergeordneten Gliederungsebene (ÖNACE-Aggregat).[23]
Im Jänner des Berichtsjahres 2009 gelangten auf diese Weise von rund 12.000 erhobenen Unternehmen nur die kleinsten 24% iterativ in die Modellbasis, um als Strukturspender für die zu schätzenden Einheiten der 300 Wirtschaftsbereiche zu dienen. Gemessen am Eckwert Beschäftigte konnte für fast 90% der Schätzbasis im Basismonat Jänner 2009 die Parameterberechnung bereits auf Ebene der jeweiligen ÖNACE-(Unter-)Klasse (4- bzw. 5-Steller) erfolgen.

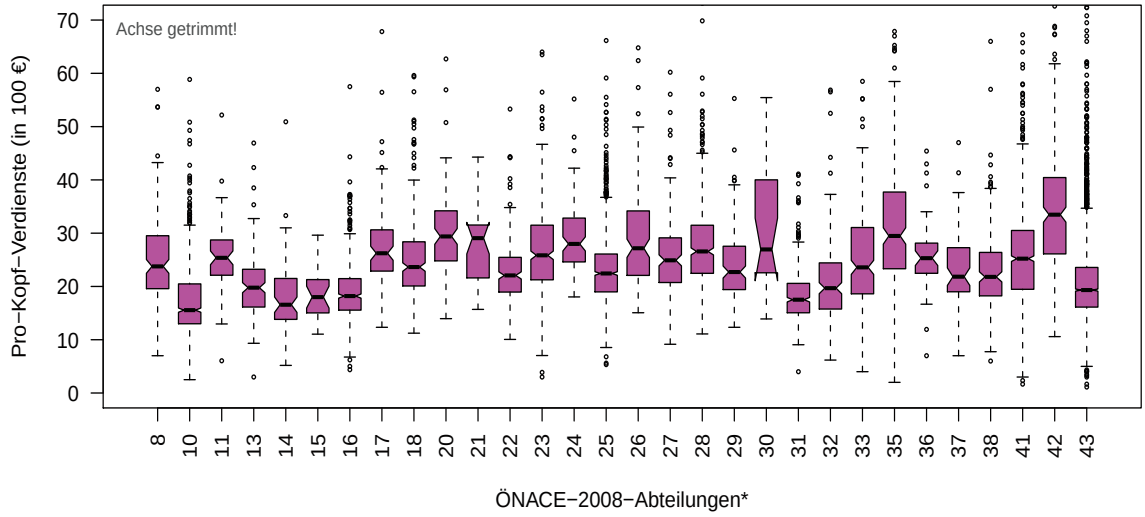
Schon am einfachen Beispiel der durchschnittlichen *Bruttoverdienste pro unselbständig Beschäftigten* in Abbildung 5 ist ersichtlich, dass sich die durch *Box-Plots* visualisierten Verteilungen der Pro-Kopf-Verdienste der ÖNACE-2008-Abteilungen deutlich unterscheiden, und die Berücksichtigung der branchenspezifischen Strukturunterschiede (also der Wirtschaftstätigkeit einer Einheit) unbedingt erforderlich ist.

[21] Entspricht bei der Wirtschaftstätigkeit den rund 300 ÖNACE-2008-Unterklassen des Produzierenden Bereichs.

[22] Ausschließlich *Einbetriebsunternehmen* dienen auf Aktivitätsebene als Strukturspender.

[23] Im Gegensatz zur Aktivitätsebene wird bei der späteren Erstellung des Güteransatzes nicht auf übergeordnete Branchen gewechselt.

Abbildung 5: Branchenspezifische Strukturunterschiede in der KJP-Erhebung
Verteilung der Pro-Kopf-Verdienste der Unternehmen der ÖNACE-2008-Abteilungen



Quelle: KJP 2009 (Berichtsmonat Jänner). - *) Ohne ÖNACE-2008-Abteilung 06, 07, 09, 12, 19, 39.

Aufgrund der relativ starken Fluktuation an der Meldeschwelle[24] vom Zeitpunkt der Stichprobenziehung bis zum Zeitpunkt der Durchführung der MDE kann innerhalb einer Branche den größeren zu schätzenden Einheiten eine vergleichbare Basisstruktur zugrunde gelegt werden. Je näher die zu schätzenden Einheiten an den Meldeschwellen liegen, umso größer ist einerseits ihr Gewicht, aber umso ähnlicher sind sie andererseits auch den Strukturspendern in der Modellbasis. Mit wachsendem Abstand zur Modellbasis verringert sich zugleich auch das Gewicht der zu schätzenden Einheiten und somit deren Einfluss auf die Merkmalssummen eines Wirtschaftsbereichs. Diesen Eigenschaften der zu schätzenden Einheiten in Kombination mit der grundsätzlich hohen primärstatistischen Merkmalsabdeckung der KJP sollten den Erwartungen nach qualitativ hochwertigen Ergebnissen pro futuro sowohl hinsichtlich zeitlicher Vergleichbarkeit als auch hinsichtlich der **Kohärenz mit anderen wirtschaftsstatistischen Projekten** (insbesondere der Leistungs- und Strukturstatistik) entsprechen - selbst wenn unterstellt wird, dass innerhalb eines Wirtschaftsbereichs **unternehmensgrößenabhängige Strukturunterschiede** vorliegen. Als problematisch muss aber eine Konstellation aus geringer Primärbesetzung und starken **branchenspezifischen Strukturunterschieden** betrachtet werden, die bei sehr detaillierter Ergebnisdarstellung auftreten kann. Hier sind aufgrund *gemischter Strukturen*[25] auch mit der vorgenommenen Modellbasisabgrenzung deutlich verzerrte Schätzer zu erwarten, und wenn zusätzlich noch eine schwache Primärabdeckung[26]

[24] Die „Fluktuation an den Meldeschwellen“ bezieht sich auf jene meldepflichtigen und nichtmeldepflichtigen Einheiten, die zum Zeitpunkt der MDE (t+90 Tage) in den jeweils anderen Bereich (unter bzw. über die Meldeschwellen) fallen.

