Institut Wohnen
und Umwelt

Olivia Denise Heuer

Einfluss regionaler Faktoren auf den Wohnungsmietpreis

Veröffentlichung der an der Hochschule Darmstadt
eingereichten Diplomarbeit

14.10.2009

Betreuer der Arbeit:

Prof. Dr. Andreas Pfeifer
Hochschule Darmstadt
Fachbereich Mathematik und Naturwissenschaften
Holzhofallee 38
D - 64295 Darmstadt

Dr. Christian v. Malottki
Institut Wohnen und Umwelt GmbH
Annastraße 15
64825 Darmstadt

ISBN 978-3-941140-10-3

IWU-Bestellnr. 06/09

Vorwort der Betreuer

Flächendeckende Daten über tatsächlich gezahlte Wohnungsmieten, d.h. die vertraglich vereinbarten Mieten, liegen in Deutschland nicht vor. Dies gilt umso mehr für Bestandsmietverträge, d.h. Verträge, die schon vor einiger Zeit abgeschlossen wurden und nun der Regulierung der Miethöhe nach dem BGB unterliegen. Für Neuvertragsmieten und Bestandsmieten ist die Wohnungsmarktanalyse deshalb auf zwei wesentliche Quellen angewiesen: Die Angaben des Immobilienverbands Deutschland (IVD), einem deutschlandweiten Zusammenschluss von Maklern, und die Auswertung von Mietspiegeln, die anhand empirischer Erhebungen oder Verhandlungen zwischen Vermieter- und Mieterverbänden die ortsübliche Vergleichsmiete bestimmen. In beiden Fällen liegen allerdings nur für ca. 300 bis 400 Städte und Gemeinden Werte vor.

In der vorliegenden Diplomarbeit, die Olivia Denise Heuer am Institut Wohnen und Umwelt schrieb und am Fachbereich Mathematik der Hochschule Darmstadt einreichte, wird deshalb eine regressionsanalytische Schätzung der Miethöhe in allen anderen Städten und Gemeinden vorgenommen. Erklärende Variablen sind verschiedene regionalstatistische Parameter, welche die regionale Struktur, Wirtschaftskraft und Attraktivität von Städten und Gemeinden beschreiben. Da ein hohes Maß an Korrelation zwischen diesen Variablen zu befürchten war, hat Olivia Denise Heuer parallel als Schätzgleichungen eine herkömmliche lineare Regression und eine Faktoranalyse gerechnet. Die Ergebnisse für die lineare Regression sind sowohl beim Mietspiegeldatensatz als auch beim IVD-Datensatz sehr plausibel, so dass die Arbeit Ergebnisse erbringt, die für die Wohnungsmarktanalyse von praktischem Wert sind. Durch die zunehmende Digitalisierung von Wohnungsannoncen, d.h. die Erfassung von Angebotsmieten, ergeben sich weitere Forschungsansätze, in denen bspw. die Vertragsmieten als Funktion der bekannten Angebotsmieten modelliert werden.

Darmstadt, im Oktober 2009

Prof. Dr. Andreas Pfeifer
Hochschule Darmstadt

Dr. Christian v. Malottki
Institut Wohnen und Umwelt

Vorwort der Autorin

Die vorliegende Diplomarbeit wurde in Zusammenarbeit mit dem Institut Wohnen und Umwelt, Darmstadt, erstellt. Ich möchte besonders Herrn Dr.-Ing. Christian von Malottki für die Betreuung meiner Diplomarbeit und die Unterstützung bezüglich deren Ausarbeitung danken.

Mein Dank gilt außerdem Herrn Prof. Dr. Andreas Pfeifer, Hochschule Darmstadt, für die Betreuung meiner Arbeit.

Ebenso bin ich Frau Galina Nuss, Institut Wohnen und Umwelt, für fachliche Betreuung, Anregungen und Verbesserungsvorschläge zu großem Dank verpflichtet.

Mein besonderer Dank gilt außerdem Frau Inge Schwarz für das Korrekturlesen meiner Arbeit.

Abschließend möchte ich noch Herrn Alexander Schwarz von ganzem Herzen für seine Unterstützung und Rücksichtnahme danken.

