Neuchâtel, Mai 2001

Schlussbericht Betriebszählung 1995 (BZ 95)

Mehrfach imputierte Umsatzzahlen

Jürg Hüsler und Samuel Müller
Institut für Mathematische Statistik und Versicherungslehre
Universität Bern

Auskunft:

Beat Hulliger

Bundesamt für Statistik
Dienst Statistische Methoden
Espace de l'Europe 10, CH-2010 Neuchâtel
Tel.: (+41 32) 71 36015
Fax: (+41 32) 71 36093
Bestellnummer: 338-0002
© BFS/OFS

Die Betriebszählung 1995 des Bundesamts für Statistik sollte unter anderem Angaben über den Umsatz der Unternehmen liefern. Da viele Umsatzangaben fehlten und die restlichen Angaben Ausreisser enthielten, wurde von einer Veröffentlichung entsprechender Ergebnisse abgesehen. Die Sektion Unternehmen und Beschäftigung des BFS hat darauf eine Studie in Auftrag gegeben, die eine Einsetzung (Imputation) der fehlenden Umsatzangaben untersuchen sollte. Professor Jürg Hüsler und Samuel Müller vom Institut für Mathematische Statistik und Versicherungslehre haben die Studie erarbeitet und den vorliegenden Schlussbericht am 27. April 1999 abgeliefert. Auf der Seite des BFS haben Dr. Iris Krummenacher von der Sektion Unternehmen und Beschäftigung und Dr. Beat Hulliger vom Dienst Statistische Methoden die Studie begleitet. Der Methodenbericht wurde von Jann Potterat, Dienst Statistische Methoden, herausgegeben.
Die Studie zeigt eine grosse Spannweite des totalen Umsatzes je nach verwendeter Methode. Offensichtlich sind nur sogenannte robuste Methoden in der Lage, plausible Schätzungen zu liefern.

INHALTSVERZEICHNIS

1 PROBLEMSTELLUNG 1

2 EINFACHE METHODEN 1

2.1 METHODE A1: HOCHRECHNEN DES MITTELWERTES PRO BERECHNUNGS-KLASSE 1
2.2 METHODE A2: ROBUSTES HOCHRECHNEN DES MITTELWERTES PRO BERECHNUNGSKLASSE 1
2.3 METHODE B1: HOCHRECHNEN DER SUMME PRO NOGA 2 KLASSE 2
2.4 METHODE B2: HOCHRECHNEN DER ROBUSTEN SUMME PRO NOGA 2 KLASSE 2
2.5 BEMERKUNGEN 2

3 MEHRFACHE ROBUSTE IMPUTATION 3

3.1 METHODE C1: MEHRFACHE ROBUSTE IMPUTATION, NICHT ROBUSTES HOCHRECHNEN DER UMSÄTZE 3
3.2 METHODE C2: MEHRFACHE ROBUSTE IMPUTATION, ROBUSTES HOCHRECHNEN DER UMSÄTZE 5

4 METHODISCHE BEMERKUNGEN 5

5 DATENPROBLEME 6

5.1 FAKTOR 10 6
5.2 FAKTOR 1000 7
5.3 PROBLEMBEHANDLUNG 8

6 FLUSSDIAGRAMM MEHRFACHE IMPUTATION 9

7 RESULTATE 9

8 FOLGERUNGEN 10

9 SCHLUSSFOLGERUNGEN 10

10 UMSATZZAHLEN 11

11 APPENDIX 14

11.1 QUANTILE UND METHODEN 14
11.2 VARIABLENNAMEN 16
11.3 UNT 17
11.4 FORMELN UND PROGRAMM 18
11.5 SCATTERPLOTS UMVZEQ / VZEQ 21
11.6 BOXPLOTS UMVZEQ 26
11.7 SCATTERPLOT DER RELATIVEN VERHÄLTNISSE 34
11.8 SCATTERPLOT METHODENVERGLEICH 35

1 Problemstellung

In der Betriebszählung 1995 (BZ 95) wurde unter anderem die Frage nach dem Umsatz des Unternehmens gestellt. Von den 277'331 Unternehmen beantworteten 57‘425 diese Frage nicht. Dies entspricht einem Anteil von 20.7%. Aufgrund der erhobenen Umsatzzahlen und weiteren Informationen zu den einzelnen Unternehmen wie Beschäftigungsstruktur, Standort, Wirtschaftsklasse, soll eine möglichst gute Schätzung des Gesamtumsatzes pro NOGA 2 Klasse gerechnet werden. Wegen den fehlenden Umsatzzahlen sind jedoch spezielle statistische Methoden zu verwenden, damit eine gute Schätzung bestimmt werden kann.

Um eine unverfälschte Schätzung des Gesamtumsatzes zu erhalten, ist es notwendig, dass jedem Unternehmen, welches den Umsatz nicht angegeben hat, ein sogenannt theoretischer Umsatz zugeordnet wird. Zur Berechnung dieser theoretischen Umsätze existieren verschiedene statistische Methoden. Wir haben in dieser Studie mit 6 verschiedenen Methoden diese theoretischen Umsätze bestimmt und entsprechende Hochrechnungen der Umsätze durchgeführt.

2 Einfache Methoden

2.1 **Methode A1:** Hochrechnen des Mittelwertes pro Berechnungsklasse

Pro Berechnungsklasse (siehe Tabelle 4) wird in dieser einfachen Methode der Mittelwert des Umsatzes pro Vollzeitäquivalenz (UMVZEQ) derjenigen Unternehmen berechnet, welche den Umsatz angegeben hatten. Anschliessend wird jedem Unternehmen, welches den Umsatz nicht angegeben hatte, sein theoretischer Umsatz zugeordnet, der sich aus der Multiplikation dieses entsprechenden Mittelwerts mit der Anzahl Vollzeitäquivalenz (VZEQ) des Unternehmens ergibt. Anschliessend wird pro NOGA 2 Klasse der Gesamtumsatz durch Summieren aller Umsätze (sowohl der angegebenen als auch der theoretischen) berechnet (Formeln siehe Appendix 11.4).

2.2 **Methode A2:** Robustes Hochrechnen des Mittelwertes pro Berechnungsklasse

Diese zweite einfache Methode funktioniert ähnlich wie Methode A1, nur werden pro Berechnungsklasse die Mittelwerte robuster bestimmt. Dabei werden extrem hohe Umsätze als fehlende Umsätze behandelt, da einerseits einige Unternehmen ihren Umsatz in Franken angaben und somit einen 1000 Mal zu grossen Umsatz haben (siehe auch Abschnitt 5); andererseits ist auch nicht auszuschliessen, dass gewisse Unternehmen aus verschiedenen Gründen ihre Umsatzzahlen viel zu hoch angaben. Diese Unternehmen haben UMVZEQ, welche mit grosser Wahrscheinlichkeit oberhalb des 99%-Quantils des UMVZEQ liegen. Pro Berechnungsklasse wurde deshalb der Mittelwert des UMVZEQ ohne das extreme obere Prozent bestimmt.

Danach wurden wieder wie in der Methode A1 den Unternehmen mit fehlendem Umsatz diese Mittelwerte multipliziert mit der Anzahl VZEQ als theoretischer Umsatz zugeordnet. Auch Unternehmen, deren UMVZEQ im extremen oberen Prozent lagen, wurden durch diese theoretischen Umsatzwerte ersetzt. Schlussendlich berechneten wir den Gesamtumsatz pro NOGA 2 Klasse wiederum durch Summieren aller Umsätze, sowohl der angegebenen, der theoretischen wie der ersetzten Umsätze (Formeln siehe Appendix 11.4).

2.3 **Methode B1**: Hochrechnen der Summe pro NOGA 2 Klasse

Pro NOGA 2 Klasse wird bei dieser Methode zuerst der Gesamtumsatz von jenen Unternehmen berechnet, welche den Umsatz angegeben hatten. Danach wird dieser Gesamtumsatz multipliziert mit der Summe der VZEQ aller Unternehmen und dividiert durch die Summe der VZEQ jener Unternehmen, welche den Umsatz angegeben hatten (Formeln siehe Appendix 11.4).

2.4 **Methode B2**: Hochrechnen der robusten Summe pro NOGA 2 Klasse

Diese Methode ist die robuste Variante von Methode B1, nur wird pro NOGA 2 Klasse der Gesamtumsatz jener Unternehmen, welche ihren Umsatz angaben, robust geschätzt. Bei Unternehmen mit UMVZEQ oberhalb des 99%-Quantils des UMVZEQ wurde der Umsatz ersetzt durch VZEQ multipliziert mit dem 99%-Quantil der UMVZEQ (Formeln siehe Appendix 11.4).

2.5 Bemerkungen

Der Vorteil dieser einfachen Methoden liegt darin, dass ohne grossen Aufwand und in Anbetracht der grossen Anzahl Unternehmen in relativ kurzer Zeit eine Schätzung für den Gesamtumsatz pro NOGA 2 Klasse erhalten wird. Diese Methoden weisen aber auch einige nicht vernachlässigbare Nachteile auf. Die wichtigsten sind:

i) Die Unsicherheit der Methode kann nicht angeben werden.
ii) Mögliche systematische Unterschiede zwischen Unternehmen, welche den Umsatz angegeben haben, und jenen, die ihn nicht angegeben haben, werden nicht berücksichtigt.
iii) Das Problem extremer Werte (Falschangaben) wird unbefriedigend gelöst.
iv) Hinter diesen Methoden verstecken sich Annahmen über die Verteilung der Umsätze in Abhängigkeit der VZEQ. Genaugenommen sollte man pro Berechnungsklasse die Voraussetzungen überprüfen und anschliessend die geeignete parametrische Methode anwenden. Dies wäre jedoch nicht immer die gleiche Methode, das heisst also, dass bei manchen NOGA 2 Klassen Methode A verwendet werden sollte, bei anderen jedoch Methode B.

Deshalb schlagen wir in den folgenden Abschnitten Methoden vor, die diese gewichtigen Nachteile nicht aufweisen. Die folgenden Methoden werden als mehrfache, robuste Imputation bezeichnet.

3 Mehrfache robuste Imputation

3.1 **Methode C1:** Mehrfache robuste Imputation, nicht robustes Hochrechnen der Umsätze

Das Problem von fehlenden Werten ist in der Statistik schon seit den Anfängen bekannt. Vor einigen Jahren entstand die Methode der mehrfachen Imputation von fehlenden Werten. Diese basiert hauptsächlich auf der Theorie der sogenannten Markoff Ketten Monte Carlo (MCMC). Dank schnellen Computern ist es heute möglich für moderate Umfänge (bis ca. 1000) diese Methode anzuwenden. Bei der BZ95 ist das Modellieren der Einflussgrössen durch Multinomialverteilungen für die diskreten Variablen und durch Multinormalverteilungen für die stetigen Variabeln, sowie das anschliessende Anwenden des EM-Algorithmus aufgrund der „leider" immer noch beschränkten Computerleistungen nicht möglich. Deshalb wird im folgenden eine vereinfachte Version, das sogenannte „Hot Deck"-Verfahren, verwendet.
Dieses „Hot Deck"-Verfahren ist für Umfänge wie in der BZ95 anwendbar. Grundgedanke dieses Verfahrens ist, dass man einem fehlenden Umsatz einen beobachteten Umsatz zuordnet. Man setzt aber nicht irgendeinen Umsatz ein, sondern versucht durch ein Modell für ein Unternehmen mit einem fehlenden Umsatz mehrere Unternehmen zu finden, welche ähnliche Strukturen aufweisen, um dann von diesen vergleichbaren Umsätzen zufällig einen Umsatz auszuwählen und diesen als imputierten Umsatz zu verwenden. Ähnliche Struktur wird in dem Sinne gemeint, dass die anderen erhobenen Variablen wie Betriebsgrösse, Wirtschaftsklasse, Export, Lage, usw., möglichst mit denjenigen des Unternehmens, dessen Umsatz fehlt, übereinstimmen.
Die Unterteilung der Unternehmen in kleinere und grössere Unternehmen soll die Strukturabhängigkeit der fehlenden Umsätze aufzeigen.

Betriebsklassen nach Anzahl Beschäftige	Anzahl Betriebe	Anzahl Betriebe mit fehlendem Umsatz	Anzahl Betriebe mit fehlendem Umsatz in %
1	78027	20906	26.8
2	51757	11227	21.7
3-4	55606	11006	19.8
5-9	50783	9088	17.9
10-19	22993	3518	15.3
20-49	11857	1245	10.5
50-99	3520	264	7.5
100-199	1600	104	6.5
200-249	296	21	7.1
250-499	537	22	4.1
500-999	214	12	5.6
1000+	141	12	8.5

Wie lassen sich aber nun die Unternehmen in verschiedene Klassen mit ähnlicher Struktur einteilen? Wie gelingt es, dass in den verschiedenen Klassen genügend Unternehmen vorhanden sind, welche ihren Umsatz angegeben hatten?

Zur Einteilung in passende Klassen stehen mehrere Variablen zur Verfügung, die gemeinsam im Sinne der mehrdimensionalen Klassierung verwendet werden sollen. Mit einer geeigneten Regression lässt sich das mehrdimensionale Klassieren auf eine Dimension vereinfachen. Dies ist insbesondere sehr hilfreich, wenn nominale, ordinale und stetige Variabeln zur Klassierung beitragen, was bei der BZ95 der Fall ist.

Mit Hilfe der logistischen Regression gelingt es, jedem Betrieb die Wahrscheinlichkeit zuzuordnen, mit der dieses Unternehmen den Umsatz angegeben hat. Als Einflussgrössen verwenden wir die verschiedenen anderen Fragestellungen der BZ 95 (siehe Appendix 11.2). Die Zielgrösse ist die dichotome Variable UMSATZ, welche den Wert 1 hat, falls der Umsatz angegeben wurde, und den Wert 0 hat, falls der Umsatz nicht angegeben wurde. Nach der Berechnung der logistischen Regression pro Berechnungsklasse hat man also für jedes Unternehmen einen Wert zwischen Null und Eins. Dabei gilt natürlich, dass wenn zwei Unternehmen gleiche erklärende Variabeln haben, dann stimmen auch die geschätzten Wahrscheinlichkeiten für das Fehlen des Umsatzes überein. Man sucht sich somit zur Imputation zu einem Unternehmen ohne Umsatz die Unternehmen mit etwa gleich grosser Wahrscheinlichkeit der Umsatzangabe. Dadurch wird das mehrdimensionale Problem auf eine eindimensionale Klassierung reduziert.

Das Intervall [0,1] der Wahrscheinlichkeiten wird nun in verschiedene Klassen unterteilt. Die Anzahl der Klassen wählen wir als die aufgerundete Wurzel aus der Anzahl der angegebenen Umsätze. Danach werden wegen der Imputation die Klassenbreiten so gewählt, dass in jedem Teilintervall etwa gleich viele Unternehmen vorhanden sind, welche ihren Umsatz angegeben haben. Somit wird also jedem Unternehmen einer Wahrscheinlichkeitsklasse zugeordnet. Zuletzt wird jedem Unternehmen, das den Umsatz nicht angegeben hat, aus seiner Klasse heraus zufällig ein möglicher Umsatz pro Vollzeitäquivalenz zugeordnet.

Da wir verhindern möchten, dass Ausreisser als Umsatzwerte zugeordnet werden, kommen Umsätze aus dem oberen extremen Prozent der Umsatzwerte pro Vollzeitäquivalenz als möglicher Wert (Donor) nicht in Frage. Dieses Verfahren wurde in jeder Berechnungsklasse verwendet. Eine detaillierte Übersicht der verwendeten Anzahl Klassen pro Berechnungsklasse findet sich im Appendix 11.1.

Spielt man diese Prozedur einmal durch, so erhält man eine einzelne Imputation. Um eine Schätzung der Unsicherheit dieser Methode zu erhalten, wird dieses Imputationsverfahren fünf Mal durchgeführt. Man erhält eine sogenannte mehrfache Imputation. Aus diesen imputierten und den angegebenen Umsatzwerten lassen sich die Umsatzzahlen jeder NOGA 2 Klasse bestimmen und daraus der Umsatz aller Unternehmen. Da fünf Datensätze durch die fünf Imputationen existieren, werden diese Umsatzzahlen fünf Mal (U1 bis U5) bestimmt, woraus der Einfluss der Imputation und die damit verbundene Unsicherheit abgeschätzt werden kann. Der definitive mehrfache Imputationsschätzer für den Umsatz ist einfach der Mittelwert

der fünf einfachen Imputationen. Für diese Schätzung lässt sich aus den fünf einfachen Schätzungen auch die Standardabweichung berechnen.

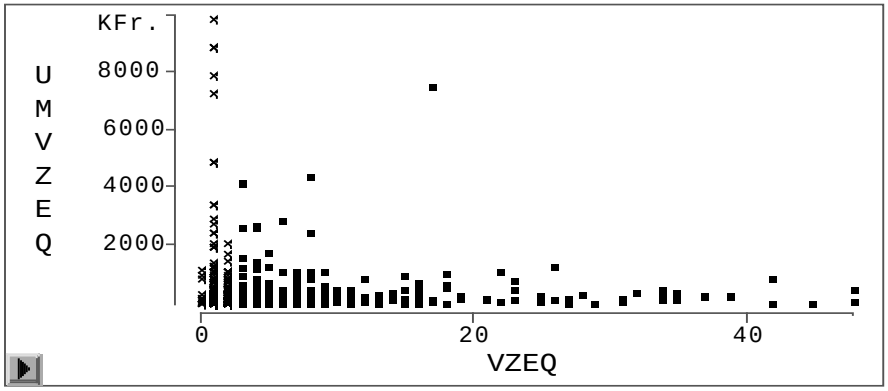
3.2 **Methode C2:** Mehrfache robuste Imputation, robustes Hochrechnen der Umsätze

Wichtig in der Beurteilung der Umsatzzahlen ist die Frage der Verwendung der extremsten Umsatzwerte pro Vollzeitäquivalenz. Um eine Mehrfache Robuste Imputation zu erhalten, entschieden wir uns, die Umsätze pro Vollzeitäquivalenz aus dem extremen oberen Prozent durch das 99%-Quantil zu ersetzen. Dadurch entdeckt man grosse Unterschiede in der Schätzung der Umsatzzahlen. Je nach Anzahl fehlender Umsatzwerte kann dieser Effekt ein Mehrfaches der robusten Schätzung ausmachen. Diese Werte findet man in der Tabelle 3. Der Vergleich der Kolonnen UR1 bis UR5 (der fünf robusten Imputationen, robust hochgerechnet) mit den Kolonnen U1 bis U5 (der fünf robusten Imputationen, nicht robust hochgerechnet) zeigt die Unterschiede sehr deutlich. Zur Wahl des 99%-Quantils siehe auch Abschnitt 5.