[25] Gemischte Strukturen sind dann anzunehmen, wenn etwa aufgrund geringer Primärbesetzung die Parameterschätzung einer ÖNACE-2008- Unterklasse (5-Steller) auf dem übergeordneten 4-Steller (ÖNACE-2008-Klasse) erfolgen muss, und die anderen 5-Steller dieser ÖNACE-2008-Klasse bspw. ein deutlich höheres Lohnniveau aufweisen als jener 5-Steller für den die Parameterschätzung erfolgen soll.

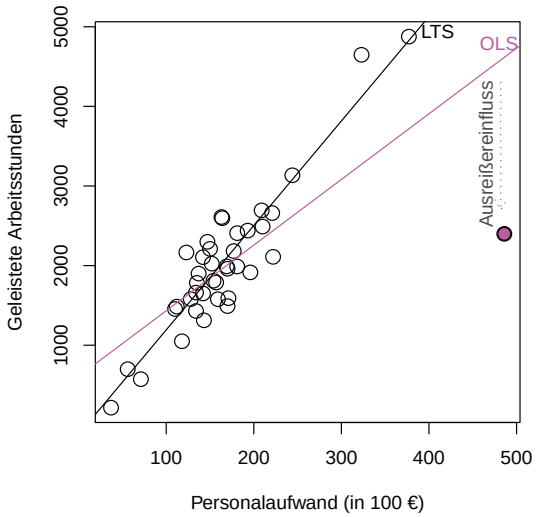
[26] Zu unterscheiden ist die Problematik einer schwachen Primärabdeckung (Anteil der MDE an Merkmalssumme ist hoch) von jener der geringen Primärbesetzung (zu wenig Strukturspender für die branchenspezifische

vorliegt, der Anteil der MDE also hoch ist, dann kann ein gewisser systematischer Effekt in den Ergebnissen kaum vermieden werden. Im Einzelfall muss deshalb zusätzliches qualifiziertes Expertenrating[27] zur Gewährleistung der erforderlichen Datenqualität beitragen.

Parameterschätzung

Wirtschaftsstatistische Zusammenhänge von Mikrodaten können häufig schon durch ein einfaches lineares Modell gut beschrieben werden. Da wirtschaftsstatistische Daten jedoch stets mit **Ausreißern**[28] behaftet sind, ist eine robuste Methode zur Gewährleistung einer stabilen Qualität der Modellanpassung unerlässlich.
Abbildung 6 veranschaulicht dies an einem einfachen Beispiel für eine der 300 ÖNACE-2008-Unterklassen der Modellbasis. Es ist ersichtlich, dass im klassischen linearen Regressionsmodell (OLS-Regression)[29] schon eine einzige Beobachtung ausreicht, um völlig unsinnige Schätzer zu erhalten - die Lösung wird quasi zum Ausreißer hingezogen. Die LTS-Regression[30] hingegen ist robust, reagiert also nicht so empfindlich auf Ausreißer und passt sich gut an die Datenmehrheit an.

Abbildung 6: Ausreißereinfluss im Regressionsmodell
(Auszug aus Modellbasis)



Quelle: KJP 2009 (Berichtsmonat Jänner).

Basierend auf der iterativ bestimmten Modellbasis werden bei der MDE deshalb alle be-

Modellbildung).

[27] Beim sogenannten „Expertenrating“ handelte es sich um Einschätzungen und Modifizierungen einzelner Werte und Parameter durch Fachexperten der Wirtschaftsstatistik aufgrund ihrer Kenntnisse branchenspezifischer Eigenheiten und Kriterien, die beim Vorliegen bekannter systematischer Abweichungsmechanismen vorgenommen werden.

[28] Als Ausreißer gelten im vorliegenden Fall jene Beobachtungen, die nicht dem (linearen) Muster der Datenmehrheit folgen.

[29] Ordinary Least Squares Regression.

[30] Least Trimmed Squares Regression.

schäftigtenbezogenen Hauptmerkmale branchenspezifisch mit Hilfe eines **robusten linearen Regressionsmodells** berechnet. Um den Einfluss additiver Effekte bei der Mikrodatenerstellung zu vermeiden, wird grundsätzlich ein lineares Modell ohne Interzept angewandt.

Für das Modell $y_i = x_{i1}\beta_{j1} + ... + x_{ip}\beta_{jp} + \epsilon_i$, mit ϵ_i als Störterm, und $i \in J_{j\alpha}$ ist die Least Trimmed Squares (LTS) Regression folgendermaßen definiert (*Rousseeuw, 1984*):[31]

$$\hat{\beta}^{LTS}_{n_{j\alpha}} = argmin_{\hat{\beta}_j} \sum_{k=1}^{h_j} (e^2)_{k:n_{j\alpha}} \quad \ldots \text{ist die Zielfunktion der LTS Optimierung.} \tag{3}$$

Die Summe der h_j kleinsten quadratischen Residuen sind dabei zu minimieren, wobei die quadrierten Residuen $e_i = (y_i - \hat{y}_i)$ in geordneter Form wie folgt definiert sind:

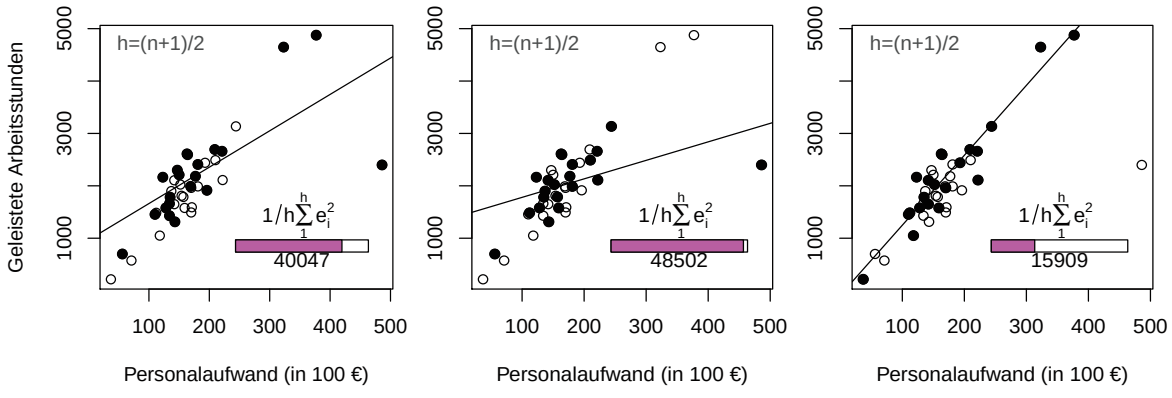
$$(e^2)_{1:n_{j\alpha}} \leq (e^2)_{2:n_{j\alpha}} \leq \ldots \leq (e^2)_{k:n_{j\alpha}} \leq \ldots \leq (e^2)_{n_{j\alpha}:n_{j\alpha}}. \tag{4}$$