Darmstadt, August 2009

Olivia Denise Heuer

Inhalt

Abbildungsverzeichnis **9**

Tabellenverzeichnis **11**

Abkürzungsverzeichnis **13**

Symbolverzeichnis **14**

1 Einleitung **17**

1.1 Problemstellung und Untersuchungsgegenstand 17

1.2 Zielsetzung und Vorgehensweise 17

2 Faktorenanalyse **19**

2.1 Einführung 19

2.2 Idee der Hauptkomponentenanalyse 20

2.3 Standardisierung der Ausgangsdatenmatrix 21

2.4 Korrelationsanalyse 22

2.4.1 Der Bravais-Pearson-Korrelationskoeffizient r 22

2.4.2 Der Phi-Koeffizient ϕ 23

2.5 Variablenauswahl 26

2.5.1 Signifikanzniveau der Korrelation 26

2.5.2 Bartlett-Test (test of sphericity) 26

2.5.3 Kaiser-Meyer-Olkin-Maß 27

2.5.4 Image-Analyse von Guttman 28

2.6 Entwicklung des Modells 29

2.6.1 Das Fundamentaltheorem der Faktorenanalyse 30

2.6.2 Grafische Interpretation der Faktoren 31

2.7 Bestimmung der Anzahl der zu extrahierenden Faktoren 33

2.7.1 Bestimmung der Kommunalitäten 34

2.7.2 Berechnung der Faktorladungsmatrix A 34

2.7.3 Reproduzierte Korrelationsmatrix $\hat{R}$ 36

2.7.4 Kaiser-Kriterium 36

2.7.5 Scree-Test 36

2.8 Faktorinterpretation 37

2.8.1 Orthogonale Rotation 38

2.9 Bestimmung der Faktorwerte 39

2.10 Beispiel einer Faktorenanalyse mit SPSS 40

3 Lineare Regressionsanalyse 47
3.1 Einführung 47
3.2 Das lineare Regressionsmodell 47
3.3 Parameterschätzungen 49
3.3.1 Schätzung der Regressionskoeffizienten 50
3.3.2 Schätzung der Störgrößenvarianz 51
3.4 Überprüfung der Modellgüte 51
3.4.1 Bestimmtheitsmaß und korrigiertes Bestimmtheitsmaß 52
3.4.2 Test auf Signifikanz des Erklärungsansatzes 53
3.4.3 Standardfehler der Schätzung 53
3.5 Überprüfung der Regressionskoeffizienten 53
3.5.1 T-Test 54
3.5.2 Konfidenzintervalle 54
3.5.3 Plausibilitätskontrolle 54
3.6 Überprüfung der Modellannahmen 55
3.6.1 Multikollinearität 55
3.6.2 Normalverteilung der Residuen 56
3.6.3 Heteroskedastizität 56
3.6.4 Autokorrelation 56
3.7 Beispiel einer Regressionsanalyse mit SPSS 57
4 Datenerhebung 65
4.1 Mietpreisinformationen durch Auswertung von Mietspiegeln 65
4.1.1 Standardwohnungstypen 65
4.1.2 Mietspiegelauswertung 69
4.2 Exkurs: Bedeutung und Arten von Mietspiegeln 70
4.2.1 Aufgaben des Mietspiegels 70
4.2.2 Arten und Verbreitung von Mietspiegeln 70
4.2.3 Tabellen- und Regressionsmietspiegel 71
4.2.4 Mietdatenbank 75
4.3 IVD – Datensatz 76
4.4 Erklärende Variablen 77
4.4.1 Kurzbeschreibung der Variablen 77
5 Untersuchung des Mietspiegel-Datensatzes 82
5.1 Übersicht über den Datensatz 82
5.2 Abgeleitete Variablen 83