4 Methodische Bemerkungen

Die Berechnungen wurden separat pro NOGA 2 Klasse durchgeführt. Um homogenere Klassen zu erhalten, wurde wenn möglich noch weiter unterteilt in Unternehmen mit Vollzeitäquivalenz 1 oder 2 und in Unternehmen mit VZEQ grösser als 2 (siehe Beispiel). Für die 1848 Unternehmen mit VZEQ 0 wurde eine separate Berechnung über alle NOGA 2 Klassen vorgenommen. Alle Unternehmen, welche eine 0 beim Umsatz angaben, wurden als Missing Values betrachtet.
Diese Methode hat apriori keine Voraussetzung an die Verteilung der UMVZEQ. Einziges Ziel ist das erzeugen von möglichst homogenen Klassen in Bezug auf die verschiedenen Einflussgrössen. Dahinter steht die Vorstellung, dass Unternehmen mit ähnlichen Einflussgrössen auch ähnliche Umsätze haben, das heisst also, dass in den durch diese Methode erzeugten Klassen die UMVZEQ identisch und unabhängig verteilt sind.

Beispiel: Scatterplot VZEQ gegen UMVZEQ (in 1000 Fr) für die NOGA 2 Klasse 67 (Beobachtungen mit UMVZEQ>10 Mio. Fr sind nicht berücksichtigt, ebenso wie Unternehmen mit VZEQ > 60).



Unternehmen mit VZEQ < 3 sind mit einem Kreuz dargestellt, jene mit VZEQ grösser gleich 3 mit einem Kästchen. Aus dieser Grafik ist ersichtlich, dass in dieser NOGA 2 Klasse die Streuung von UMVZEQ abhängig von VZEQ ist. Weitere Beispiele und Grafiken siehe Appendix 11.5.

Die oben beschriebene logistische Regression wurde mit so vielen Einflussgrössen wie möglich berechnet. Bei NOGA 2 Klassen mit weniger als 3000 Unternehmen konnten nicht alle der 27 Einflussgrössen (Beschreibung der Einflussgrössen im Appendix 11.2) berücksichtigt werden. Deshalb ist in diesen Fällen die Anzahl der Einflussgrössen abhängig von der Anzahl Unternehmen in der Berechnungsklasse. Bei ganz kleinen NOGA 2 Klassen (weniger als 20 Unternehmen), war es gar nicht sinnvoll eine logistische Regression durchzuführen. Bei diesen kleinen Klassen wurde den Unternehmen mit fehlendem Umsatz zufällig ein beobachteter Umsatz (umgerechnet aus dem Umsatz pro Vollzeitäquivalenz) zugeordnet. In den meisten Fällen der logistischen Regression konnten alle Einflussgrössen verwendet werden, auch wenn nicht alle einen wesentlichen Einfluss ausüben. Die zeitliche Belastung wäre wesentlich grösser gewesen, hier das logistische Modell passend zu vereinfachen. Die Vereinfachung des Modelles hätte auch keine wesentliche Veränderung des Imputationsverfahrens zur Folge und auch die Resultate würden nicht wesentlich beeinflusst. Im Appendix 11.1 (Tabelle 4) ist das genaue Vorgehen pro Berechnungsklasse angegeben.

5 Datenprobleme

Neben dem Fehlen der Umsatzwerte bestehen noch andere Probleme mit den Daten der Betriebszählung, welche einen grossen Einfluss auf die Umsatzberechnungen ausüben.

5.1 Faktor 10

Beim Einlesen der Umsatzzahlen mit dem Scanner kam es vor, dass zusätzlich eine 1 rechts von der Zahl gelesen wurde (z. B. 1001 an Stelle von 100). Bei 6446 (2.3%) Unternehmen findet sich eine solche 1. Ein Teil dieser 1 ist durch fehlerhaftes Einlesen entstanden, was durch ein passendes Verfahren korrigiert werden sollte. Wie häufig dieser Fehler auftritt, ist unbekannt. Dieses Problem wurde wie folgt gelöst:

Jedem Unternehmen wurde eine geschätzte Wahrscheinlichkeit p(Umsatz) zugeordnet, mit der die letzte Stelle seines angegebenen Umsatzes gestrichen werden sollte. Danach wurde die letzte Stelle (eine 1) des Umsatzes von den 6446 für dieses Problem in Frage kommende Unternehmen mit der entsprechend geschätzten Wahrscheinlichkeit p(Umsatz) zufällig gestrichen.

Die Wahrscheinlichkeiten p(Umsatz) berechneten wir unter folgender Überlegung. Wir nehmen an, dass die Wahrscheinlichkeit, dass ein Unternehmen seinen Umsatz mit einer 1 an der letzten Stelle um Tausend Franken abrundet gleich der Wahrscheinlichkeit ist, dass ein Unternehmen seinen Umsatz mit einer 9 an der

letzten Stelle um Tausend Franken aufrundet. Die Wahrscheinlichkeit für einen Umsatz von *i* KFr. wurde folgendermassen geschätzt:

$$p(i) = \frac{a(i) - a(i+8)}{a(i)}, \qquad \text{falls } a(i) - a(i+8) > 0 \text{ und } (i \text{ Modulo } 10) = 1,$$

$$p(i) = 0, \qquad \text{falls } a(i) - a(i+8) \leq 0 \text{ oder } (i \text{ Modulo } 10) \neq 1,$$

wobei $a(i)$ gleich der Anzahl der Unternehmen mit Umsatz gleich KFr i ist.

Beispiel: Wahrscheinlichkeiten für Umsatzangaben zwischen 1000 und 1010 KFr.

i = 1000, also (i Modulo 10) = 0
i = 1001, also (i Modulo 10) = 1
...
i = 1009, also (i Modulo 10) = 9
i = 1010, also (i Modulo 10) = 0

Also gilt, dass p(1000) = p(1002) = p(1003) = ... = p(1010) = 0
Für i = 1001 benötigen wir die Anzahl Unternehmen, welche diesen Umsatz angegeben haben, sowie die Anzahl Unternehmen welche Umsatz 1009 angaben:

a(1001) = 88 und a(1009) = 20

Somit gilt: p(1001) = (88-20)/88 = 0.77.

Anmerkung

Da wir fünf einfache Imputationen durchführten, wiederholten wir das zufällige Streichen der 1 auch fünf Mal und erzeugten somit fünf verschiedene Datensätze, um den Einfluss dieser Korrektur zu beurteilen. Mit diesen Datensätzen führten wir dann die oben beschriebene logistische Regression durch.

Dieses Problem ist nicht so schwerwiegend wie das nächste, das wesentlich grösseren Einfluss auf die Umsatzschätzung hat.

5.2 Faktor 1000

Die Umsatzzahlen sollten in 1000 Franken angegeben werden. Dies wurde allerdings nicht von jedem Unternehmen beachtet und es kam vor, dass Unternehmen ihren Umsatz in Franken angegeben haben. Dies hat zur Folge, dass bei Nichtbeachten dieser Tatsache, der Gesamtumsatz völlig überschätzt wird. Zur Illustration und Abschätzung des Einflusses dieses Datenproblems dient folgende einfache Überlegung:
530 (0.2%) Unternehmen gaben mehr als 20 Millionen Umsatz pro Vollzeitäquivalenz an. Nehmen wir an, dass alle diese Unternehmen ihren Umsatz irrtümlicherweise in Franken notiert haben, dann würden wir diese Umsatzzahlen entsprechend um den Faktor 1000 korrigieren. Dadurch würde der Gesamtumsatz nur ein Drittel des Gesamtumsatzes ohne Korrektur betragen.

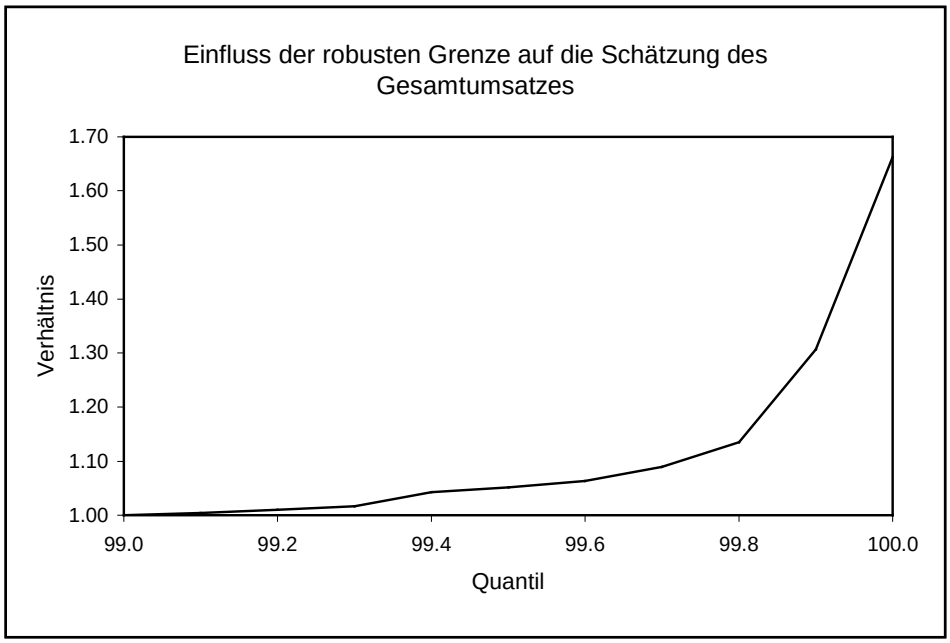
Anmerkung:
Dieser Effekt ist auch im Vergleich der beiden einfachen Methoden A1 und A2 ersichtlich.

5.3 Problembehandlung

Es wäre aber eine zu einfache und höchstwahrscheinlich auch falsche Lösung gewesen, alle extremen angegebenen Umsätze um den Faktor 1000 nach unten zu korrigieren. Deshalb wurde dieses Problem „gelöst“, indem wir die extremen 1 Prozent der Umsatzwerte pro Vollzeitäquivalenz auf das 99%-Quantil setzten. Dies ist natürlich eine willkürliche Wahl. Anstelle des 99%-Quantils liesse sich auch ein anderes Quantil wählen.

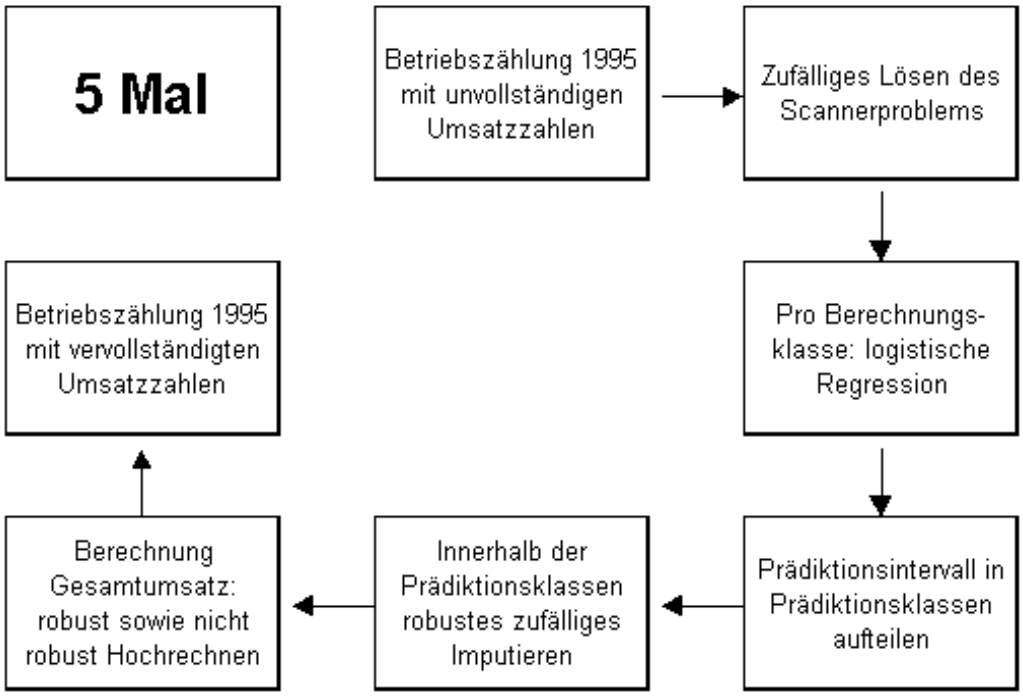
Zur Beurteilung des 99%-Quantils wurden die extremen Werte der UMVZEQ pro NOGA 2 Klasse mit Extremwertverteilungen modelliert und analysiert, um einen passenden Hinweis auf die zu verwendende Grenze zu erhalten. Es zeigte sich jedoch, dass nur in wenigen NOGA 2 Klassen ein Grenze ersichtlich wurde. Um ein ziemlich einheitliches und einfach verständliches Vorgehen in der Imputation anzuwenden, wurde deshalb das 99%-Quantil gewählt.

Anstelle des 99%-Quantils wurden auch andere Quantile verwendet, um den Effekt dieser Korrektur zu erkennen. Zwischen den Umsatzschätzungen mit den 99%- und 98.5%-Quantilen besteht nur ein minimer Unterschied. Etwas grösser (ca. 5%) ist der Unterschied für den Gesamtumsatz, wenn das 99.5%-Quantil verwendet wird, da das 99.5%-Quantil je nach Berechnungsklasse bis zu einem Faktor 43 grösser ist als das 99%-Quantil (siehe Appendix 11.1, Tabelle 4, NOGA 2 Klasse 37). Zur Veranschaulichung dient folgende Grafik, welche den Einfluss der Wahl der robusten Grenze auf den Gesamtumsatz zeigt. Berechnet wurde der Einfluss auf die erste Imputationsschätzung U1 in Intervallschritten von 0.1. Der Wert an der Stelle 99.0 entspricht UR1 und entsprechend steht der Wert an der Stelle 100.0 für U1. Die Berechnungspunkte wurden linear verbunden.



6 Flussdiagramm Mehrfache Imputation

Zum besseren Verständnis des gewählten Vorgehens soll das folgende Flussdiagramm dienen. Es zeigt die verschiedenen Auswertungsschritte auf dem Weg zu einem Imputationsschätzer (mit robuster sowie nicht robuster Hochrechnung) des Gesamtumsatzes pro NOGA 2 Klasse.



7 Resultate

Die Resultate der mehrfachen robusten Imputation ohne robuste Hochrechnung (Methode C1), der robusten multiplen Imputation mit robuster Hochrechnung (Methode C2), sowie der einfachen Methoden sind in der Tabelle 2 zusammengefasst. In der Tabelle 3 finden sich die Werte der 5 robusten Imputationen, sowohl der robust hochgerechneten Imputationen (UR1 bis UR5) wie auch der nicht robust hochgerechneten Imputationen (U1 bis U5).

8 Folgerungen

Aufgrund der Resultate der Tabellen stellen wir folgendes fest:

i) Die fünf verschiedenen, zufälligen, einzelnen Imputationen zeigen nur kleine Unterschiede in den Umsatzschätzungen, sie weisen eine kleine Variabilität auf. Die einzelnen Schätzungen des gesamten Umsatzes variieren um ± 10 Mia. Fr. Der Mittelwert der fünf einzelnen Imputationen kann gut verwendet werden.

ii) Die Methodik, das heisst hier die Behandlung der extremen Werte übt einen grossen Einfluss aus. Die Hochrechnung des Umsatzes liefert in der weniger robusten Behandlung der extremen Werte fast den doppelten Wert (von 0.6 Bio. auf 1 Bio. Fr.).

iii) Aus i) folgt, dass die Imputationsmethode in der Behandlung der fehlenden Werte eine mögliche und theoretisch richtige Schätzung für den Gesamtumsatz liefert, welche eine erstaunlich kleine Variabilität hat. Aus ii) ziehen wir den Schluss, dass die Frage nach der Behandlung der extremen Werte entscheidend für die Güte der Schätzung ist. Diese Problematik kann nicht allein mit Hilfe statistischer Methoden gelöst werden.

iv) Auf der Stufe der einzelnen NOGA 2 Klassen zeigen sich zum Teil erheblich grössere Unterschiede zwischen den verschiedenen Methoden zur Bestimmung des Umsatzes (siehe Abbildungen in Appendix 11.7 und 11.8).