Der Umfang h_j ist ein Subset aus den gegebenen $n_{j\alpha}$ Beobachtungen der Branche j, wobei h_j definiert ist als

$$h_j \in \left[\frac{n_{j\alpha}+1}{2}, n_{j\alpha}\right]. \tag{5}$$

Die unter (3) bis (5) formal beschriebene LTS-Technik kann vereinfacht als iteratives Verfahren erklärt werden wie es in Abbildung 7 demonstriert wird. Wird der Umfang des Subsets beispielsweise auf $h = (n + 1)/2$ gesetzt, so werden wiederholt zufällig 21 Beobachtungen (dunkle Punkte) aus den $n = 40$ Beobachtungen ausgewählt und basierend darauf die OLS-Regression und die Summe der h kleinsten quadratischen Residuen berechnet. Ziel ist es, jene Teilmenge h aus den n Beobachtungen zu finden, bei der die Summe der h kleinsten quadratischen Residuen minimiert wird. Dem LTS-Schätzer entspricht dann die OLS-Anpassung an diese h Punkte.

Abbildung 7: Demo der LTS-Technik - zufällig gewählte Subsets (Greedy Algorithmus)



Quelle: KJP 2009 (Berichtsmonat Jänner).

[31] *P.J. Rousseeuw (1984).* "Least median of squares regression". Journal of the American Statistical Association.

Diese Technik ist extrem rechenintensiv, da theoretisch der „ganze Raum“ abgesucht werden müsste, um das Minimum zu finden. Deshalb werden spezielle Algorithmen eingesetzt, die die Lösung (für größere Datensätze) approximieren. Die beschriebene LTS-Prozedur kann über SAS/IML (Interactive Matrix Language) aufgerufen werden, wobei standardmäßig der FAST-LTS Algorithmus[32]) implementiert ist.
Mit der LTS-Methode werden stets $n - h$ Beobachtungen ausgeschlossen, wodurch relevante Information verloren gehen kann. Mit wachsendem Subset h gewinnt der Schätzer zwar mehr Information aus den Daten, verliert aber an Robustheit und vice versa. Wird $h = 0.75n$ gewählt, so gilt dies als guter Kompromiss zwischen Robustheit und Effizienz. Es ist jedoch kaum zu erwarten, dass in der Modellbasis alle Wirtschaftsbereiche und Merkmalszusammenhänge bis zu 25% ausreißerbehaftet sind. Andererseits ist die tatsächliche Ausreißeranzahl im Vorhinein nicht bekannt, d.h. auch nicht exakt über h bestimmbar. Deshalb wird für die MDE eine **gewichtete LS-Regression** angewandt, die in einem zusätzlichen Schritt - basierend auf dem besten LTS-Subset und der zugehörigen LTS-Funktion - nur noch Beobachtungen mit großen Residuen ausschließt und basierend auf dem neu gesetzten Subset h' den finalen LTS-Schätzer berechnet (im Regelfall gilt: $h' > h$). Für die Beispieldaten in Abbildung (7) gilt $h' = 39$, d.h. alle bis auf eine der Beobachtungen werden in die finale Berechnung aufgenommen.

Mit Hilfe der berechneten branchenspezifischen Regressionsparameter erfolgt, gebunden an die Eckdaten der 46.000 zu schätzenden statistischen Einheiten, die Berechnung von Merkmalen wie *Bruttolöhne, Teilzeitbeschäftigte, geleistete Arbeitsstunden* usw.

4.3. Schätzmodell auf Güterebene

Für die Ergänzung des primärstatistischen Datenkörpers sind auch die Produktionsarten (wie *abgesetzte Produktion* oder *durchgeführte Lohnarbeit*, um zwei Beispiele zu nennen) einer statistischen Einheit zu berechnen. Zusätzlich ist die **Aufgliederung der Produktion nach PRODCOM-Positionen** (genauer: 8-Steller des nationalen Güterverzeichnisses ÖPRODCOM, da das nationale Güterverzeichnis ÖPRODCOM über die Erfordernisse der europäischen PRODCOM-Liste hinausgeht) erforderlich, um eine entsprechende Ergebnisdarstellung im Rahmen der Güterproduktion (Güteransatz) zu ermöglichen.
Die Schätzung erfolgt auch hier wiederum aktivitätsbezogen und unter Einbindung der Eckdaten aus Verwaltungsquellen. Dabei kann aufgrund des hohen statistischen Zusammenhanges das Produktionsvolumen einer statistischen Einheit über dessen Unternehmensumsatz bestimmt werden.
Darüber hinaus ist auch eine plausible ÖPRODCOM-Zuordnung über die wirtschaftliche Aktivität einer statistischen Einheit möglich. Die Matrixdarstellung nach der Systematik der Wirtschaftsaktivitäten (ÖNACE 2008) und der statistischen Güterklassifikation in Verbindung mit den Wirtschaftszweigen (ÖCPA 2008)[33] weist generell eine hohe Konzentration des