5.2.1 Bildung von Dummy-Variablen ... 83
5.2.2 Bildung von Referenzkategorien ... 84
5.3 Erstes Überprüfen auf Multikollinearität ... 85
5.4 Analyse verschiedener Modellvarianten ... 86
5.5 Regressionsanalyse für Modell 1 ... 87
5.5.1 Festlegung des Signifikanzniveau α ... 87
5.5.2 Überprüfung der Voraussetzungen ... 87
5.5.3 Überprüfung der Modellgüte ... 90
5.5.4 Überprüfung der Koeffizienten ... 92
5.5.5 Überprüfung auf räumliche Korrelation ... 93
5.6 Faktorenanalyse für Modell 2 ... 95
5.6.1 Korrelationsanalyse ... 95
5.6.2 Variablenauswahl ... 96
5.6.3 Bestimmung der Anzahl der zu extrahierenden Faktoren ... 101
5.6.4 Auswertung der Ergebnisse ... 104
5.6.5 Überprüfung der Merkmale anhand einer 50%-Zufallsstichprobe ... 109
5.7 Regressionsanalyse für Modell 2 ... 110
5.7.1 Überprüfung der Voraussetzungen ... 111
5.7.2 Überprüfung der Modellgüte ... 114
5.7.3 Überprüfung der Koeffizienten ... 115
5.7.4 Überprüfung auf räumliche Korrelation ... 116
5.8 Überprüfung der Modellstrukturen anhand einer 50%-Zufallsstichprobe ... 118
5.8.1 Vergleich der Voraussetzungen ... 118
5.8.2 Vergleich der Modellgüte ... 118
5.8.3 Vergleich der Koeffizienten ... 118
6 Untersuchung des IVD-Datensatzes ... 120
6.1 Vergleich IVD-Daten mit Mietspiegel-Daten ... 120
6.2 Übersicht über den Datensatz ... 122
6.3 Abgeleitete Variablen ... 123
6.3.1 Bildung von Dummy-Variablen ... 123
6.3.2 Bildung von Referenzkategorien ... 123
6.4 Analyse verschiedener Modellvarianten ... 124
6.5 Regressionsanalyse für Modell 1 ... 124
6.5.1 Überprüfung der Voraussetzungen ... 124
6.5.2 Überprüfung der Modellgüte ... 129

6.5.3 Überprüfung der Koeffizienten ... 130
6.5.4 Überprüfung auf räumliche Korrelation ... 132
6.6 Faktorenanalyse für Modell 2 ... 134
6.6.1 Korrelationsanalyse ... 134
6.6.2 Variablenauswahl ... 136
6.6.3 Bestimmung der Anzahl der zu extrahierenden Faktoren ... 142
6.6.4 Auswertung der Ergebnisse ... 144
6.6.5 Überprüfung der Merkmale anhand einer 50%-Zufallsstichprobe ... 149
6.7 Regressionsanalyse für Modell 2 ... 150
6.7.1 Überprüfung der Voraussetzungen ... 150
6.7.2 Überprüfung der Modellgüte ... 153
6.7.3 Überprüfung der Koeffizienten ... 154
6.7.4 Überprüfung auf räumliche Korrelation ... 155
6.8 Überprüfung der Modellstrukturen anhand einer 50%-Zufallsstichprobe ... 157
6.8.1 Vergleich der Voraussetzungen ... 157
6.8.2 Vergleich der Modellgüte ... 157
6.8.3 Vergleich der Koeffizienten ... 157
7 Auswertung der Untersuchungsergebnisse ... 159
7.1 Auswertung der Ergebnisse des Mietspiegel-Datensatzes ... 159
7.2 Auswertung der Ergebnisse des IVD-Datensatzes ... 164
8 Zusammenfassung und Ausblick ... 168
9 Literaturverzeichnis ... 170
9.1 Bücher ... 170
9.2 Internetquellen ... 171
Anhang ... 173
Mietspiegelverbreitung nach Erstellungsform ... 173
Grafische Darstellung der erklärenden Variablen ... 174
Variablenliste 1: Datei Mietspiegel.sav ... 185
Variablenliste 2: Datei IVD.sav ... 188