9 Schlussfolgerungen

1) **Je nach gewählter Methode resultieren recht unterschiedliche Werte für die Summe der Umsatzwerte aller Unternehmen in der Betriebszählung 1995.**

2) **Werden nur die statistisch sinnvolleren, robusten Imputationsmethoden verwendet, liegt der totale Umsatz etwa zwischen 620 Mia. Fr. bis 1'040 Mia. Fr., je nach Behandlung der extremsten Werte.**

3) **Werden andere, zu einfache Methoden benützt, wird dieser Bereich des Gesamtumsatz auf 550 Mia. Fr. bis 1'450 Mia. Fr. vergrössert.**

10 Umsatzzahlen

Tabelle 1: Wichtige Grössen NOGA 2 Klassen

NOGA 2 Klasse	Anzahl Betriebe	Anzahl Missing	Anzahl Donors	Missing in %	Summe Donors VZEQ	Summe Miss ing VZEQ	Summe VZEQ	Missing VZEQ in %
10	4	1	3	25.0	28	7	35	20.0
11	9	1	8	11.1	87	14	101	13.9
13	3	0	3	0.0	6	0	6	0.0
14	349	47	302	13.5	4'746	654	5'400	12.1
15	2'843	499	2'344	17.6	55'793	3'485	59'299	5.9
17	924	129	795	14.0	18'599	992	19'597	5.1
18	1'123	239	884	21.3	9'597	801	10'412	7.7
19	350	77	273	22.0	3'574	451	4'030	11.2
20	6'741	1'350	5'391	20.0	35'157	6'054	41'226	14.7
21	262	17	245	6.5	14'974	723	15'698	4.6
22	4'596	711	3'885	15.5	49'349	3'869	53'250	7.3
23	19	3	16	15.8	761	7	768	0.9
24	804	106	698	13.2	59'066	6'696	65'768	10.2
25	822	103	719	12.5	20'865	1'282	22'152	5.8
26	1'395	207	1'188	14.8	19'222	1'590	20'817	7.6
27	305	36	269	11.8	14'896	1'133	16'029	7.1
28	7'391	1'249	6'142	16.9	77'684	7'615	85'322	8.9
29	3'693	582	3'111	15.8	106'517	4'867	111'399	4.4
30	171	27	144	15.8	3'741	285	4'026	7.1
31	1'224	158	1'066	12.9	45'663	1'705	47'370	3.6
32	667	88	579	13.2	14'857	1'940	16'800	11.5
33	3'088	480	2'608	15.5	56'603	5'163	61'782	8.4
34	222	46	176	20.7	4'070	398	4'468	8.9
35	357	62	295	17.4	5'812	277	6'090	4.5
36	4'024	678	3'346	16.8	26'903	2'483	29'416	8.4
37	299	72	227	24.1	1'733	370	2'105	17.6
40	262	28	234	10.7	17'906	1'013	18'921	5.4
41	85	15	70	17.6	694	34	734	4.7
45	32'374	6'536	25'838	20.2	280'819	36'078	316'973	11.4
50	13'747	3'101	10'646	22.6	63'386	11'397	74'826	15.2
51	19'158	3'338	15'820	17.4	154'848	15'209	170'242	8.9
52	40'863	7'645	33'218	18.7	253'348	32'088	285'695	11.2
55	23'233	5'542	17'691	23.9	151'726	29'894	181'650	16.5
60	6'998	1'772	5'226	25.3	46'094	39'800	85'936	46.3
61	111	23	88	20.7	2'215	387	2'602	14.9
62	146	44	102	30.1	16'965	702	17'667	4.0
63	2'439	498	1'941	20.4	28'964	4'253	33'233	12.8
64	312	87	225	27.9	3'769	72'397	76'168	95.1
67	1'188	404	784	34.0	4'472	1'316	5'803	22.7
70	3'296	1'001	2'295	30.4	10'784	2'719	13'587	20.1
71	725	134	591	18.5	2'740	292	3'039	9.6
72	5'454	966	4'488	17.7	24'825	3'225	28'105	11.5
73	266	61	205	22.9	2'735	636	3'374	18.9
74	45'981	8'808	37'173	19.2	172'780	28'006	201'190	13.9
80	3'293	915	2'378	27.8	10'169	2'410	12'623	19.2
85	16'818	4'394	12'424	26.1	48'496	13'242	61'865	21.4
90	516	144	372	27.9	3'610	968	4'584	21.1
92	4'900	1'325	3'575	27.0	14'287	4'314	18'658	23.2
93	13'481	3'676	9'805	27.3	27'459	7'960	35'602	22.5
Summe	**277'331**	**57'425**	**219'906**	**20.7**	**1'993'394**	**361'201**	**2'356'443**	**-**

Tabelle 2: Umsatzzahlen (in Mio. Franken)

NOGA 2 Klasse	Umsatz ohne Modell	Umsatz mit Methode A1	Umsatz mit Methode B1	Umsatz mit Methode C1	Umsatz mit Methode A2	Umsatz mit Methode B2	Umsatz mit Methode C2
10	5	8	6	7	8	6	7
11	18	16	21	20	16	21	20
13	5	5	5	5	5	5	5
14	3'209	4'594	3'525	3'308	1'844	1'689	1'670
15	25'918	36'991	26'583	26'257	24'703	25'074	24'832
17	4'118	7'494	4'209	4'166	3'026	3'538	3'527
18	2'389	3'936	2'515	2'454	1'171	1'527	1'536
19	826	1'097	900	874	507	678	674
20	20'109	26'593	22'825	20'907	6'062	6'868	6'995
21	3'898	6'085	3'934	3'896	4'107	3'906	3'870
22	13'564	26'208	13'392	13'285	9'861	11'378	11'409
23	231	257	233	234	257	233	234
24	28'254	48'073	30'844	29'248	19'646	25'492	24'391
25	4'598	7'762	4'767	4'804	4'777	4'570	4'618
26	5'468	8'887	5'421	5'384	5'985	5'031	5'022
27	3'740	4'004	3'894	3'827	3'453	3'812	3'751
28	16'221	23'909	16'836	16'915	13'255	14'875	15'114
29	44'097	53'029	45'415	44'792	25'148	27'529	27'655
30	1'156	1'794	1'228	1'241	1'195	1'220	1'233
31	13'271	24'202	13'488	13'313	8'893	12'439	12'300
32	9'586	14'176	10'605	9'897	3'634	3'603	3'620
33	17'024	22'685	17'959	17'624	12'044	14'175	14'132
34	1'019	915	1'053	1'056	841	1'040	1'045
35	2'085	11'252	2'177	2'157	1'107	1'200	1'223
36	11'491	10'599	12'138	11'595	4'373	5'924	5'410
37	3'953	4'245	4'618	3'985	475	618	583
40	16'059	17'747	16'242	16'312	15'367	16'170	16'244
41	208	453	210	214	392	204	207
45	99'376	139'774	107'962	103'153	47'887	49'528	50'693
50	48'753	47'709	52'195	49'507	23'706	33'827	33'567
51	149'496	215'558	151'430	149'168	102'575	122'736	122'829
52	104'824	157'666	110'788	108'861	64'297	79'225	80'487
55	85'255	152'910	98'207	88'219	19'795	20'443	21'450
60	22'626	37'286	32'596	29'173	13'437	12'056	15'137
61	721	1'246	828	936	1'120	822	931
62	16'123	109'354	16'733	16'695	10'885	7'061	7'393
63	13'786	26'053	15'042	15'094	16'490	13'197	13'458
64	672	23'096	1'236	13'822	17'060	1'214	13'811
67	2'879	4'538	3'468	3'219	1'778	2'105	2'109
70	7'062	15'952	8'091	7'496	4'636	3'777	3'905
71	1'179	1'701	1'172	1'134	804	857	848
72	7'943	13'535	8'527	8'284	4'954	6'251	6'246
73	1'047	2'833	1'147	1'072	618	890	856
74	150'749	99'033	169'543	155'055	31'298	39'906	41'289
80	2'983	6'161	3'478	3'275	1'431	1'451	1'573
85	12'575	19'117	14'864	14'343	9'953	9'102	9'595
90	1'084	1'455	1'289	1'437	1'185	1'263	1'416
92	5'039	10'935	6'085	5'670	3'238	3'033	3'190
93	7'573	11'557	9'011	8'298	3'071	3'048	3'429
Summe	**994'267**	**1'464'485**	**1'078'735**	**1'041'689**	**552'374**	**604'617**	**625'539**

Tabelle 3: Umsatzzahlen bei den Methoden C2 und C1 (in Mio. Franken)

NOGA 2 Klasse	Anzahl Betriebe	UR1	UR2	UR3	UR4	UR5	U1	U2	U3	U4	U5
10	4	7	7	7	7	7	7	7	7	7	7
11	9	19	22	19	19	21	19	22	19	19	21
13	3	5	5	5	5	5	5	5	5	5	5
14	349	1'653	1'654	1'689	1'665	1'691	3'291	3'292	3'327	3'303	3'329
15	2'843	24'756	24'976	24'899	24'707	24'820	26'181	26'401	26'325	26'132	26'245
17	924	3'517	3'544	3'562	3'532	3'482	4'155	4'183	4'200	4'171	4'120
18	1'123	1'535	1'533	1'522	1'554	1'537	2'453	2'451	2'440	2'471	2'455
19	350	692	648	666	685	677	892	848	866	885	877
20	6'741	6'987	7'038	6'937	7'007	7'004	20'899	20'952	20'851	20'920	20'914
21	262	3'882	3'907	3'816	3'874	3'869	3'909	3'934	3'842	3'901	3'896
22	4'596	11'352	11'455	11'452	11'375	11'413	13'230	13'330	13'327	13'249	13'289
23	19	232	233	233	234	235	232	233	233	234	235
24	804	24'406	24'488	24'248	24'852	23'959	29'264	29'345	29'105	29'709	28'816
25	822	4'571	4'577	4'616	4'683	4'643	4'757	4'763	4'802	4'869	4'830
26	1'395	4'990	5'027	5'035	5'101	4'960	5'351	5'389	5'397	5'462	5'321
27	305	3'736	3'836	3'716	3'687	3'780	3'812	3'912	3'792	3'764	3'856
28	7'391	15'026	15'112	15'142	15'171	15'121	16'827	16'915	16'939	16'969	16'924
29	3'693	27'723	27'545	27'596	27'578	27'832	44'860	44'682	44'733	44'716	44'969
30	171	1'207	1'280	1'202	1'278	1'199	1'215	1'288	1'210	1'286	1'207
31	1'224	12'309	12'377	12'254	12'319	12'240	13'322	13'389	13'267	13'332	13'253
32	667	3'668	3'709	3'493	3'702	3'528	9'946	9'987	9'770	9'979	9'805
33	3'088	14'102	14'209	14'270	14'014	14'063	17'595	17'702	17'763	17'507	17'555
34	222	1'070	1'017	1'078	1'002	1'057	1'082	1'028	1'089	1'013	1'068
35	357	1'249	1'222	1'212	1'221	1'210	2'183	2'156	2'146	2'155	2'144
36	4'024	5'403	5'376	5'410	5'414	5'450	11'586	11'559	11'592	11'605	11'630
37	299	598	581	581	583	574	4'000	3'983	3'983	3'985	3'976
40	262	16'230	15'988	16'371	16'385	16'243	16'299	16'056	16'439	16'454	16'311
41	85	211	208	204	208	205	217	214	210	215	212
45	32'374	50'619	50'676	50'829	50'804	50'535	103'080	103'140	103'293	103'261	102'989
50	13'747	33'744	33'678	33'391	33'354	33'668	49'683	49'617	49'331	49'293	49'609
51	19'158	122'861	122'015	123'105	122'368	123'798	149'199	148'353	149'443	148'706	150'136
52	40'863	80'480	80'506	80'403	80'952	80'096	108'853	108'881	108'777	109'326	108'469
55	23'233	21'593	21'573	21'490	21'350	21'246	88'367	88'341	88'261	88'124	88'004
60	6'998	12'311	14'980	14'090	14'362	19'943	26'347	29'016	28'126	28'398	33'979
61	111	972	932	827	840	1'084	977	937	832	845	1'089
62	146	7'434	7'430	7'281	7'347	7'474	16'736	16'732	16'582	16'649	16'776
63	2'439	12'752	14'075	13'570	13'896	12'998	14'388	15'710	15'206	15'531	14'633
64	312	16'007	12'560	5'212	13'359	21'917	16'018	12'571	5'223	13'370	21'929
67	1'188	2'109	2'139	2'079	2'137	2'077	3'220	3'250	3'190	3'248	3'188
70	3'296	3'943	3'933	3'968	3'882	3'801	7'534	7'524	7'559	7'472	7'392
71	725	850	849	850	848	840	1'137	1'136	1'137	1'135	1'127
72	5'454	6'256	6'316	6'215	6'245	6'197	8'297	8'355	8'252	8'284	8'234
73	266	882	848	861	855	834	1'098	1'065	1'077	1'071	1'050
74	45'981	41'244	41'175	41'511	41'320	41'197	155'012	154'938	155'273	155'091	154'961
80	3'293	1'585	1'581	1'520	1'608	1'573	3'286	3'282	3'221	3'311	3'276
85	16'818	9'521	9'632	9'525	9'644	9'653	14'265	14'379	14'270	14'399	14'403
90	516	1'449	1'405	1'399	1'380	1'446	1'470	1'427	1'420	1'401	1'467
92	4'900	3'288	3'140	3'143	3'234	3'143	5'765	5'624	5'628	5'713	5'620
93	13'481	3'495	3'361	3'513	3'395	3'379	8'364	8'230	8'383	8'265	8'248
Summe	**277'331**	**624'533**	**624'376**	**616'015**	**625'043**	**637'727**	**1'040'687**	**1'040'531**	**1'032'164**	**1'041'212**	**1'053'853**

11 Appendix

11.1 Quantile und Methoden

Jede Zeile entspricht einer Berechnungsklasse, wobei in den ersten zwei Spalten zu sehen ist, welche Betriebe in die entsprechende Berechnungsklasse fallen.

Tabelle 4: Berechnungsklassen und Quantile UMVZEQ

NOGA 2 Klasse	VZEQ	98.5 % Quantil	99 % Quantil	99.5 % Quantil	100 % Quantil	Methode	Anzahl Klassen	Einfluss-grössen	Anzahl Betriebe	Missing in %
10	>0	-	-	-	347	Random	1	keine	4	25
11	>0	-	-	-	252	Random	1	keine	9	11
13	>0	-	-	-	4'339	Random	1	keine	3	0
14	>0	2'467	2'498	6'856	145'454	log Reg	17	b	349	13
15	1 und 2	2'200	2'726	5'847	250'000	log Reg	28	c	1'086	24
15	>2	1'531	1'790	4'833	83'333	log Reg	38	d	1'736	13
17	>0	696	950	1'500	72'500	log Reg	27	e	918	14
18	>0	1'125	1'500	3'800	136'666	log Reg	29	e	1'109	21
19	>0	708	1'700	5'000	33'333	log Reg	16	e	345	22
20	1 und 2	950	1'300	4'000	500'500	log Reg	47	c	3'038	25
20	>2	501	833	1'347	400'000	log Reg	55	d	3'688	16
21	>0	2'900	3'800	6'400	25'000	log Reg	15	b	261	7
22	1 und 2	2'000	2'150	7'580	500'000	log Reg	40	c	2'074	20
22	>2	956	1'389	1'965	155'000	log Reg	46	d	2'490	12
23	>0	-	-	-	1'132	Random	1	keine	19	16
24	>0	1'661	2'667	7'615	249'235	log Reg	26	e	798	13
25	>0	959	1'100	1'429	85'000	log Reg	26	e	817	12
26	>0	2'500	3'350	7'000	37'600	log Reg	34	b	1'390	15
27	>0	1'627	1'727	3'463	3'604	log Reg	16	e	305	12
28	1 und 2	705	800	10'000	105'000	log Reg	48	c	3'080	22
28	>2	667	776	1'565	91'333	log Reg	60	d	4'288	13
29	1 und 2	1'300	1'800	3'998	132'000	log Reg	32	c	1'373	21
29	>2	1'059	1'350	2'339	160'438	log Reg	44	d	2'305	12
30	>0	1'800	2'500	15'000	15'000	log Reg	11	b	171	16
31	>0	900	1'211	3'472	148'504	log Reg	32	e	1'222	13
32	>0	1'506	2'143	3'580	248'000	log Reg	23	e	664	13
33	1 und 2	2'000	2'500	6'467	39'416	log Reg	33	c	1'381	19
33	>2	1'157	1'375	1'808	126'762	log Reg	38	d	1'691	13
34	>0	730	786	1'500	1'800	log Reg	13	b	222	21
35	>0	1'250	1'716	36'000	450'000	log Reg	17	e	356	17
36	1 und 2	1'250	2'000	7'000	73'000	log Reg	45	c	2'561	19
36	>2	875	1'036	2'167	97'406	log Reg	35	d	1'433	13
37	>0	1'543	1'924	83'333	320'702	log Reg	14	e	297	24
40	>0	4'239	6'090	6'418	19'364	log Reg	15	b	260	10
41	>0	3'571	3'571	6'340	6'340	log Reg	8	b	79	18
45	1 und 2	886	1'230	2'970	269'800	log Reg	97	c	12'925	26
45	>2	650	1'000	1'800	500'000	log Reg	126	d	19'373	17
50	1 und 2	2'500	2'636	4'790	500'000	log Reg	69	c	6'637	26
50	>2	1'690	2'177	3'700	475'000	log Reg	75	d	7'067	19