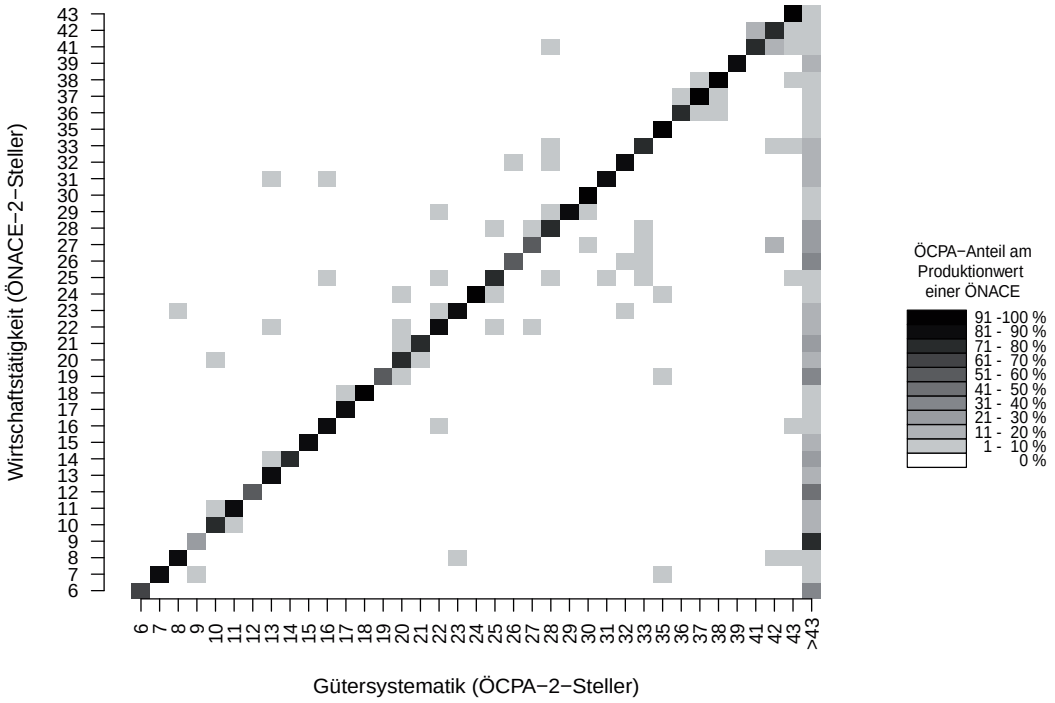
[32] *Rousseeuw, P.J. and Van Driessen, K.*: “Computing LTS Regression for Large Data Sets”, Springer, Netherlands 2006.

[33] Vgl. Anhang zur Verordnung (EG) Nr. 451/2008 des Europäischen Parlaments und des Rates vom 23. April 2008 zur Schaffung einer neuen Güterklassifikation in Verbindung mit den Wirtschaftszweigen (CPA) und zur Aufhebung der Verordnung (EWG) Nr. 3696/93 des Rates.

relevanten Produktionsvolumens an den Hauptachsen auf. Abbildung 8 veranschaulicht die Gliederung des Produktionsvolumens nach Wirtschaftsbereichen und Gütergruppen. An der Intensität der Farbe ist die Konzentration der Produktion eines Wirtschaftsbereichs in den verschiedenen Gütergruppen ersichtlich. In der ersten Zeile der Makematrix ist beispielsweise die Konzentration der *abgesetzten Produktion* für die ÖNACE-2008-Abteilung 43 (*Vorbereitende Baustellenarbeiten, Bauinstallationen und sonstiges Ausbaugewerbe*) veranschaulicht. Für die ÖNACE-2008-Abteilung 43 liegt die Konzentration der charakteristischen Produktion (ÖCPA-2008-Abteilung 43) bei rund 92%, die restlichen 8% fallen auf *Produktbegleitende Umsätze* (ÖCPA-2008-Abteilung> 43).

Abbildung 8: Makematrix nach ÖNACE-2008-Abteilungen und ÖCPA-2008-Abteilungen

Konzentration der abgesetzten Produktion in KJP-Erhebung



Quelle: KJP 2009 (Berichtsmonat Jänner).

Schätzung der Parameter zur Gütergliederung

Als Ausgangsbasis für die Aufgliederung der Produktionswerte nach Gütern dient eine Matrix, in der den ~ 300 Wirtschaftsbereichen (ÖNACE-2008-Unterklassen) plausible ÖPRODCOM-Codes zugeordnet sind. Dies kann wie zuvor auf Aktivitätsebene beschrieben, grundsätzlich durch empirisch beobachtbare Zuordnungen innerhalb einer Branche erfolgen.

Nur in einigen nicht primärstatistisch erfassten Wirtschaftsbereichen wird ein ÖPRODCOM-Set unter Einbindung der Klassifikationsdatenbank der STATISTIK AUSTRIA zugewiesen.[34] Für alle Gütercodes eines Wirtschaftsbereichs werden empirische Gewichte wie etwa *Besetzungsgewicht, mittlerer Mengenanteil, mittleres Codegewicht* usw., berechnet.

Für die $j = 1, ..., 300$ Wirtschaftsbereiche erfolgt die Berechnung der Parameter zur Gütergliederung des Produktionswertes basierend auf den jeweils $n_{j\alpha}$ erhobenen Einheiten der Modellbasis. Es gilt:

$c = 1, ..., T$ ... laufender Gütercodeindex

$w_{i,c}$ ... Produktionswert eines Gütercods c einer statistischen Einheit i

$w_i = \sum_{c=1}^{T} w_{i,c}$ ... Produktionswert einer statistischen Einheit i

$$g_{j,c} = \frac{1}{n_{j\alpha}} \sum_{i \in J_{j\alpha}} I_{i,c}, \quad \text{... Besetzungsgewicht eines Gütercodes } c \text{ einer Branche } j. \tag{6}$$

$$\text{Mit } I_{i,c} = \begin{cases} 1 & \text{wenn Einheit } i \text{ den Code } c \text{ aufweist,} \\ 0 & \text{sonst.} \end{cases}$$

Für den branchenspezifischen mittleren Anteil des Gütercodes c am Produktionsvolumen w gilt:

$$\overline{w}_{j,c} = \frac{\sum_{i \in J_{j\alpha}} \frac{w_{i,c}}{w_i}}{\sum_{i \in J_{j\alpha}} I_{i,c}} \tag{7}$$

Der in (6) und (7) beschriebenen Parameterberechnung wird wieder eine iterativ gewählte Modellbasis kleinerer primärstatistisch erhobener Einheiten zugrunde gelegt, wodurch sich auch die zu berücksichtigenden ÖPRODCOM-Codes von mehr als 3.200 unterschiedlichen Codes der Primärerhebung der KJP auf etwa 1.200 Gütercodes der Modellbasis reduzieren. Die Vorgehensweise bei der Modellbasisauswahl entspricht im Wesentlichen der bereits auf Aktivitätsebene beschriebenen Methode (vgl. dazu den Abschnitt Auswahl der Modellbasis), wobei jedoch die Berechnungen beim Güteransatz ausschließlich auf Ebene der ÖNACE-2008-Unterklasse erfolgen.[35]

Bei nicht ausreichender primärstatistischer Besetzung wird beim Güteransatz (im Gegensatz zum Aktivitätsansatz) für die Parameterberechnung nicht auf eine übergeordnete Wirtschaftsebene gewechselt, sondern es werden innerhalb der betreffenden ÖNACE-2008-Unterklasse schrittweise auch jene großen erhobenen Einheiten (große Einbetriebsunternehmen und Betriebe von Mehrbetriebsunternehmen)[36] in die Modellbasis aufgenommen, die beim Aktivitätsansatz nicht berücksichtigt werden. Die Einbindung beschränkt sich in diesen Fällen jedoch auf

[34] Automatische Übernahme jener Gütercodes der Klassifikationsdatenbank, deren Codierung auf Ebene der 4- und 5-Steller mit der betreffenden ÖNACE 2008-(Unter)Klasse übereinstimmen, und nachträgliche Überprüfung der erfolgten Zuordnung (Hinweis: der symmetrische Aufbau von CPA-2008 und ÖNACE-2008 ist in einigen wenigen Fällen nicht eindeutig und muss entsprechend berücksichtigt werden (etwa beim ÖNACE-3-Steller Bau von Gebäuden (412)).

[35] Über ein höheres ÖNACE-Aggregat kann in der Regel keine plausible ÖPRODCOM-Zuweisung erfolgen.

[36] Betriebe eines Mehrbetriebsunternehmens fließen in die Berechnung der Modellgewichte nur als eine zusammengefasste Einheit ein.

Gütercodes, die dem Kriterium des „wirtschaftlichen Ursprungs“ folgen (deren Codierung auf Ebene der 4- und 5-Steller mit der betreffenden ÖNACE-2008-(Unter)Klasse übereinstimmt). Gütercodes, die in der primärstatistischen Datenmasse nur vereinzelt auftreten, werden grundsätzlich nicht auf die zu ergänzenden Einheiten übertragen, um eine mögliche Überzeichnung dieser Codes zu vermeiden (außer es handelt sich um Hauptcodes, und für die Branche liegen keine weiteren relevanten Codes vor).

Parameterzuweisung und Gütergliederung

Mit der im Abschnitt zuvor beschriebenen Parameterschätzung erhält man für jeden Wirtschaftsbereich ein Gütercode-Set mit zugehörigen Gewichten. In Abbildung (9) ist am Beispiel der ÖNACE-2008-Unterklasse 10711 (*Herstellung von Schwarz- und Weißbackwaren*) eines dieser 300 Gütercode-Sets mit einigen Parametern veranschaulicht.

Abbildung 9: Ausgewählte Parameter zur Gütergliederung am Beispiel der ÖNACE 10711 (*Herstellung von Schwarz- und Weißbackwaren* (Auszug aus Modellbasis))

ÖPROCDOM 2008			Gewicht in Modellbasis			Anzahl der Betriebe in Ergänzung
Bezeichnung	CODE	Eh	g (in %)	HC (in %)	$\overline{w}$	
Schwarzbrot	1071110001	kg	99	14	0,16275	1.148
Weißbrot	1071110002	kg	83	6	0,11716	965
Spezialbrot	1071110003	kg	83	4	0,09771	965
Weißgebäck	1071110005	kg	91	64	0,29394	1.065
Brot, Gebäck ... für diätische Verwendung	1071110006	kg	9	1	0,09310	100
Anderes frische Brot ohne Zusatz von ...	1071110007	kg	13	1	0,09561	150
Weichwaren (ohne Dauerbackwaren), gesüßt	1071120001	kg	4	3	0,10340	466
Feingebäck (ohne Dauerbackwaren), gesüßt	1071120002	kg	89	6	0,12784	1.032
Sonstiger Einzelhandel, a. n. g.	4700002000		83	0	0,17156	965
Dienstleistungen von Restaurants, ..., Cafés	5610101000		37	0	0,14143	433
Q: KJP 2009 (Berichtsmonat Januar)						

Für die zu ergänzenden statistischen Einheiten eines Wirtschaftsbereiches wird das über den Unternehmensumsatz berechnete Produktionsvolumen $\hat{w}$ nun anhand dieser branchenspezifischen Gewichte nach ÖPRODCOM-Positionen aufgegliedert. Die ÖPRODCOM-Zuweisung erfolgt dabei über die jeweilige schwerpunktmäßige wirtschaftliche Tätigkeit einer statistischen Einheit und unter Berücksichtigung des Besetzungsgewichtes $g_{j,c}$ anhand eines probabilistischen Zuordnungsverfahrens.[37]
Von den 1.179 zu schätzenden statistischen Einheiten der in der Abbildung 9 beispielhaft angeführten ÖNACE-2008-Unterklasse 10711 erhalten also 99% den Code 1071110001, 83% den Code 1071110002 usw.

Der Anteil eines Gütercodes c am über den Unternehmensumsatz geschätzten Produktionswert $\hat{w}$ hängt letztendlich davon ab, welche Code-Kombination der statistischen Einheit per Zufallsauswahl innerhalb der Branche zugewiesen wurde. Jede statistische Einheit hat aber zumindest einen Hauptcode (HC) und $k \geq 0$ Nebencodes.