Tabelle 4: Fortsetzung

NOGA 2 Klasse	VZEQ	98.5 % Quantil	99% Quantil	99.5% Quantil	100% Quantil	Methode	Anzahl Klassen	Einfluss-grössen	Anzahl Betriebe	Missing in %
51	1 und 2	5'900	8'000	15'000	3'030'000	log Reg	86	c	9'827	22
51	>2	5'333	6'667	9'750	1'002'300	log Reg	89	d	9'146	12
52	1 und 2	1'800	3'200	6'500	400'000	log Reg	134	c	23'437	22
52	>2	1'050	1'333	2'200	400'000	log Reg	120	d	17'167	15
55	1 und 2	525	751	2'000	242'000	log Reg	66	c	6'360	30
55	>2	500	730	1'150	10'227'143	log Reg	114	d	16'843	22
60	1 und 2	1'801	1'860	5'000	120'000	log Reg	54	c	4'227	29
60	>2	996	1'219	2'582	223'432	log Reg	46	d	2'729	19
61	>0	3'750	4'333	4'333	4'693	log Reg	9	b	111	21
62	>0	4'990	11'289	11'289	558'456	log Reg	10	b	146	30
63	1 und 2	3'701	6'000	8'000	47'000	log Reg	29	c	1'163	26
63	>2	2'220	3'002	5'350	333'333	log Reg	32	d	1'260	16
64	>0	2'000	2'403	5'000	7'500	log Reg	14	e	310	28
67	>0	3'500	7'400	10'100	150'000	log Reg	27	e	1'173	34
70	1 und 2	7'500	8'000	14'000	530'409	log Reg	37	c	2'139	33
70	>2	1'667	2'071	3'250	200'000	log Reg	28	d	1'073	24
71	>0	2'565	7'300	15'000	64'333	log Reg	24	e	718	19
72	1 und 2	783	934	2'000	1'000'000	log Reg	54	c	3'740	19
72	>2	922	1'167	1'600	37'271	log Reg	37	d	1'659	14
73	>0	1'582	2'810	40'000	92'325	log Reg	14	e	263	22
74	1 und 2	1'200	2'000	5'000	1'023'487	log Reg	150	c	29'096	21
74	>2	1'200	1'493	3'000	842'696	log Reg	117	d	16'481	16
80	1 und 2	1'000	1'001	3'000	284'000	log Reg	42	c	2'605	29
80	>2	987	1'167	2'164	120'000	log Reg	22	d	644	23
85	1 und 2	600	651	1'000	215'000	log Reg	82	c	9'469	27
85	>2	367	424	1'009	265'333	log Reg	73	d	7'222	24
90	>0	1'589	1'923	3'870	12'000	log Reg	19	e	510	28
92	1 und 2	2'500	5'000	8'000	750'000	log Reg	49	c	3'426	28
92	>2	890	1'000	1'836	25'000	log Reg	32	d	1'417	24
93	1 und 2	1'111	2'000	6'000	314'135	log Reg	81	c	9'576	29
93	>2	400	480	1'000	216'333	log Reg	53	d	3'722	22
alle	0	2'870	6'046	9'000	36'375	log Reg	37	a	1'848	24

Einflussgrössen a:
agglo, oefh, urefo, esoh, efbv, ever, impo, expo, noga2, verkfl, buaz, edat, kanton

Einflussgrössen b:
edat, agglo, vzkl, vzeq, expo, impo, noga3

Einflussgrössen c:
kanton, edat, buaz, bklt, verkfl, noga3, expo, impo, ever, efbv, esoh, urefo, oefh, vzeq, agglo, bif, be1, be2

Einflussgrössen d:
kanton, edat, buaz, bklt, verkfl, noga3, expo, impo, ever, efbv, esoh, urefo, oefh, vzeq, blm, bkn, bkt, bkv, bka, bks, agglo, bif, be1, be2, vzkl

Einflussgrössen e:
edat, verkfl, ever, agglo, vzkl, vzeq, expo, impo, noga3, bif, be1, be2, evbf, esoh

11.2 Variablennamen

Die folgenden Variabeln wurden als Einflussgrössen bei der logistischen Regression verwendet:

kanton	Kantonszugehörigkeit, nominale Variable
edat	Eröffnungsperiode des Unternehmens, ordinale Variable
buaz	Wöchentliche Arbeitszeit, stetige Variable
bklt	Anzahl Lehrlinge (Männer+Frauen) Total, stetige Variable
verkfl	Verkaufsfläche Total, stetige Variable
noga3	NOGA 3 Klassen, nominale Variable
expo	Exporttätigkeit des Unternehmens, nominale Variable
impo	Importtätigkeit des Unternehmens, nominale Variable
ever	Einnahmen Umsatz aus Verkauf (Ja/Nein), nominale Variable
efbv	Einnahmen Umsatz aus Beiträgen/Vermögen (Ja/Nein), nominal Variable
esoh	Einnahmen Umsatz aus Subventionen/öffentlicher Hand (Ja/Nein), nominale Variable.
urefo	Rechtsform des Unternehmens, nominale Variable
oefh	Öffentliche Hand am Unternehmen Beteiligt, nominale Variable
vzeq	Anzahl Vollzeitäquivalenzen des Unternehmens, stetige Variable
blm	Verhältnis Lehrlinge Männer zu Lehrlinge Total, stetige Variable
bkn	Verhältnis niedergelassene Ausländer Total zu Beschäftigte Total, stetige Variable
bkt	Verhältnis temporär vermittelte Personen Total zu Beschäftigte Total, stetigeVariable.
bkv	Verhältnis Personen AHV-Alter Total zu Beschäftigte Total, stetige Variable
bka	Verhältnis Asylbewerber Total zu Beschäftigte Total, stetige Variable
bks	Verhältnis Saisonniers Total zu Beschäftigte Total, stetige Variable
agglo	Standort Agglomeration (Ja/Nein), nominale Variable
bif	Verhältnis Beschäftigte Inhaber Total zu Beschäftigte Total, stetige Variable
be1	Verhältnis Beschäftigte 50-90 Total zu Beschäftigte Total, stetige Variable
be2	Verhältnis Beschäftigte <50 Total zu Beschäftigte Total, stetige Variable
vzkl	Vollzeitäquivalenzklassen, ordinale Variable

11.3 UNT

Eine Variante zu der Methode B ist das folgende Modell UNT, gerechnet vom BFS. Wir kritisieren es bewusst nicht im statistischen Sinn, da Erfahrung manchmal zu besseren Resultaten führen kann, als strenges Modellieren. Zu beachten gilt hier vorallem, dass durch das feste Wählen der Abschneidegrenze je nach Berechnungsklasse robuster oder gar nicht robust gerechnet wird.

Berechnungsmethode UNT:
obere Abschneidegrenze: UMVZEQ > Fr. 5'000‘000.-
untere Abschneidegrenze: UMVZEQ < Fr. 30'000.-

1) Total VZEQ pro NOGA-5steller (ohne Abschneidegrenzen)
2) Total VZEQ pro NOGA-5steller (unter Berücksichtigung der Abschneidegrenzen)
3) Total Umsatz pro NOGA-5steller (unter Berücksichtigung der Abschneidegrenzen)
4) Umsatz proVollzeitäquivalent pro NOGA-5steller: 3)/2)
5) Berechneter Umsatz pro NOGA-5steller: 4)*1)
6) Berechneter Umsatz pro NOGA-2steller: Summierung NOGA-5steller aus 5)

NOGA 2 Klasse	Umsatz nach UNT	NOGA 2 Klasse	Umsatz nach UNT
10	5	36-37	7’274
11	21	40-41	17’623
13	6	45	55’357
14	1’753	50	38’404
15	26’705	51	120’650
17	3’885	52	89’584
18	1’685	55	24’530
19	712	60	15’460
20	7’941	61	863
21	4’289	62	6’839
22	12’390	63	14’693
23	259	64	14’775
24	27’165	67	2’088
25	5’169	70	3’999
26	5’837	71	883
27	4’463	72	7’040
28	17’142	73	1’030
29	30’667	74	44’089
30	1’230	80	1’735
31	14’277	85	10’398
32	4’117	90	1’423
33	18’388	92	3’417
34	1’364	93	3’587
35	1’484		
Summe Umsatz nach UNT:			**676‘693**

11.4 Formeln und Programm

In diesem Appendix finden sich die notwendigen Formeln zu den Methoden (A1, A2, B1, B2), sowie das „Imputationsprogramm“ welches in S-Plus geschrieben ist.

In den Formeln gilt folgende Notation:

n = Anzahl Unternehmen innerhalb einer Berechnungsklasse
i = Laufindex für die Unternehmen innerhalb einer Berechnungsklasse
U = Umsatz
V = Umsatz pro VZEQ, falls VZEQ > 0 sonst Umsatz
VQ = VZEQ, falls VZEQ > 0 sonst 1
Q = 99% Quantil von V
1{ } bezeichne die Indikatorfunktion, welche den Wert 1 hat, falls die Bedingung in { } erfüllt ist, und 0 sonst.

Methode A1:

$$A1 = \sum_{i=1}^{n} U_i \cdot 1\{U_i > 0\} + \sum_{i=1}^{n} VQ_i \cdot 1\{U_i = 0\} \cdot \left(\frac{\sum_{i=1}^{n} V_i \cdot 1\{U_i > 0\}}{\sum_{i=1}^{n} 1\{U_i > 0\}} \right)$$

Methode A2:

$$A2 = \sum_{i=1}^{n} U_i \cdot 1\{0 < V_i \le Q\} + \sum_{i=1}^{n} Q \cdot VQ_i \cdot 1\{V_i > Q\} + \sum_{i=1}^{n} VQ_i \cdot 1\{U_i = 0\} \cdot \left(\frac{\sum_{i=1}^{n} V_i \cdot 1\{0 < V_i \le Q\}}{\sum_{i=1}^{n} 1\{0 < V_i \le Q\}} \right)$$

Methode B1:

$$B1 = \left(\frac{\sum_{i=1}^{n} VQ_i}{\sum_{i=1}^{n} VQ_i \cdot 1\{U_i > 0\}} \right) \cdot \sum_{i=1}^{n} U_i \cdot 1\{U_i > 0\}$$

Methode B2:

$$B2 = \left(\frac{\sum_{i=1}^{n} VQ_i}{\sum_{i=1}^{n} VQ_i \cdot 1\{U_i > 0\}} \right) \cdot \sum_{i=1}^{n} \min\{U_i, Q \cdot VQ_i\} \cdot 1\{U_i > 0\}$$

Imputationsprogramm (S-Plus 4.0):

```
# IMPUTATION NACH LOGISTISCHER REGRESSION INNERHALB
# DER PRAEDIKITIONSKLASSEN
#
# Eingelesen wird ein DataFrame mit dem Namen d0, welcher
# mindestens folgende Variabeln enthält (Variabelnamen kommen
# 5 Mal vor, da wir 5 Mal Imputieren: Bsp M1,M2,M3,M4,M5)
# M1,...,M5        Codierung ob Beob. ein MissingValue ist oder
#                  nicht (bei Missing hat M Wert 0 sonst 1)
# V1,...,V5        UMVZEQ (letzte Stelle schon zufällig gestrichen)
# PRE.1,...,PRE.5  Prädiktion aus logistischer Regression
# I1,...,I5        Codierung ob Beob. Ein extremer Wert ist oder
#                  nicht (hier 0 falls im extr. oberen 1, sonst 1)

n  <- dim(d0)[1];          #ANPASSEN AN DEN DATENSATZ

for (loop in 1:5){
if (loop ==1){ u <- d0$V1; m <- d0$M1; p <- d0$PRE.1; I1<- d0$I1;}
if (loop ==2){ u <- d0$V2; m <- d0$M2; p <- d0$PRE.2; I1<- d0$I2;}
if (loop ==3){ u <- d0$V3; m <- d0$M3; p <- d0$PRE.3; I1<- d0$I3;}
if (loop ==4){ u <- d0$V4; m <- d0$M4; p <- d0$PRE.4; I1<- d0$I4;}
if (loop ==5){ u <- d0$V5; m <- d0$M5; p <- d0$PRE.5; I1<- d0$I5;}

kl<-1;
i<-1; iu<-0; ip<-0; in1<-0;  #Erzeugung der Missing Matrix
s<-0;
while (i <=n ) {
  if (m[i] == 0) {s<-s+1; iu[s]<-0; ip[s]<-p[i]; in1[s]<-i;}
  i<-i+1;}
s<-s+1; s2<-s-1;
iu[s]<-0; ip[s]<-2; in1[s]<-n+1; mat1<-c(iu,ip,in1);
dim(mat1)<-c(s,3); mat1 <- mat1[order(mat1[,2]),1:3];
rm(s,iu,ip,in1,i);

n2 <- sum(m)+sum(I1)-n;      #Erzeugung der Donors Matrix
klasse <- ceiling(sqrt(n2));
pred <- 0; um1 <- 0; k<-1; i<-1;
while (i <= n) {
  if (1==(0.5*m[i]+0.5*I1[i])){pred[k]<-p[i];um1[k]<-u[i];k<-k+1;}
  i<-i+1;}
n2<-n2+1; um1[n2]<-0; pred[n2]<-2; imp<-c(um1, pred);
dim(imp)<-c(n2, 2); imp <- imp[order(imp[,2]),1:2];
mw <- mean(um1); rob<- max(um1); rm(um1, i, k);
quant <- quantile(pred, probs=seq(0,1,(1/klasse)));
quant[1]<-0; quant[klasse+1]<-1;

q<-1; umsa<-rep(0,n2);       #klasse
while (kl <= klasse){
  s <-0;
  while (imp[q,2] <= quant[kl+1]) {
    s<-s+1;
    umsa[s]<-imp[q,1];
    q<-q+1;}

  #Imputier-Step fuer den Umsatz
  i<-1;
  while (mat1[i,2] <= quant[kl+1]) {z<-ceiling(s*runif(1,0,1));
        mat1[i,1]<-umsa[z]; i<-i+1;}
  rm(s,i);
  print(klasse-kl);
  kl<-kl+1;} #END WHILE SCHLEIFE KLASSE

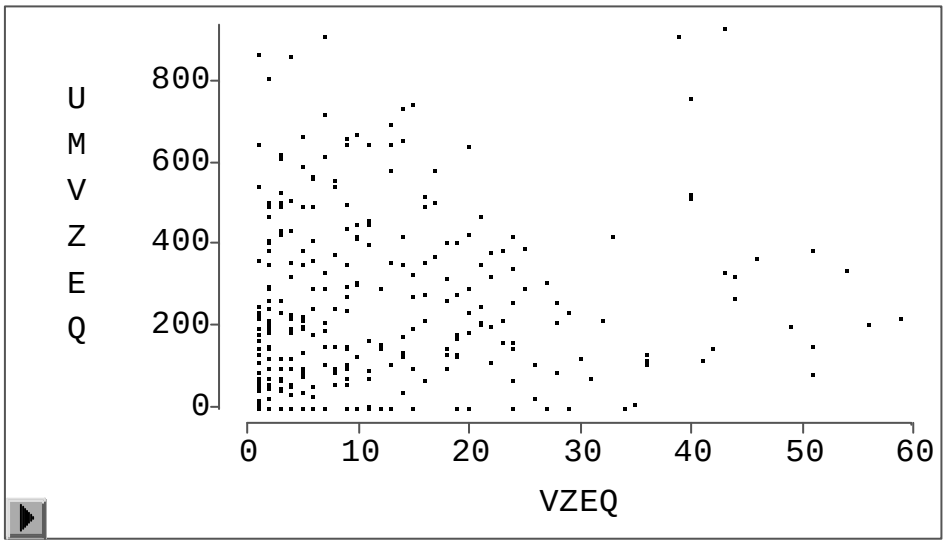
rm(I1,imp,kl,umsa,klasse,m,n2,p,pred,q,quant,u,z);
i<-1;
if (loop ==1) {while (i <= s2){
  w<-mat1[i,3];
  d0$V1[w]<-mat1[i,1];
  i<-i+1;}
```

```
  rm(mat1,s2,i,w);}
if (loop ==2) {while (i <= s2){
  w<-mat1[i,3];
  d0$V2[w]<-mat1[i,1];
  i<-i+1;}
  rm(mat1,s2,i,w);}
if (loop ==3) {while (i <= s2){
  w<-mat1[i,3];
  d0$V3[w]<-mat1[i,1];
  i<-i+1;}
  rm(mat1,s2,i,w);}
if (loop ==4) {while (i <= s2){
  w<-mat1[i,3];
  d0$V4[w]<-mat1[i,1];
  i<-i+1;}
  rm(mat1,s2,i,w);}
if (loop ==5) {while (i <= s2){
  w<-mat1[i,3];
  d0$V5[w]<-mat1[i,1];
  i<-i+1;}
  rm(mat1,s2,i,w);}
}
d0$U1<-d0$VZEQ*d0$V1;
d0$U2<-d0$VZEQ*d0$V2;
d0$U3<-d0$VZEQ*d0$V3;
d0$U4<-d0$VZEQ*d0$V4;
d0$U5<-d0$VZEQ*d0$V5;
d0$UM<-d0$VZEQ*rep(mw,n);
d0$UR<-d0$U1;
i<-1;
while (i<=n){
d0$UR[i]<-min(d0$UR[i],d0$VZEQ[i]*rob);
i<-i+1;
}
rm(n,i,rob,mw);
sum(d0$U1); # liefert U1
sum(d0$U2); # liefert U2
sum(d0$U3); # liefert U3
sum(d0$U4); # liefert U4
sum(d0$U5); # liefert U5
```

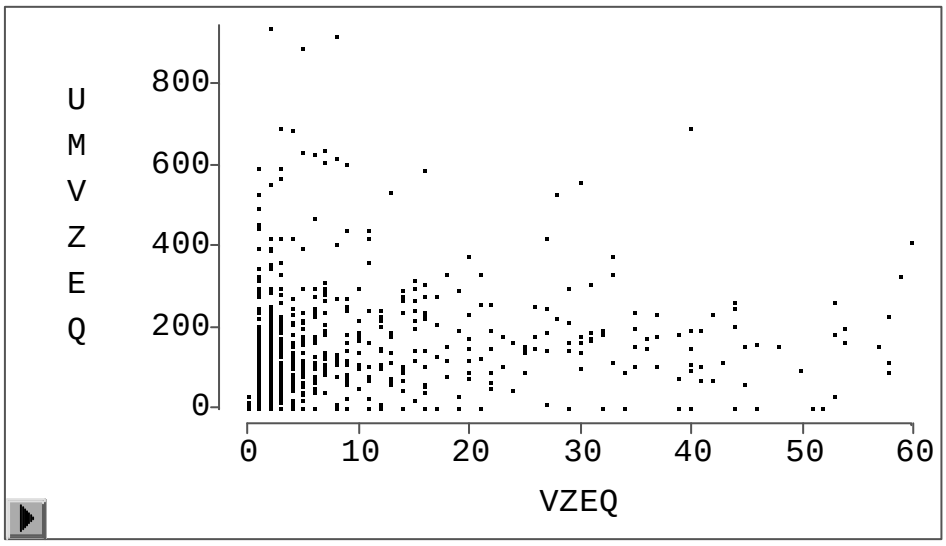
11.5 Scatterplots UMVZEQ / VZEQ

Scatterplots von VZEQ gegen UMVZEQ (in 1000 Fr) für einige nicht zu grosse NOGA 2 Klasse, helfen die Modellannahmen zu diskutieren. Es ist gut zu sehen, dass die Verteilung der UMVZEQ in Abhängigkeit der VZEQ je nach NOGA 2 Klasse verschieden ist. Das heisst für die Wahl einer geeigneten einfachen Methode, dass bei einigen NOGA 2 Klasse Methode A, respektive Methode B die geeignetere Methode ist, und somit weder Methode A, noch Methode B, für alle NOGA 2 Klassen die passende Methode ist.
Scatterplots für grosse NOGA 2 Klassen sind nicht sinnvoll, da bei diesen zu viele Punkte gezeichnet würden und somit nur ein schwarzes Feld zu erkennen wäre. Bei den folgenden Scatterplots wird der Ausschnitt VZEQ kleiner gleich 60 und UMVZEQ kleiner gleich 1 Mio. Fr gezeichnet. Ein Punkt kann unter Umständen für mehrere Unternehmen stehen (bei kleinem UMVZEQ und zugleich tiefen VZEQ ist dies in den grossen NOGA 2 Klassen häufig der Fall). Der Umsatz pro Vollzeitäquivalenz ist in den folgenden Figuren in KFr. angegeben.