[37] Liegen jedoch aus Vorperioden (früheren Erhebungen) brauchbare historische Informationen über die Gütergliederung einer statistischen Einheit vor, so wird diese Struktur auf die Einheit übertragen.

Es gilt:

$\hat{w}_{i,c}$ ... geschätzter Produktionswert eines Codes c einer stat. Einheit i

$\hat{w}_i = \sum_{c=1}^{T} \hat{w}_{i,c}$ ... geschätzter Produktionswert einer statistischen Einheit i

$C_{j\alpha} := \{c \in C_j : g_{j,c} > 0\}$ ... Menge der Gütercodes c in Branche j der Modellbasis

C_j umfasst alle primärstatistisch auftretenden Gütercodes einer Branche j, und $C_{j\alpha}$ alle zulässigen Codes der Modellbasis. Es gilt: $\hat{C}_i \subset C_{j\alpha} \subset C_j$, mit $\hat{C}_i$ als die einer statistischen Einheit zugeordneten Gütercodemenge.

Die Funktion zur Verteilung der Gütercodes c einer Branche j auf die statistische Einheit i ist definiert als:

$$\hat{w}_{i,c} = \hat{w}_i \frac{\overline{w}_{j,c} I_{i,c}}{\sum_{c' \in C_{j\alpha}} \overline{w}_{j,c'} I_{i,c'}}, \tag{8}$$

$$\text{mit } I_{i,c} = \begin{cases} 1 & \text{wenn Code } c \text{ Einheit } i \text{ zugewiesen,} \\ 0 & \text{sonst.} \end{cases}$$

Für die in der Abbildung 9 beispielhaft angeführte ÖNACE-Unterklasse 10711 gilt etwa $|C_{j=25}| = 42$ und $|C_{j=25\ \alpha=2}| = 10$, d.h. es sind nur noch zehn aus ursprünglich 42 möglichen Gütercodes der Primärerhebung in der Modellbasis.

Von der Möglichkeit, die in der Erhebungsmasse vorkommende *Kombination von Güterkategorien* gemäß ihrer Auftrittswahrscheinlichkeit auf die Schätzbasis zu übertragen wurde abgesehen, da diese für die vorgesehenen statistischen Zwecke nicht unbedingt erforderlich ist, und ein multivariater Ansatz - aufgrund der Modellvorgaben (Beschränkungen durch Modellbasis, Restriktionen bei Gütercodeübertragung, Spenderstruktur kann nicht wie bei Hot-Deck einfach übertragen werden, usw.) - zu nicht abschätzbaren Effekten führen kann. Die Übertragung der univariaten Auftrittswahrscheinlichkeit wurde als ausreichend erachtet.

5. Unterjährige Fortschreibung

Zu Beginn eines Berichtsjahres (Berichtsmonat Jänner) erfolgt im Zuge der modellbasierten Ergänzung des primärstatistischen Datenköpers die Überprüfung einflussreicher Fälle[38] der Schätzbasis durch Experten der Fachstatistik - insbesondere hinsichtlich der übernommenen Verwaltungsdaten an die alle übrigen Merkmale gebunden sind.

Unterjährig (Berichtsmonat Februar bis Dezember) werden die mit dem Berichtsmonat Jänner erzeugten Daten in einem automatisierten Prozess fortgeschrieben.[39] Grundsätzlich wieder - wie im Abschnitt zuvor beschrieben - durch die Einbindung von Verwaltungsquellen und durch die Berücksichtigung von Branchenentwicklungen. Die Beschäftigtenstruktur einer statistischen Einheit kann dabei über Registerverknüpfungen zum HV, die Umsatzhöhe

[38] Als einflussreiche Fälle gelten bei der MDE jene Einheiten, die von der Definition der Klein- und Kleinstunternehmen abweichen oder aus bestimmten Kriterien auffällig sind (bspw. durch einen hohen Anteil an der Merkmalssumme eines Gütercodes oder durch eine starke Umsatzänderung im Vergleich zur Vorjahresstatistik).

[39] Für unterjährig neu auftretende Einheiten wird ein Vormonats-Dummydatensatz angelegt (mittels branchenspezifischer Spenderstruktur aus Unternehmen der Vormonatsergänzung), inaktive Einheiten werden gelöscht.

durch UVA-Meldungen der Finanzbehörde monatlich aktualisiert werden. Zur Berücksichtigung arbeitstägiger Konstellationen und saisonaler Muster werden zudem über die unterjährige Entwicklung der Primärdaten branchenspezifische (mediane) Änderungsraten berechnet. Die Vorgehensweise entspricht im Wesentlichen der LOCF-Methode[40], welche in der KJP für Meldeausfälle mit historischen Informationen entwickelt wurde. Da es sich bei den ergänzten Daten aber nicht um zufällige Meldeausfälle, sondern um eine systematisch abgeschnittene Masse handelt, muss den zu schätzenden Branchengewichten eine andere Basis als den Meldeausfällen zugrunde gelegt werden (vgl. dazu den Abschnitt Auswahl der Modellbasis). Zudem wird im Gegensatz zur LOCF-Subsitution die monatliche Umsatzänderung nicht basierend auf der Primärerhebung der KJP gewonnen, sondern kann, wie zuvor beschrieben, über die unterjährige UVA-Entwicklung der zu ergänzenden Einheiten selbst berücksichtigt werden.[41]

Es gilt:

$v = 1, ..., r$ ... Index der statistischen Merkmale
x ... Wert eines statistischen Merkmals mit Index $v^\bullet$, mit $v^\bullet \in V^\bullet$
y ... Wert eines statistischen Merkmals mit Index v°, mit $v^\circ \in V^\circ$
$V^\bullet = \{v : \text{aus Verwaltungsquellen verfügbar }\}$
$V^\circ = \{v : \text{nicht aus Verwaltungsquellen verfügbar }\}$

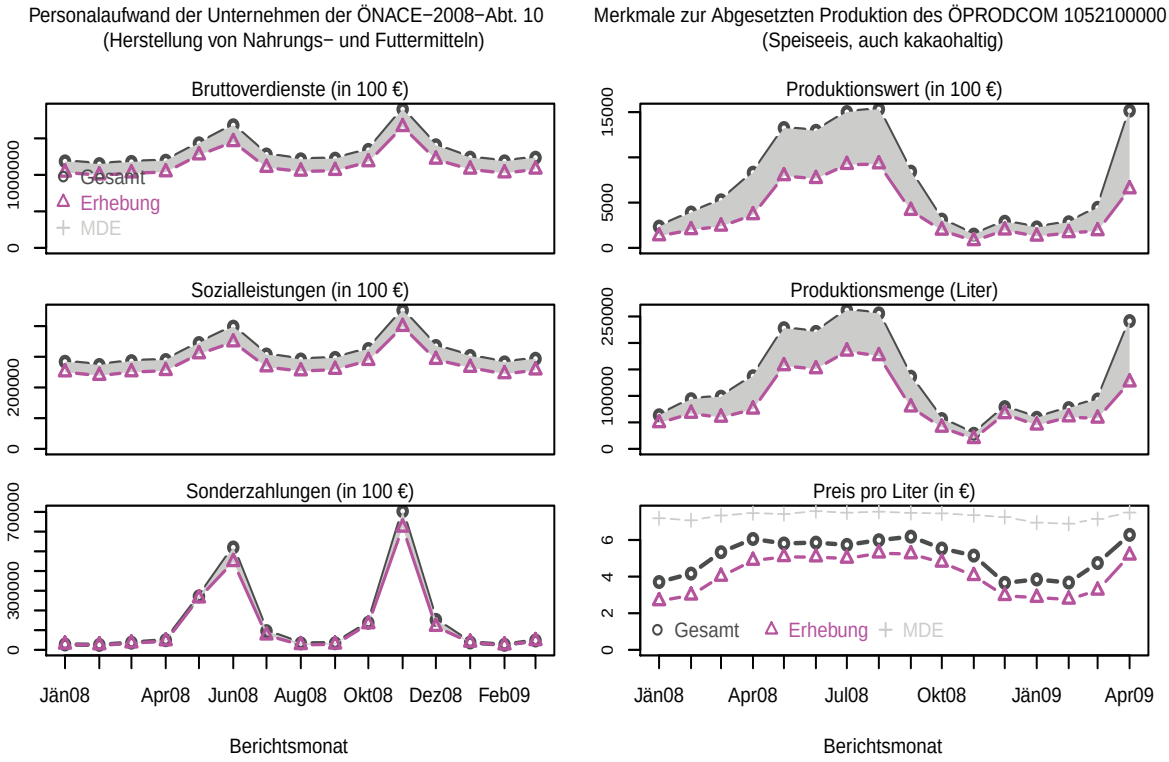
$$\hat{y}_{i^*,m} = \left[\hat{y}_{i^*,(m-1)} \left(\frac{x_{i^*,m}}{x_{i^*,(m-1)}}\right)\right]^{I_x^*} \left[\left(\underset{i \in J_{j\alpha}^\circ}{median} \left(\frac{y_{i,m}}{x_{i,m}}\right)\right) x_{i^*,m}\right]^{1-I_x^*} \left[\underset{i \in J_{j\alpha}^\circ}{median} \left(\frac{\frac{y_{i,m}}{x_{i,m}}}{\frac{y_{i,(m-1)}}{x_{i,(m-1)}}}\right)\right]^{I_x I_x^*} \tag{9}$$

Mit $i^* \in \{i : i \in J_j^{(erg)}\}$ und $J_{j\alpha}^\circ := J_{j\alpha} \backslash \{i : x_{i,(m-t)} = 0,\ t \in \{0,1\}\}$
$I_x = 1$ wenn Quelle für x_y der HV, $I_x = 0$ sonst.
$I_x^* = 1$ wenn $x_{i,(m-1)}^y > 0$, $I_x^* = 0$ sonst.

In Abbildung 10 ist am Beispiel ausgewählter beschäftigtenbezogener Merkmale der ÖNACE-2008-Abteilung 10 (*Herstellung von Nahrungs- und Futtermitteln*) die Entwicklung der ergänzten Datenmasse auf Aktivitätsebene ersichtlich. Hervorzuheben wäre etwa der Aspekt, dass Sonderzahlungen im Modell Großteils zu den Zeitpunkten Juni und November Berücksichtigung finden (Urlaubsgeld, Weihnachtsrenumeration).
Abbildung 10 veranschaulicht zudem an der unterjährigen Entwicklung der ÖPRODCOM-Position 1052100000 (*Speiseeis, auch kakaohaltig*) die Datenergänzung auf Güterebene von Jänner 2008 bis April 2009. Hier ist nicht nur bei Produktionswert und -menge ein sehr ausgeprägter saisonaler Effekt ersichtlich, sondern auch beim Literpreis. Ursache der Preisschwankungen ist, dass die industrielle Eiserzeugung ganzjährig mit sehr niedrigen Literpreisen erfolgt, und die Vielzahl an Konditoren ihr Speiseeis vorwiegend saisonal produzieren - zu deutlich höheren Preisen. Dadurch steigt der aus verschiedenen Wirtschaftsbereichen „zusammengesetzte“ Literpreis des betreffenden Gütercodes saisonal stark an, obwohl die Preise innerhalb der jeweiligen Wirtschaftsbereiche relativ konstant bleiben.

[40] Last Observation Carried Forward
[41] Einige Merkmale (z.B. Sonderzahlungen) können nicht auf diese Weise fortgeschrieben werden, hier erfolgt die monatliche Berechnung über branchenspezifische mittlere Anteilswerte.

Abbildung 10: Beispiel für Fortschreibung auf Aktivitätsebene und Güterebene



Quelle: KJP (01/2008 bis 04/2009).