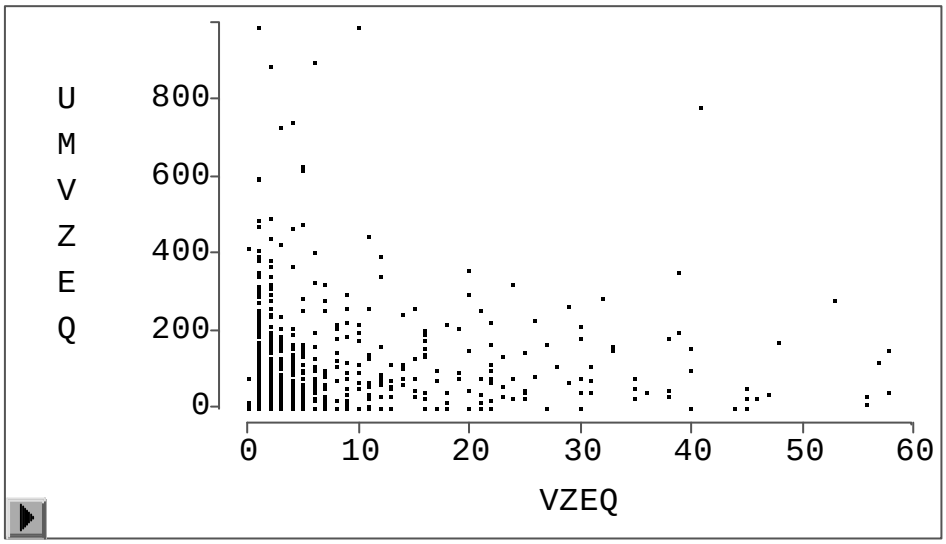
NOGA 2 = 14



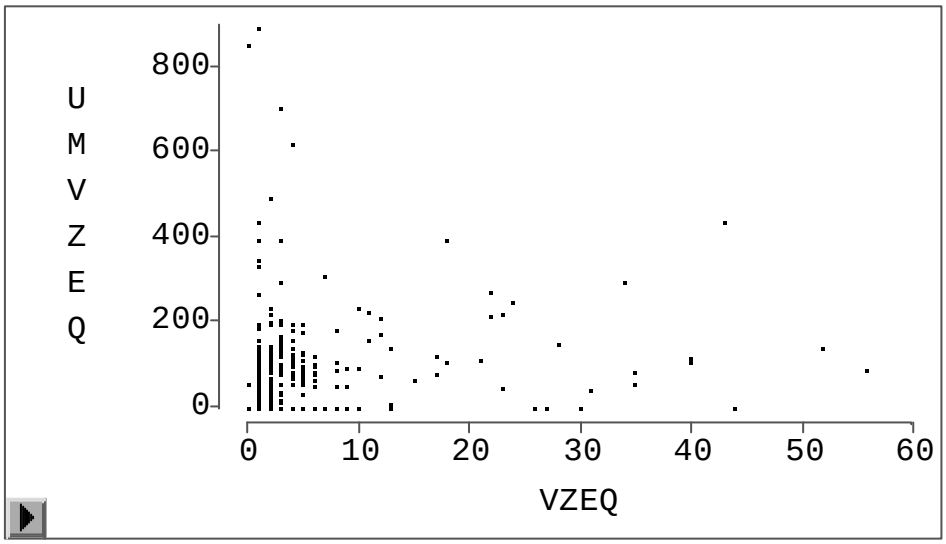
NOGA 2 = 17



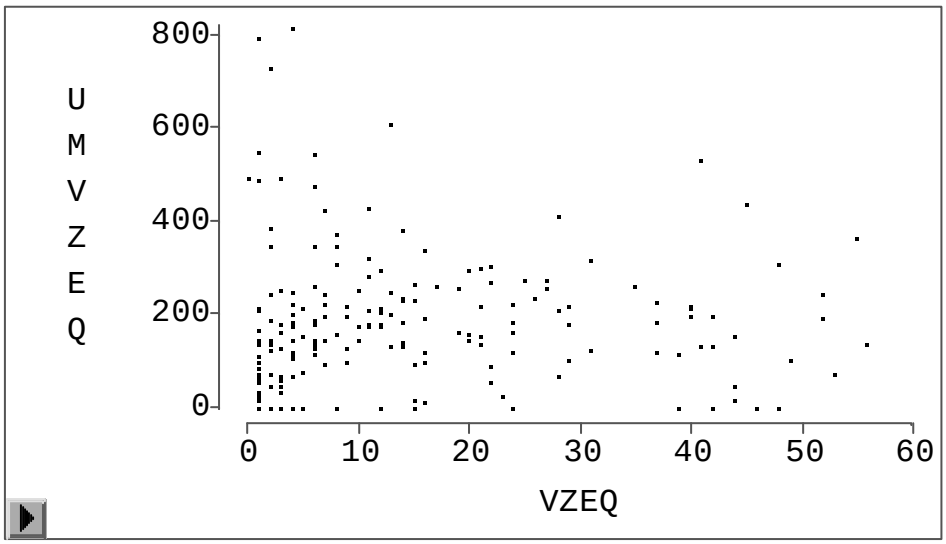
NOGA 2 = 18



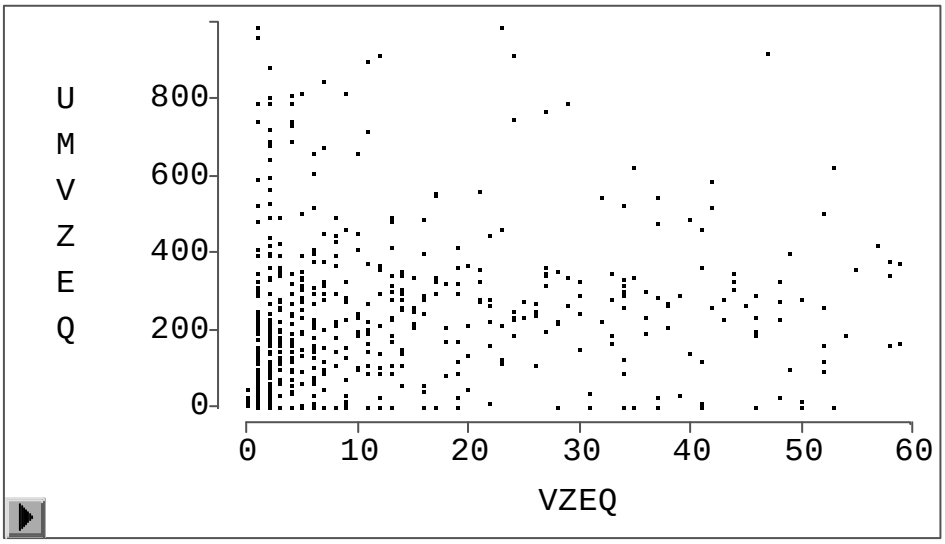
NOGA 2 = 19



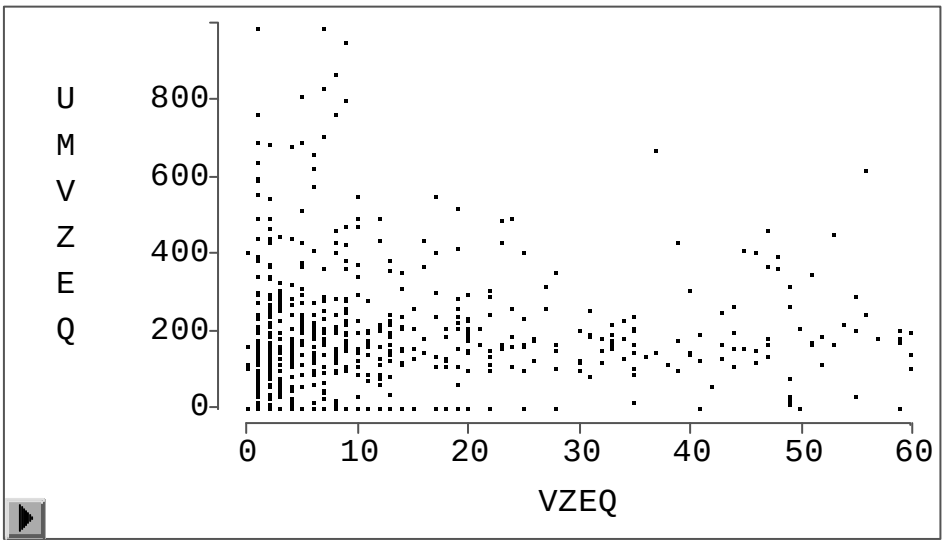
NOGA 2 = 21



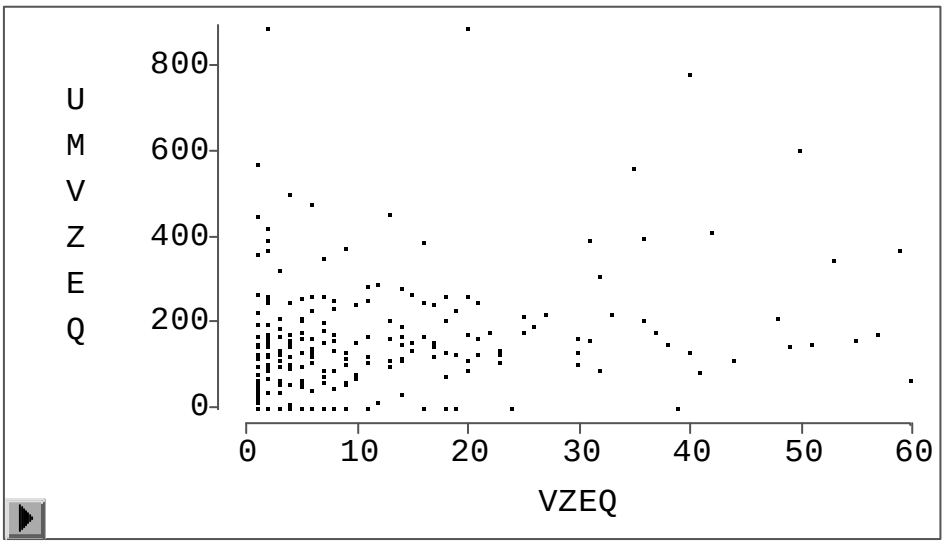
NOGA 2 = 24



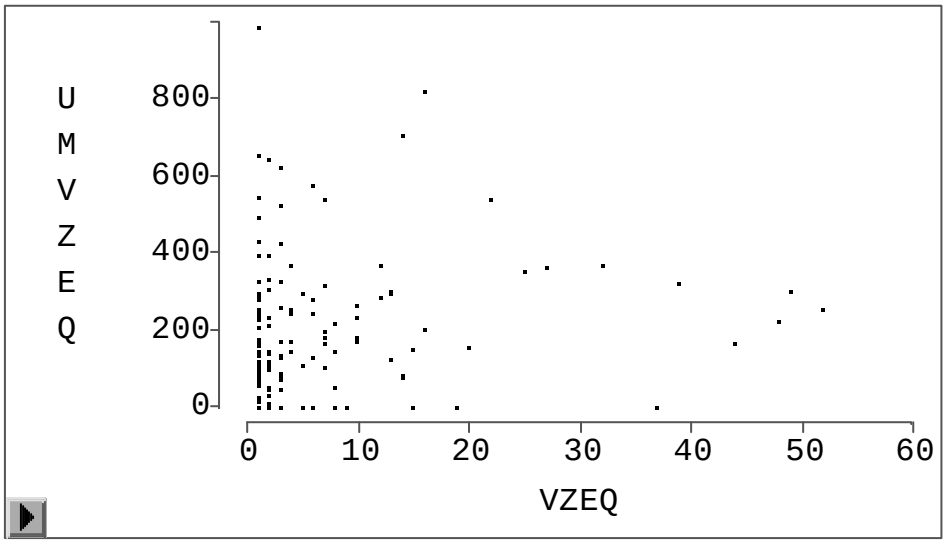
NOGA 2 = 25



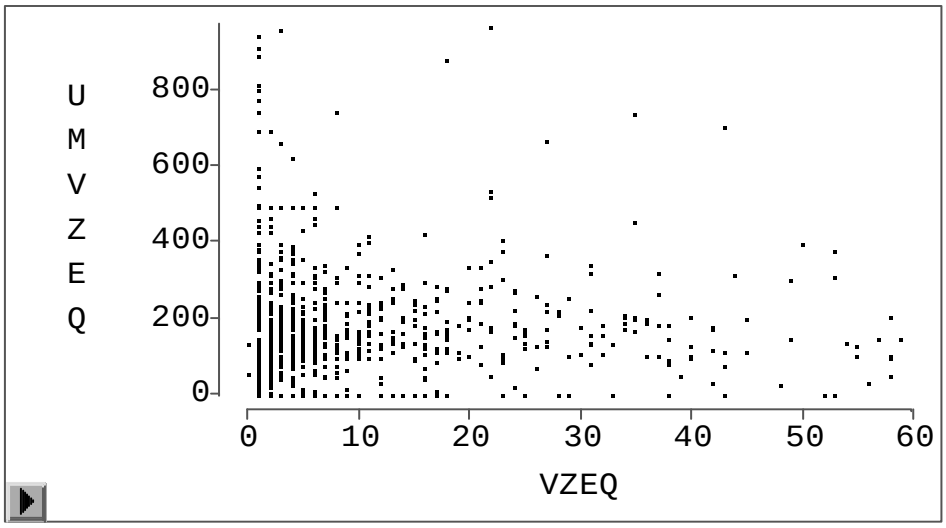
NOGA 2 = 27



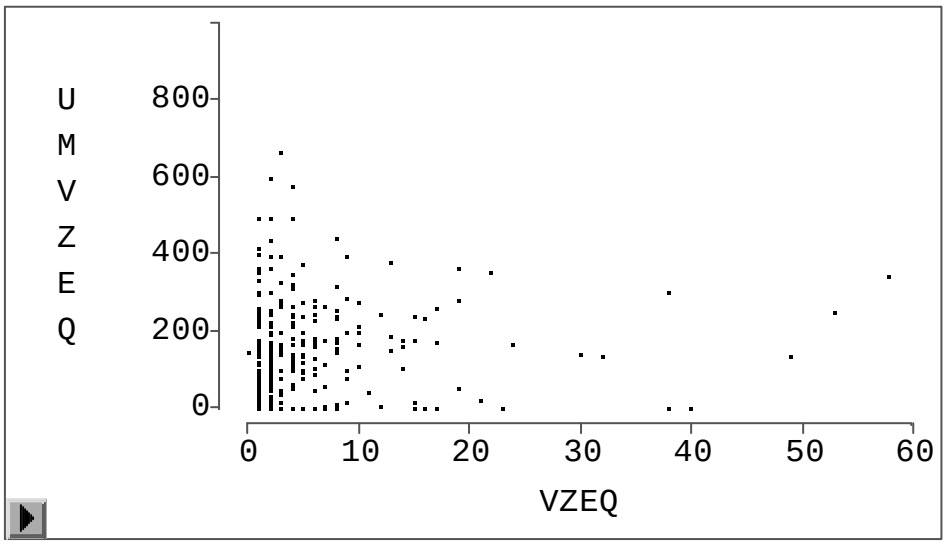
NOGA 2 = 30



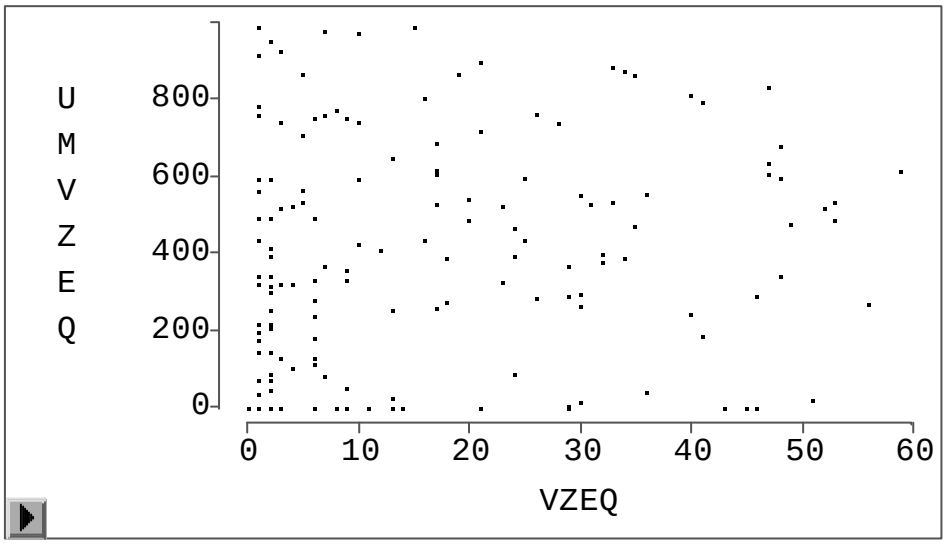
NOGA 2 = 31



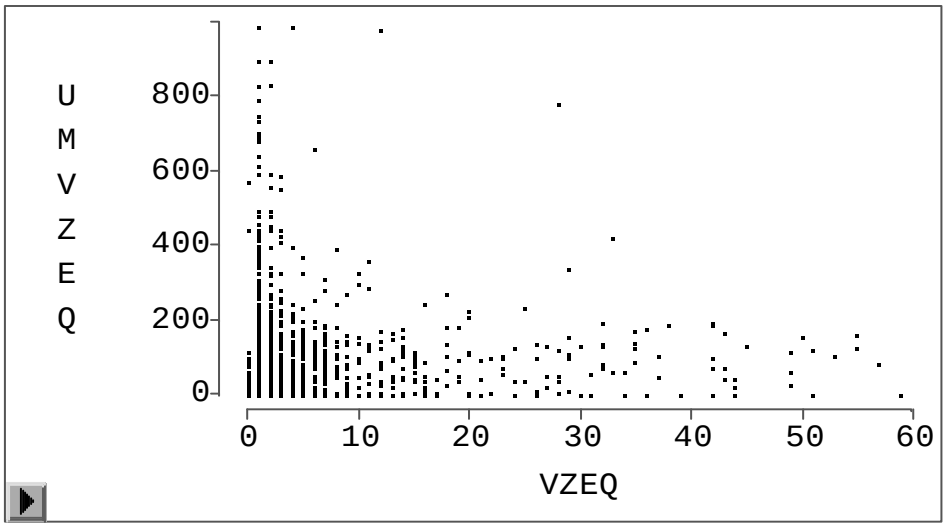
NOGA 2 = 35



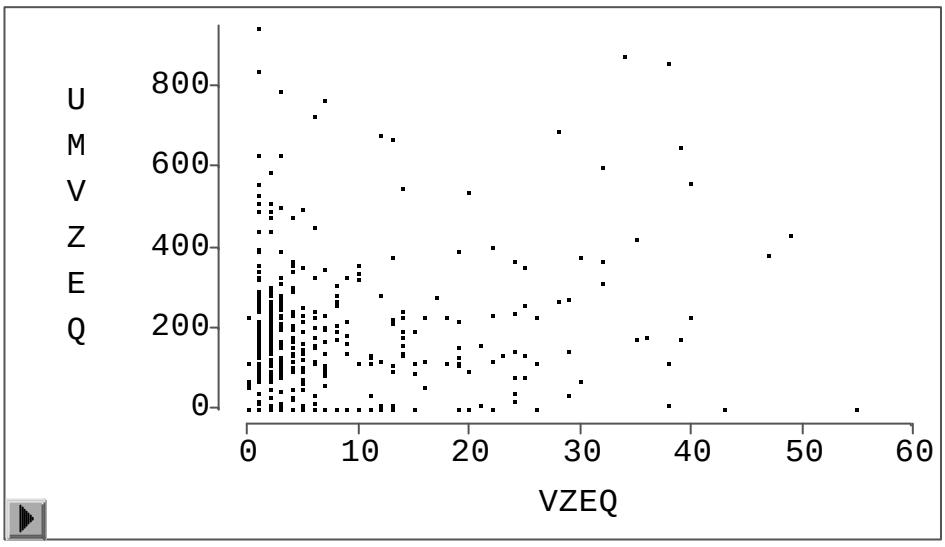
NOGA 2 = 40



NOGA 2 = 80



NOGA 2 = 90

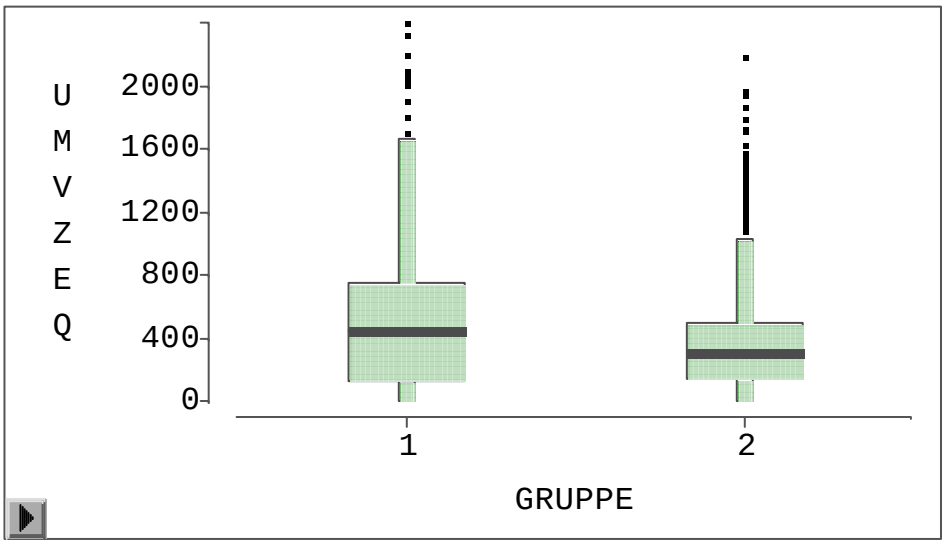


11.6 Boxplots UMVZEQ

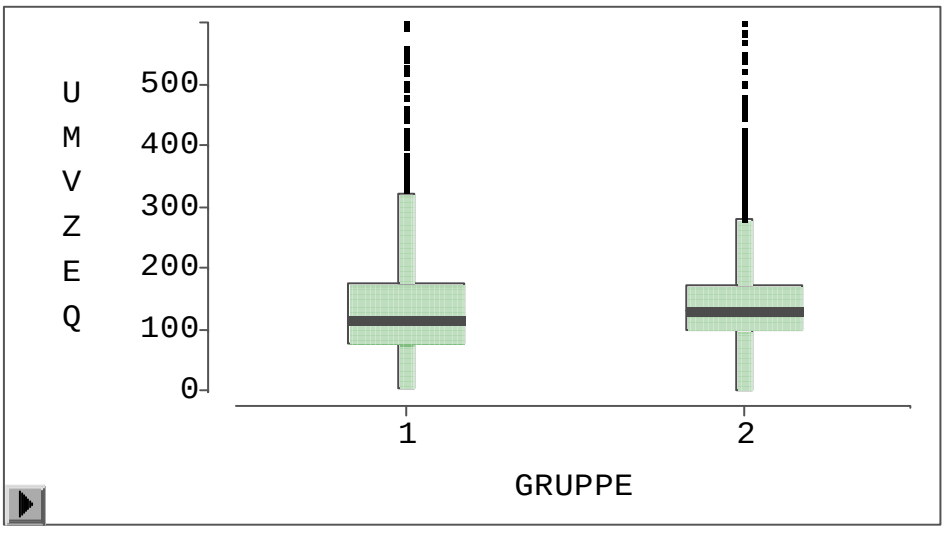
Boxplots von UMVZEQ derjenigen NOGA 2 Klassen, für welche 2 Berechnungsklassen gemacht wurden

Gruppe 1 steht für Unternehmen mit VZEQ gleich 1 oder 2. Gruppe 2 steht für Unternehmen mit VZEQ grösser gleich 3. Die extremsten Werte sind nicht eingezeichnet, da man sonst von den nicht extremen Werten nichts mehr erkennen könnte. Der fette Strich in der Mitte des breiten Kästchens stellt den Median dar. Der obere (untere) Strich des breiten Kästchens ist das 75-Quantil (25-Quantil). Das obere schmale Kästchen ist so lang wie der 1.5-fache Quartilsabstand.
Diese Boxplots sind geeignet, um auf einfache Art und Weise die Verteilungsunterschiede von UMVZEQ der Unternehmen in Gruppe 1 zu den Unternehmen in Gruppe 2 zu zeigen. Bei allen NOGA 2 Klassen ist die schiefe Verteilung der UMVZEQ deutlich ersichtlich.

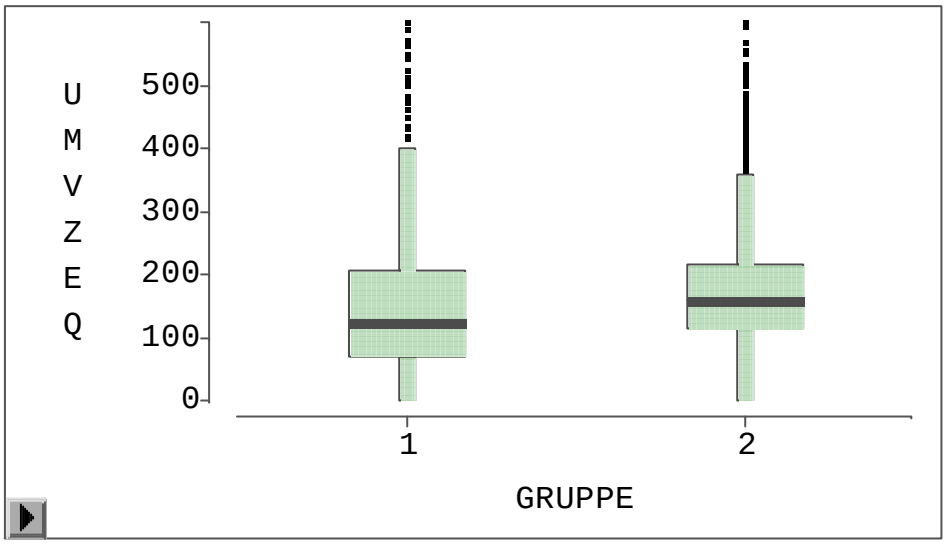
NOGA 2 = 15



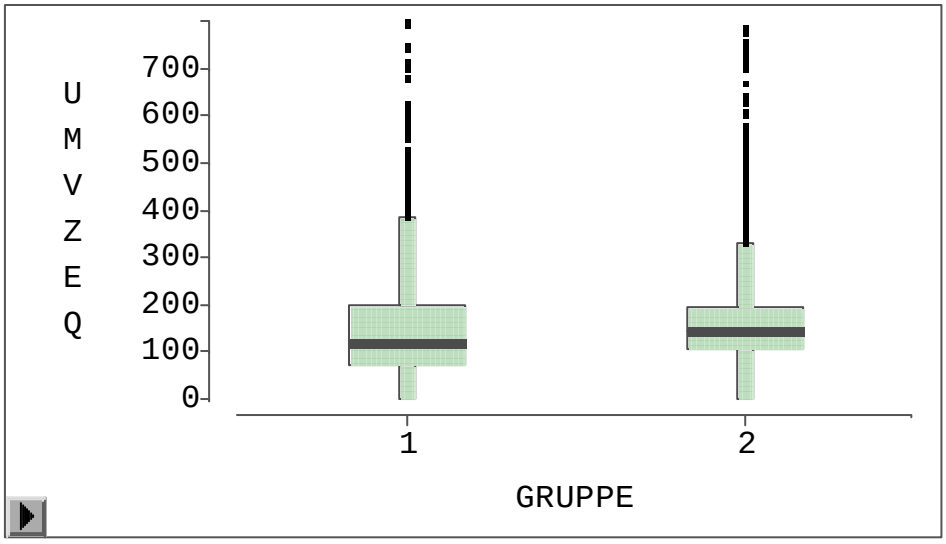
NOGA 2 = 20



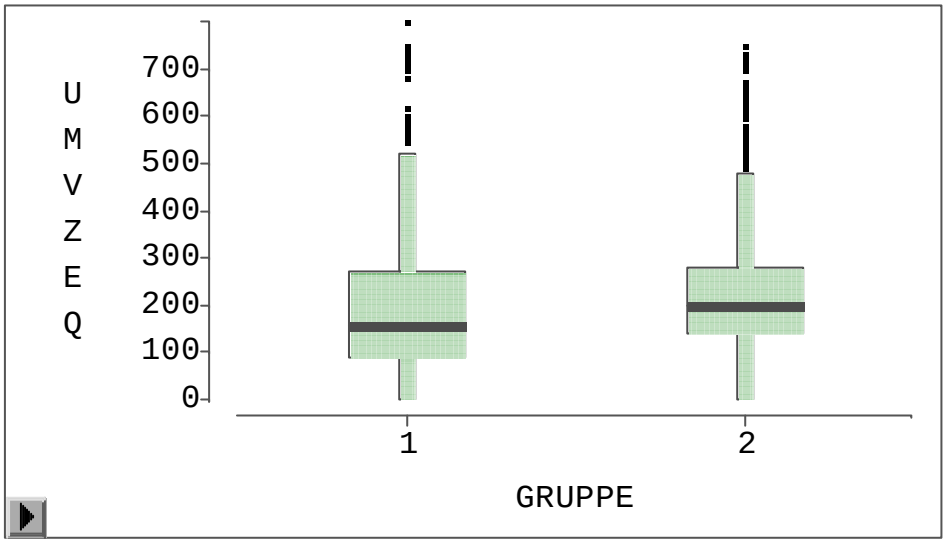
NOGA 2 = 22



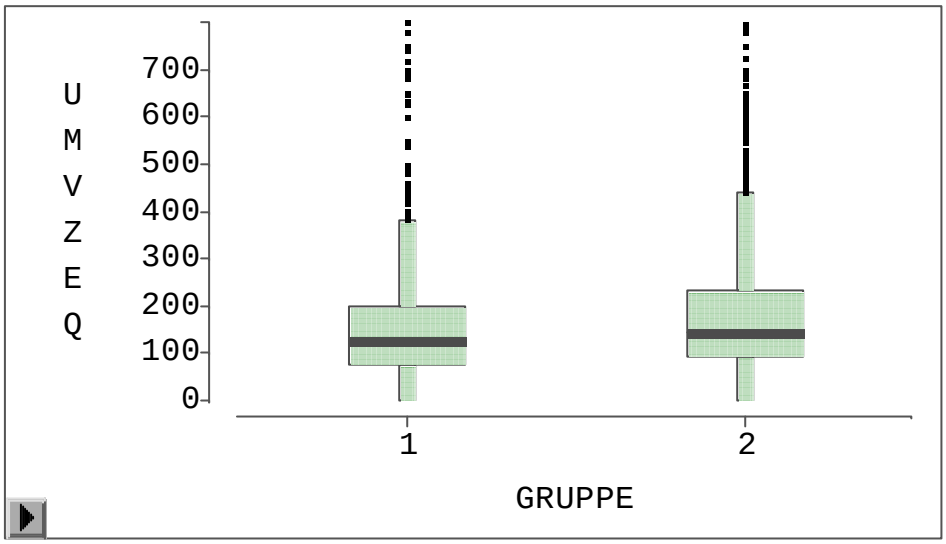
NOGA 2 = 28



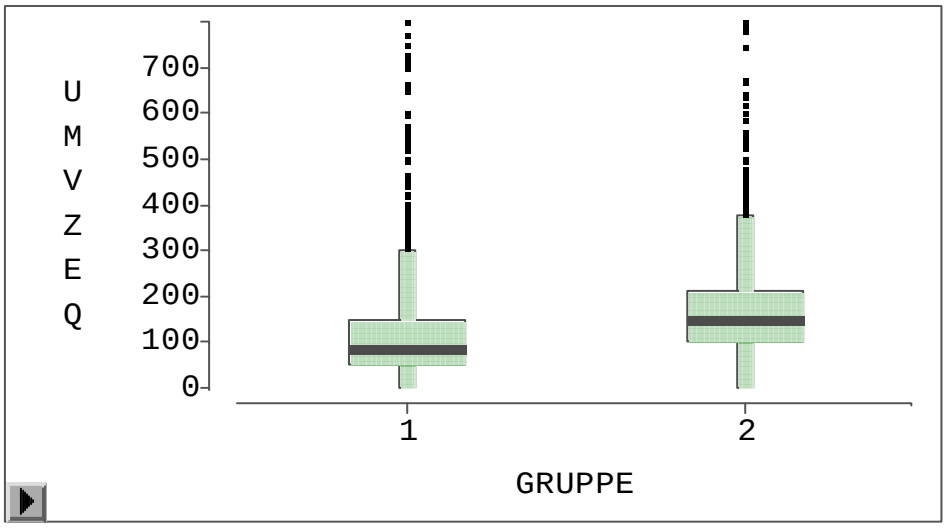
NOGA 2 = 29



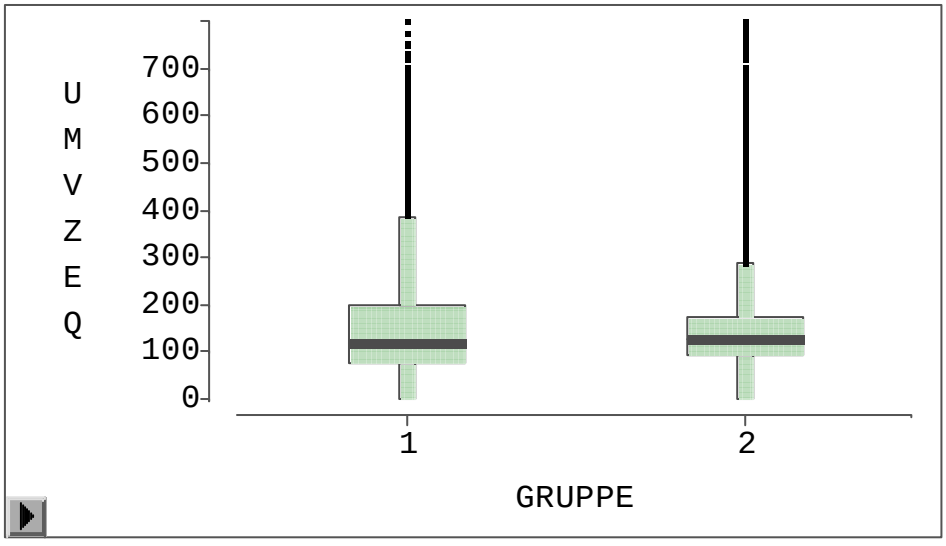
NOGA 2 = 33



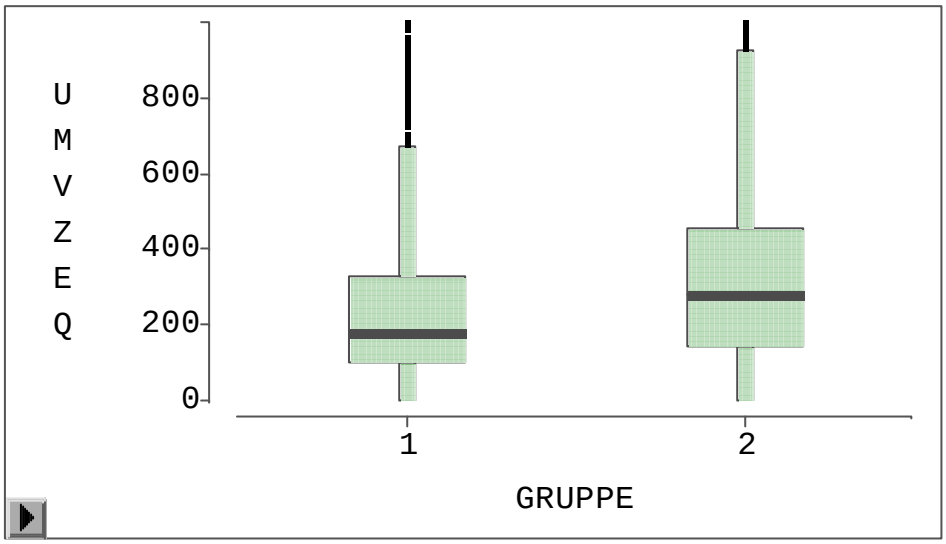
NOGA 2 = 36



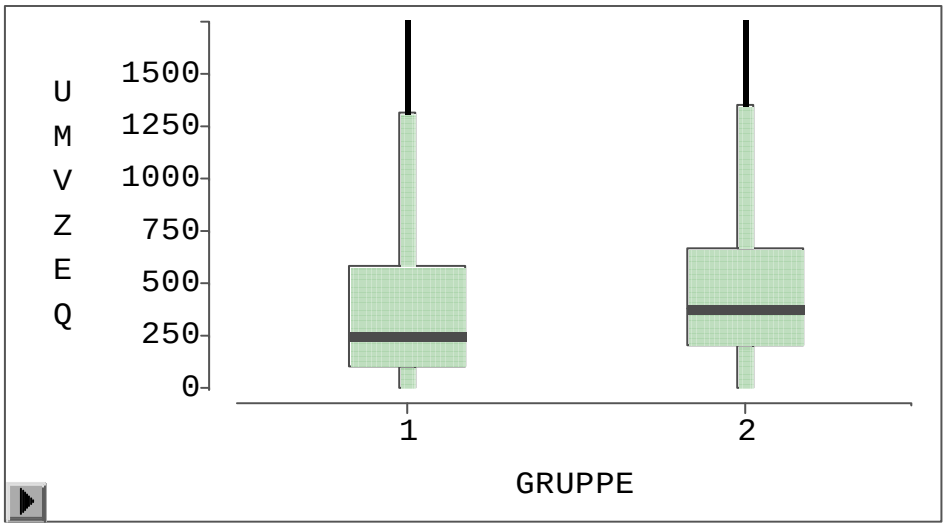
NOGA 2 = 45



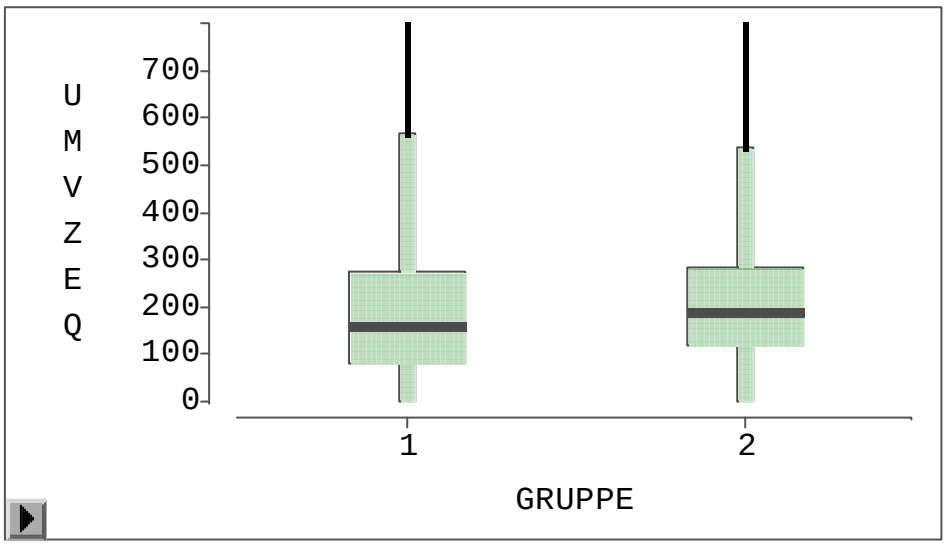
NOGA 2 = 50



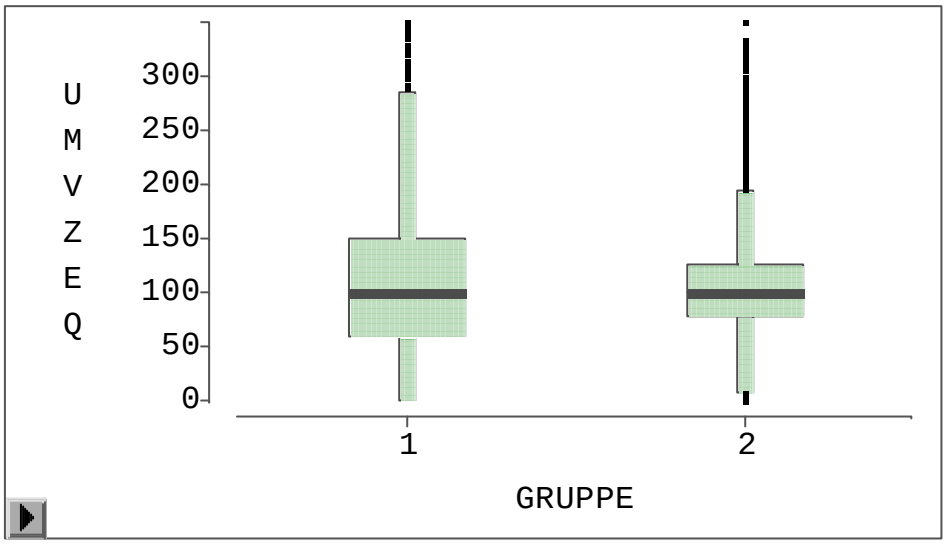
NOGA 2 = 51



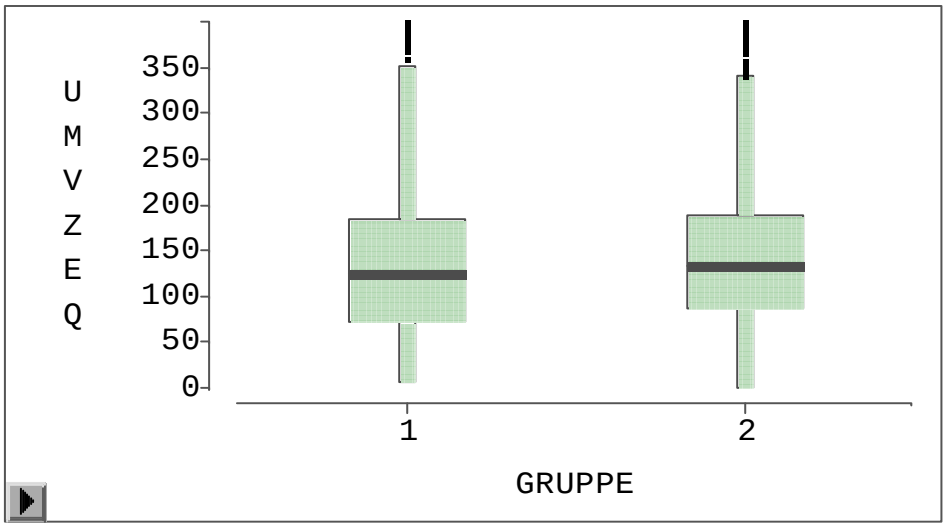
NOGA 2 = 52



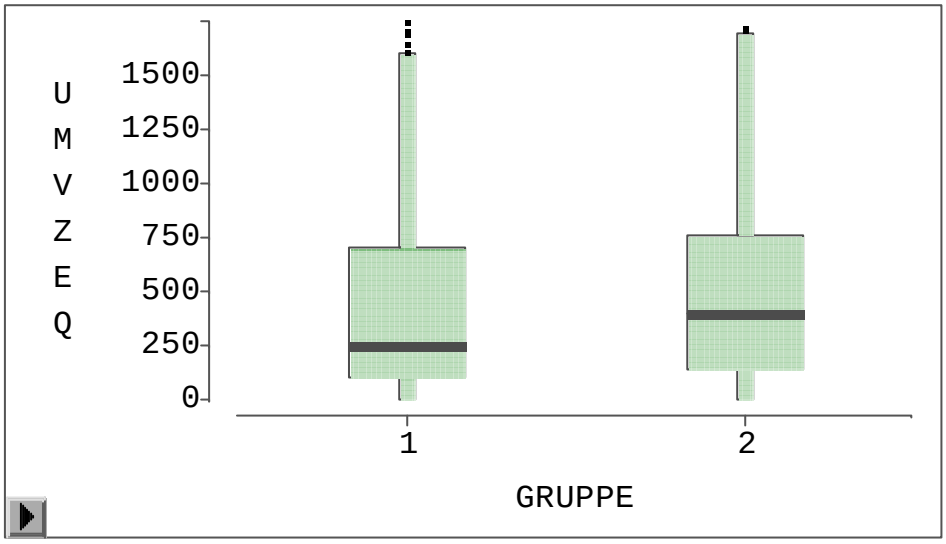
NOGA 2 = 55



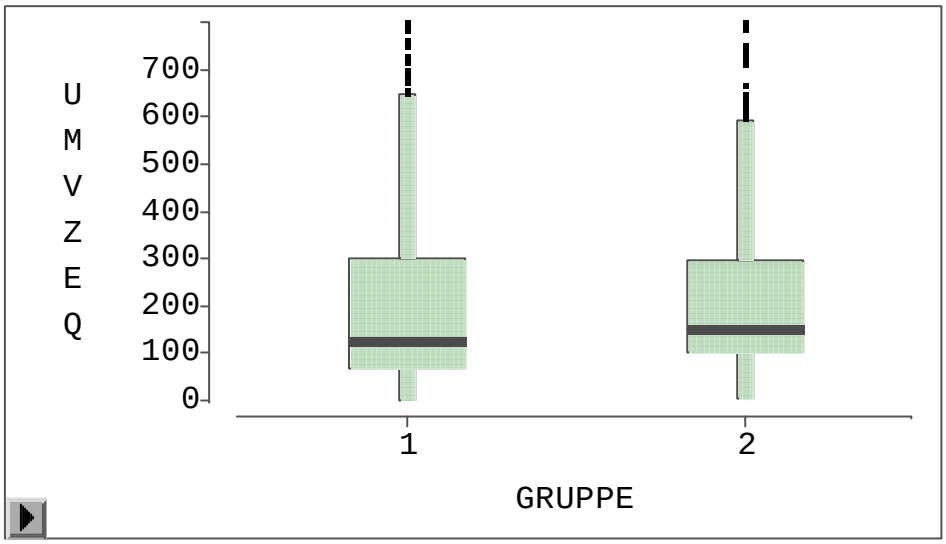
NOGA 2 = 60



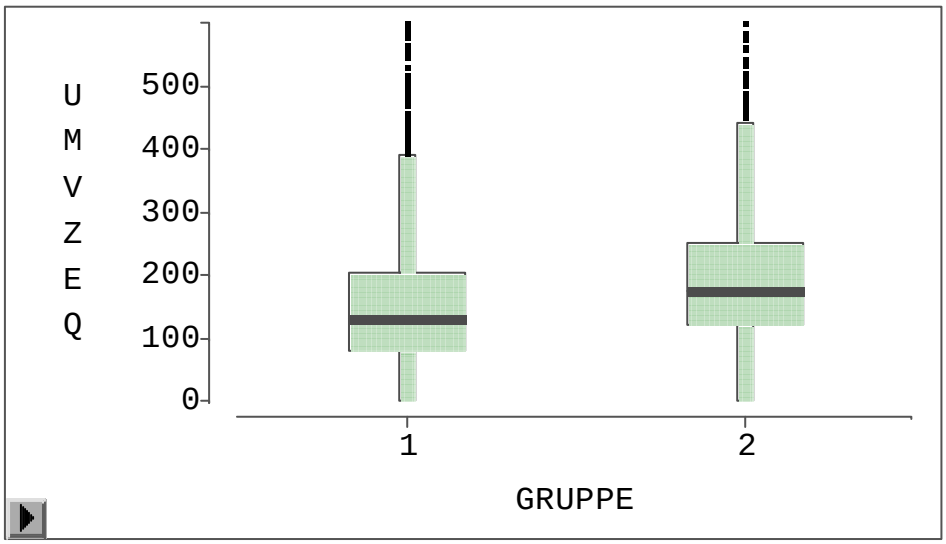
NOGA 2 = 63



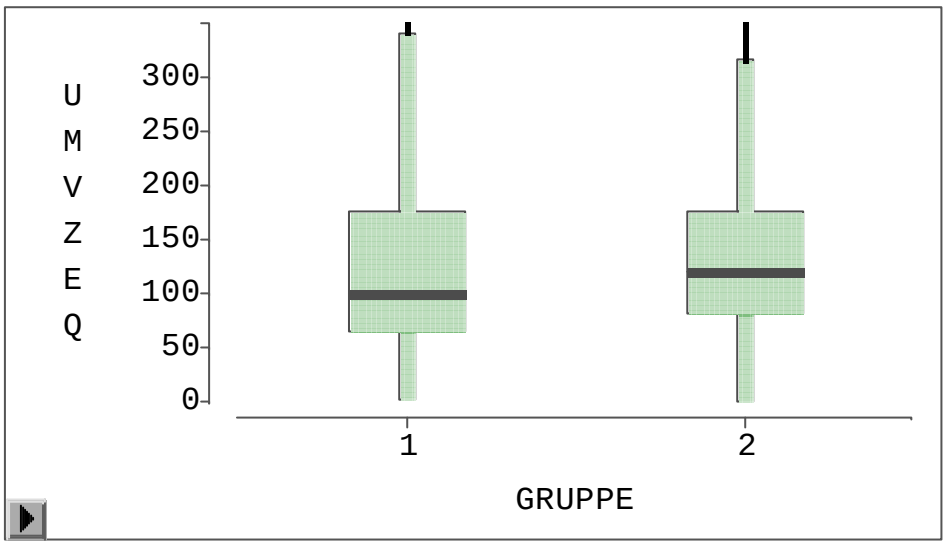
NOGA 2 = 70



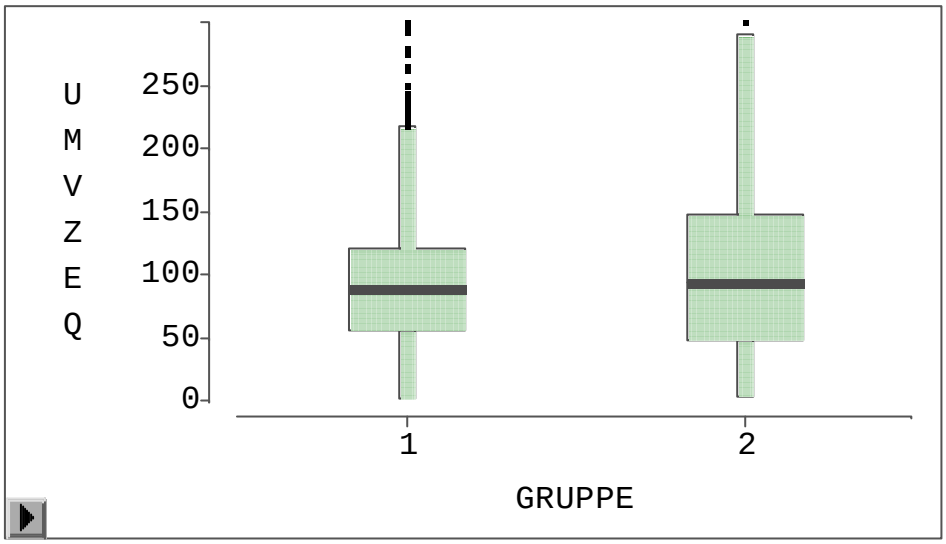
NOGA 2 = 72



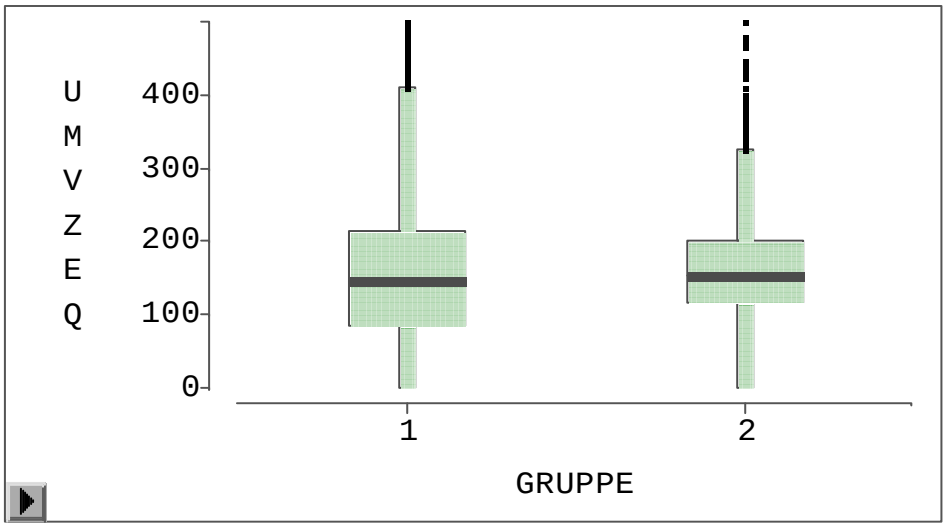
NOGA 2 = 74



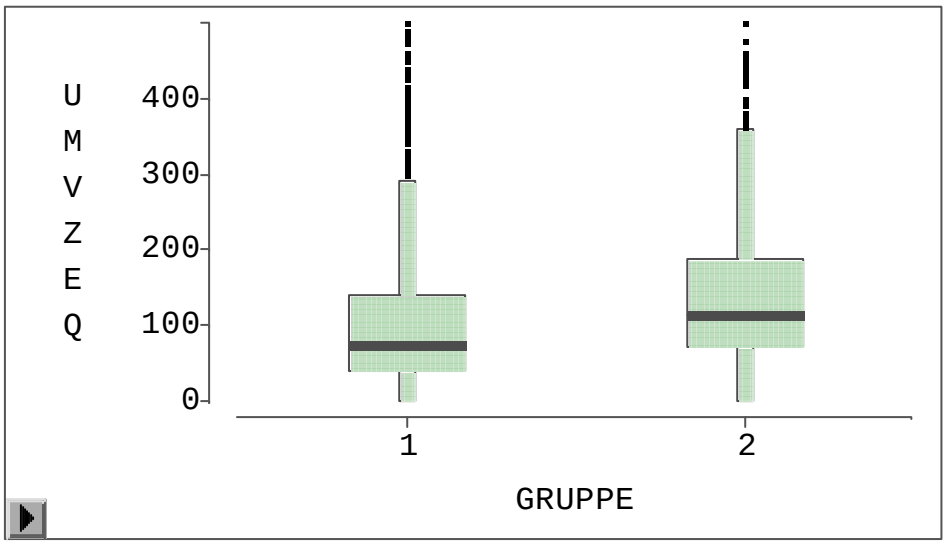
NOGA 2 = 80



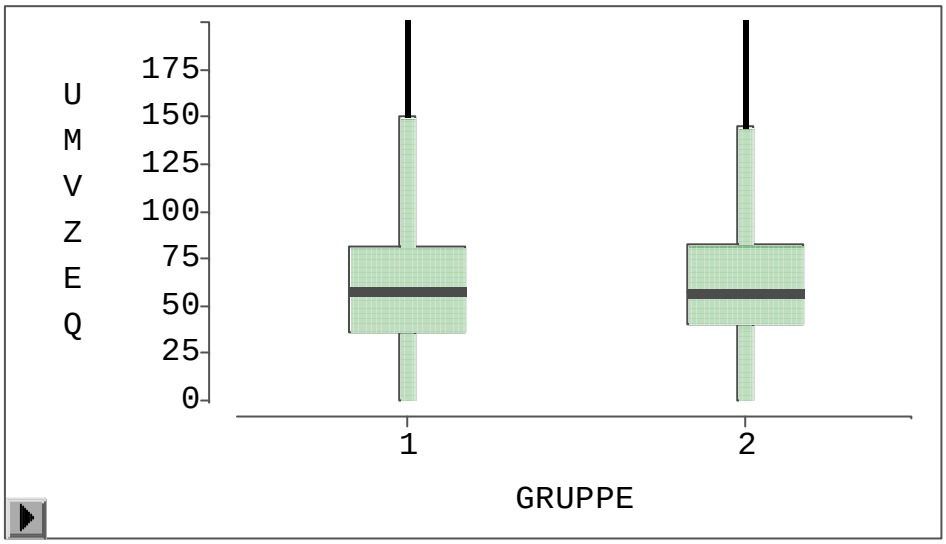
NOGA 2 = 85



NOGA 2 = 92



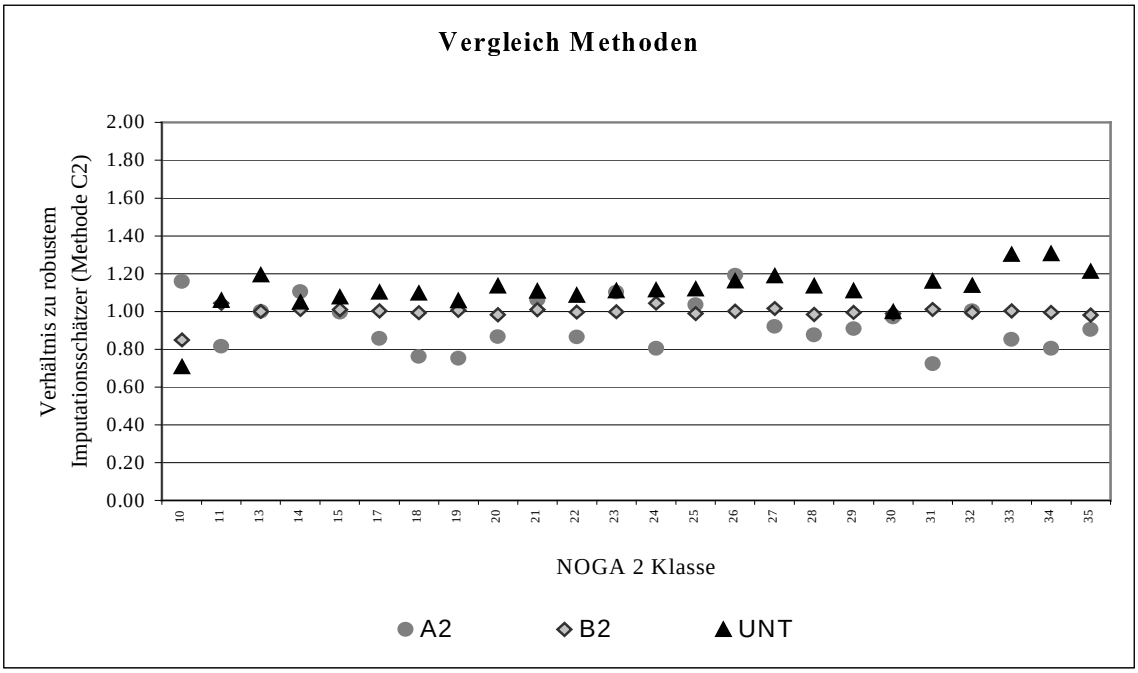
NOGA 2 = 93

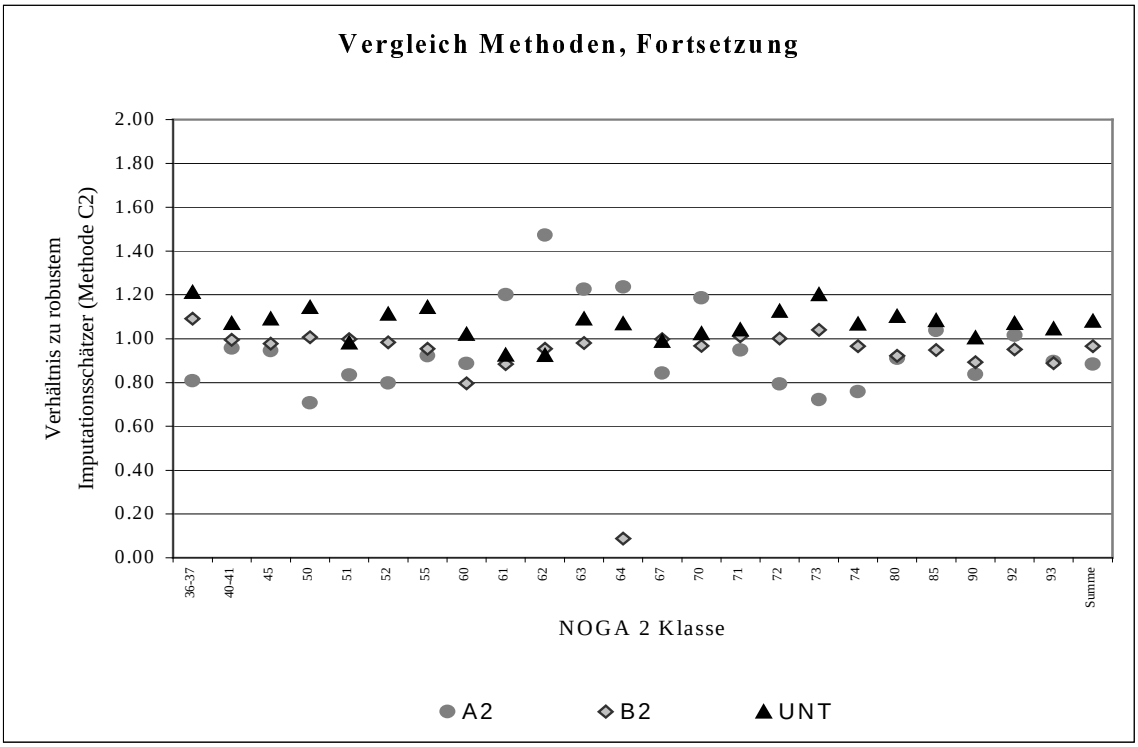


11.7 Scatterplot der relativen Verhältnisse

Scatterplot prozentuale Aenderungen zu Methode C2 für NOGA 2 Klassen

Zeigt die relativen Umsatzberechnungen der zwei einfachen robusten Methoden A2 und B2 und dem Modell UNT sowie der von uns vorgeschlagenen mehrfachen robusten Imputation C2.

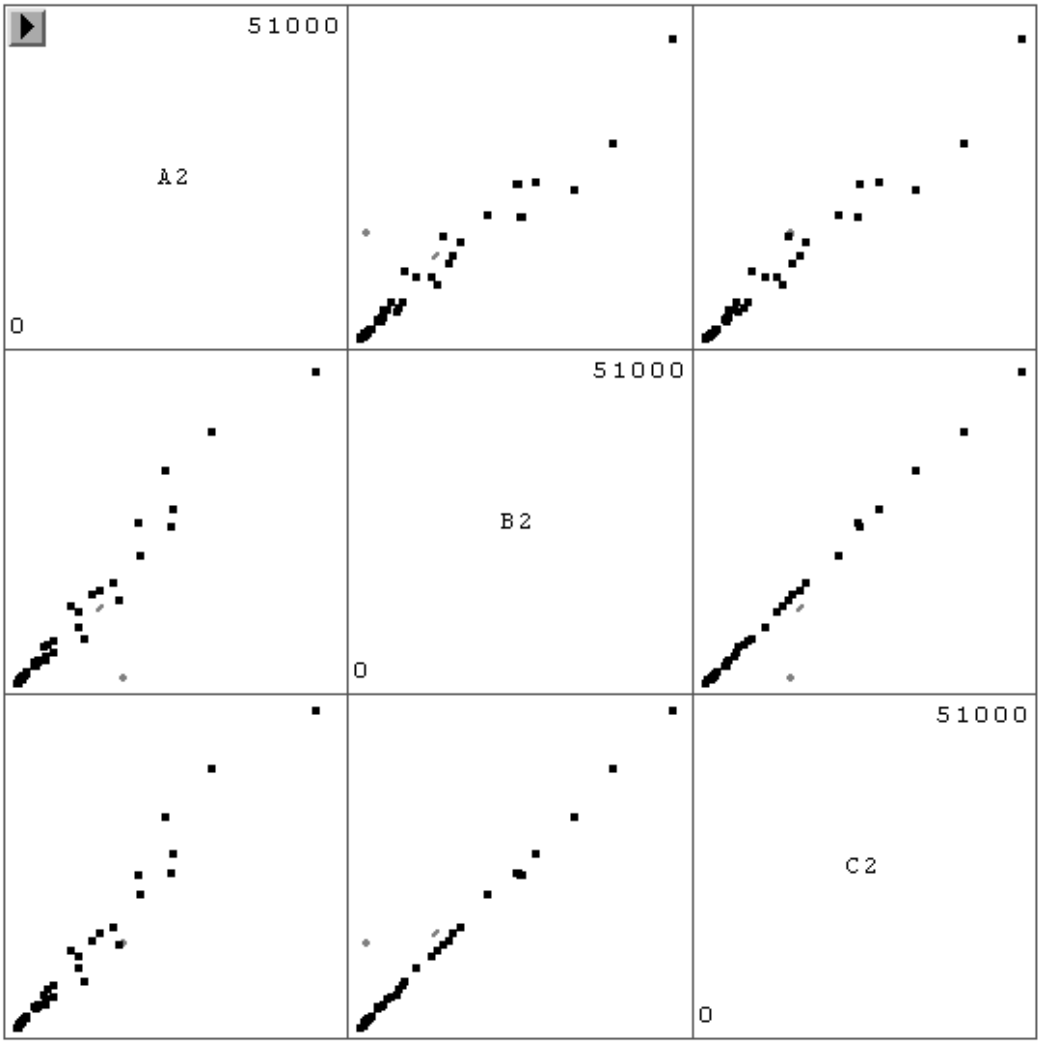




11.8 Scatterplot Methodenvergleich

Scatterplotmatrix Methode A2, B2 und C2

In dieser Grafik sieht man sehr gut, wie die hochgerechneten Umsatzzahlen der verschiedenen Methoden untereinander korrelieren. Die Methoden B2 und C2 liefern sehr ähnliche Werte, bis auf die NOGA 2 Klassen 60 (flaches Rechteck) und 64 (Karo). Beide Achsen sind gleich skaliert, wobei die NOGA 2 Klassen 51 und 52 nicht gezeichnet sind (siehe Grafik vorher), da sie sehr grosse Gesamtumsätze haben. Dadurch werden die Unterschiede der Umsatzberechnungen aller NOGA Klassen deutlich. Aus