6. Datenqualität

Die im vorliegenden Beitrag beschriebene Methode der KJP-Datenergänzung illustriert die Vielzahl an Faktoren, welche bei der Beurteilung der Datenqualität zu berücksichtigen sind. Die Beachtung verschiedenster Qualitätsfaktoren ist zum einen in den konzeptionellen Vorgaben begründet, die keine konventionelle Hochrechnung zulassen, und zum anderen auf die Komplexität des zu erstellenden Datenkörpers selbst zurückzuführen. Die Darstellung der KJP-Grundgesamtheit erfordert neben geeigneter mathematischer Verfahren auch das Einbeziehen von Informationen aus verschiedenen primär- und sekundärstatistischen Datenquellen. Neben der Schlüsselrolle, die dabei dem Unternehmensregister der STATISTIK AUSTRIA zukommt, sind es vor allem Verfügbarkeit und konzeptionelle Eignung der externen Verwaltungsquellen, die in die **Gesamtqualität der Daten** einfließen. Umfangreiche Voranalysen und Testberechnungen zeigten, dass die sekundärstatistisch ergänzten Eckdaten grundsätzlich als „voll erhoben“ betrachtet werden können und die Güte der sonstigen beschäftigten- und umsatzbezogenen Merkmale davon abhängt, inwieweit ausreichende primärstatistische Informationen zur Modellbildung vorliegen. Überall dort, wo das eigentliche Ziel einer MDE umgesetzt werden kann - nämlich die modellbasierte Ergänzung eines vorwiegend primärstatistisch erfassten Datenkörpers - kann grundsätzlich von hoher Datenqualität ausgegangen werden. Diese Annahme ist einerseits darin begründet, dass für die Datenergänzung ausreichend primärstatistische Informationen zur Modellentwicklung vorliegen, und andererseits, dass selbst in Wirtschaftsbereichen mit deutlicher Strukturabweichung mögliche Modelleffek-

te mit wachsender Primärabdeckung gegen Null tendieren.
Während die MDE auf Aktivitätsebene für die Mehrzahl von Hauptaggregaten auch ohne ressourcenintensive Arbeitsschritte zufriedenstellende Ergebnisse erbringt, stellt sich die Ergänzung im Rahmen der Güterproduktion ungleich schwieriger schwieriger dar. Beim Güteransatz sind der methodischen Vorgehensweise auch „natürliche Grenzen“ gesetzt. Güter, die von Betrieben in der Spendermasse nicht hergestellt werden, können auch nicht imputiert werden. Es fehlen somit auch bei einer MDE grundsätzlich jene Güter, die ausschließlich von kleinen Einheiten unterhalb der Cut-Off-Grenze produziert werden. Darüber hinaus zeigten in primärstatistisch schwach besetzten, nicht homogenen Wirtschafts-(teil-)bereichen und in Bereichen mit deutlicher Strukturabweichung die im Vorfeld durchgeführten Analysen, dass die Ergänzung nicht ausschließlich anhand automationsunterstützter Verfahren erfolgen kann. Trotz aller Informationen, die bereits in dieses System der modellbasierten Datenergänzung auf Mikrodatenebene eingehen und automatisch verarbeitet werden, ist in Einzelfällen qualifiziertes Expertenrating unerlässlich.

7. Resümee und Ausblick

In Österreich führte die politische Intention, Klein- und Kleinstunternehmen in höchstmöglicher Weise von der Auskunftspflicht auszunehmen, zur Situation, dass in einigen Wirtschaftsbereichen weder das Repräsentanzkriterium der nationalen Durchführungsverordnung, noch jenes der PRODCOM-Verordnung ausreichend umgesetzt werden konnte. Um ein Maximum an Repräsentanz und damit einhergehend, ein Optimum an Effizienz und Qualität des Datenkörpers zu gewährleisten, können jedoch gemäß den europäischen Bestimmungen Mitgliedstaaten die erforderlichen Daten nach dem Grundsatz der verwaltungstechnischen Vereinfachung durch eine Kombination verschiedener Quellen erstellen, sofern die übermittelten Daten die Struktur der Grundgesamtheit der statistischen Einheiten widerspiegeln. Dies ist mit der Entwicklung des Konzepts der modellbasierten Datenergänzung zur Darstellung der Grundgesamtheit der KJP seit dem Referenzjahr 2009 sowohl nach dem Aktivitäts- als auch nach dem Güteransatz in vollem Umfang gewährleistet.
Beginnend mit den endgültigen Ergebnissen des Berichtsjahres 2008 werden zusätzlich zum primärstatistischen Datenkörper auch die Ergebnisse der KJP Grundgesamtheit veröffentlicht. Weitere Entwicklungen, wie die Berücksichtigung weiterer sekundärstatistischer Quellen zur Verstärkung der Eckdaten und der an sie gebundenen Modellrechnung werden im Sinne einer ständigen Qualitätsverbesserung angestrebt. Derzeit besteht die Möglichkeit, im Zuge des **MEETS-Project 2010**[42] eine künftige Einbindung von Lohnzetteldaten zu prüfen. Weiters sind methodische Arbeiten geplant, um zu prüfen, ob es durch den Einsatz statistisch mathematischer Methoden möglich ist, dem Nutzer zur Abschätzung der Datengenauigkeit geeignete Kennzahlen zur Verfügung zu stellen.

[42] *Feasibility study of implementing wage tax data in structural business statistics.*MEETS (Modernisation of European Enterprise and Trade Statistics) projects on modernisation of Business Statistics - Use of administrative data.

Stand: Dezember, 2009

Projekt-Team

- Mag. Johann Hameseder (Direktion Unternehmen / Stv. Direktor)
- ADir. RR. Leopold Milota (Direktion Unternehmen / Projektleiter KJP)
- Mag. Martin Haitzmann (Direktion Unternehmen / Analyse)
- Dr. Mag. Martin Hirsch (Direktion Unternehmen / Analyse)
- Mag. Markus Fröhlich (Methodik / LOCF-Substitution)
- Mag. Gerlinde Dinges (Methodik / Modellbasierte Datenergänzung)

Statistik Austria, Guglgasse 13, A-1110 Wien.

Dieser Methodenbericht ist online abrufbar unter

http://www.statistik.at/web_de/downloads/methodik/kjp.pdf