Tabelle 2 ist ersichtlich, dass die Umsätze von NOGA 2 Klasse 51 und 52 mit Methode B2 und C2 fast identisch sind, jedoch wesentlich höher als mit A2. Es muss beachtet werden, dass die NOGA 2 Klassen 51 und 52 zusammen etwa einen Drittel des Gesamtumsatzes bestimmen. Die Behandlung der fehlenden Umsatzzahlen in diesen zwei Klassen beeinflusst somit massgeblich die Bestimmung des Gesamtumsatzes.



Methodenberichte des Dienstes Statistische Methoden
Rapports de méthodes du Service de méthodes statistiques
Methodological reports of the Statistical Methods Unit

Hulliger, B., Peters, R. (1994). Enquête sur la structure des salaires en Suisse: stratégie d'échantillonnage pour le secteur privé.

Peters, R., Hulliger, B. (1994). La technique de pondération des données: application à l'enquête suisse sur la santé.

Schwendener, P. (1995). Verbrauchserhebung 1990 - Vertrauensintervalle.

Hulliger, B. (1995). Konjunkturelle Mietpreiserhebung: Stichprobenplan und Schätzverfahren.

Peters, R. (1995). Enquête de structure sur les loyers: établissement d'un plan d'échantillonnage stratifié.

Salamin, P.A. (1995). Estimation des flux pour le module II des comptes globaux du marché de travail.

Hulliger, B. (1996). Gütertransportstatistik 1993: Schätzverfahren mit Kompensation der Antwortausfälle.

Hulliger, B., Peters, R. (1996). Enquête sur le comportement de la population suisse en matière de transport en 1994: plan d'échantillonnage et pondération.

Peters, R. (1996). Schéma de pondération des indices PAUL.

Peters, R., Hulliger, B. (1996). Schätzverfahren für die Lohnstruktur-Erhebung 1994 / Procédure d'estimation pour l'enquête de 1994 sur la structure des salaires.

Hulliger, B. (1996). Haushalterhebung Familie 1994: Stichprobenplan, Stichprobenziehung und Reservestichproben.

Comment, T., Hulliger, B., Ries, A. (1996). Gewichtungsverfahren für die Schweizerische Arbeitskräfteerhebung (1991-1995).

Peters, R. (1996). Pondération des données de l'enquête sur la famille en Suisse.

Peters, R. (1997). Enquête 1996 sur la structure des salaires en Suisse: établissement du plan d'échantillonnage.

Peters, R. (1997). Etablissement du plan d'échantillonnage pour l'enquête 1996 sur la recherche et le développement dans l'économie privée en Suisse.

Salamin, P.A. (1997). Evaluation de la Statistique de l'emploi.

Bender, A., Hulliger, B. (1997). Enquête suisse sur la population active: rapport de pondération pour 1996.

Peters, R. (1998). Extrapolation des données de l'enquête de structure sur les loyers.

Salamin, P.A. (1998). Etablissement d'une clef de passage pondérée entre l'ancienne (NGAE 85) et la nouvelle nomenclature (NOGA 95) générale des activités économiques.

Hulliger, B., Kassab, M. (1998). Evaluation of Estimation Methods for the Survey on Environment Protection Expenditures of Swiss Communes.

de Rossi, F.-X. (1998). Méthodes statistiques pour le compte routier suisse.

Hulliger, B., Eichenberger, P. (2000). Stichprobenregister für Haushalterhebungen: Umstellung auf Telefonnummern ohne Namen und Adressen, Abläufe für Erstellung und Stichprobenziehung. Bestellnummer: 338-0000

Renaud, A. (2001). Statistique suisse des bénéficiaires de l'aide sociale. Plan d'échantillonnage des communes. Numéro de commande: 338-0001

Hüsler, J., Müller, S. (2001). Schlussbericht Betriebszählung 1995 (BZ 95), Mehrfach imputierte Umsatzzahlen. Bestellnummer: 338-0002

Rapports de méthodes du Service de méthodes statistiques

Une liste complète des titres de cette série de publications figure à la fin de ce rapport. Les rapports de méthodes peuvent être commandés auprès de:

Office fédéral de la statistique
Service des publications
Espace de l'Europe 10
CH-2010 Neuchâtel
Tél.: (+41 32) 71 36060
Fax: (+41 32) 71 36061
Internet: Order@bfs.admin.ch

Methodenberichte des Dienstes Statistische Methoden

Eine vollständige Liste der Titel dieser Serie befindet sich am Ende dieses Berichts. Die Methodenberichte können bestellt werden bei:

Bundesamt für Statistik
Publikationsdienst
Espace de l'Europe 10
CH-2010 Neuchâtel
Tel.: (+41 32) 71 36060
Fax: (+41 32) 71 36061
Internet: Order@bfs.admin.ch