Abbildungsverzeichnis

Abbildung 1: Idee der Hauptkomponentenanalyse 20
Abbildung 2: Vektordarstellung einer Korrelation 31
Abbildung 3: Zwei-Variablen-Zwei-Faktor-Lösung (vgl. BACK S.287) 32
Abbildung 4: Beispiel eines Scree-Plots 37
Abbildung 5: Auswertung des Scree-Tests 44
Abbildung 6: Streudiagramm mit Regressionsgerade 49
Abbildung 7: Streudiagramm für Linearitätstest 58
Abbildung 8: Streudigramm für Test auf Varianzungleichheit und Ausreißer (1) 59
Abbildung 9: Streudigramm für Test auf Varianzungleichheit und Ausreißer (2) 60
Abbildung 10: Histogramm und P-P-Plot der standardisierten Residuen 63
Abbildung 11: Streudiagramme 64
Abbildung 12: Vereinfachte Darstellung Regressionsmietspiegel - Basistabelle 74
Abbildung 13: Zu- und Abschlagsmerkmale 75
Abbildung 14: Ausschnitt aus dem IVD-Wohn-Preisspiegel 2006/2007 76
Abbildung 15: Häufigkeitsverteilungen der siedlungsstrukturellen Gemeindetypen 84
Abbildung 16: Bivariate Streudiagramme 86
Abbildung 17: Histogramm der standardisierten Residuen 88
Abbildung 18: PP-Plot der standardisierten Residuen 89
Abbildung 19: Streudiagramm für den Test auf Varianzgleichheit 90
Abbildung 20: Darstellung der Residuen aus dem Mietspiegel-Modell 1 94
Abbildung 21: Scree-Plot und eingezeichnetes Kaiser-Kriterium 103
Abbildung 22: Komponentendiagramm im rotierten Raum 108
Abbildung 23: Histogramm und PP-Plot der standardisierten Residuen 112
Abbildung 24: Streudiagramm für den Test auf Varianzgleichheit 113
Abbildung 25: Darstellung der Residuen aus dem Mietspiegel-Modell 2 117
Abbildung 26: Histogramm der standardisierten Residuen (1. und 2. Teilstichprobe) 118
Abbildung 27: Differenz Mietspiegel-Mieten und IVD-Mieten 121
Abbildung 28: Häufigkeitsverteilung der siedlungsstrukturellen Gemeindetypen 123
Abbildung 29: Histogramm und PP-Plot der standardisierten Residuen 126
Abbildung 30: Histogramm und PP-Plot der standardisierten Residuen 128
Abbildung 31: Streudiagramm für den Test auf Varianzgleichheit 129
Abbildung 32: Darstellung der Residuen aus dem IVD-Modell 1 133
Abbildung 33: Scree-Plot mit eingezeichnetem Kaiser-Kriterium 143

Abbildung 34: Komponentendiagramm im rotierten Raum 147
Abbildung 35: Histogramm der standardisierten Residuen 151
Abbildung 36: Streudiagramm für den Test auf Varianzgleichheit 152
Abbildung 37: Darstellung der Residuen aus dem IVD-Modell 2 156
Abbildung 38: Histogramm der standardisierten Residuen (1. und 2. Teilstichprobe) 157
Abbildung 39: Mietpreise in Deutschland laut Mietspiegel - Modell 1 161
Abbildung 40: Mietpreise in Deutschland laut Mietspiegel-Modell 2 162
Abbildung 41: Differenz Mieten Mietspiegel - Modell 1 und 2 163
Abbildung 42: Mietpreise in Deutschland laut IVD-Modell 1 165
Abbildung 43: Mietpreise in Deutschland laut IVD-Modell 2 166
Abbildung 44: Differenz Mieten IVD-Modell 1 und 2 167
Abbildung 45: Mietspiegelverbreitung 173
Abbildung 46: Einwohnerzahl (Stand: 31.12.2006) 174
Abbildung 47: Einwohnerzahl pro Haushalt (Stand 31.12.2006) 175
Abbildung 48: Einwohner je km² (Stand: 31.12.2006) 176
Abbildung 49: Einwohner in 30km Umkreis (Stand 31.12.2006) 177
Abbildung 50: Sozialversicherungspflichtig Beschäftigte in 30km Umkreis (Stand 30.06.2006) 178
Abbildung 51: Anteil Wohnungen in Einfamilienhäusern (Stand: 31.12.2006) 179
Abbildung 52: Anteil Wohnungen in Zweifamilienhäusern (Stand: 31.12.2006) 180
Abbildung 53: Anteil Wohnungen in Mehrfamilienhäusern (Stand: 31.12.2006) 181
Abbildung 54: Kaufkraftindex (Stand: 31.12.2006) 182
Abbildung 55: Arbeitslosenquote (Stand: 30.06.2006) 183
Abbildung 56: Siedlungsstrukturelle Gemeindetypen 184

Tabellenverzeichnis

Tabelle 1: Beurteilung des Ergebnis des KMO-Maß nach Kaiser 28
Tabelle 2: SPSS-Ausgabe der deskriptiven Statistiken 40
Tabelle 3: SPSS-Tabelle der Korrelationen und ihrer Signifikanzniveaus 41
Tabelle 4: SPSS-Ausgabe des KMO-Maßes und des Bartlett-Tests 42
Tabelle 5: Ausgabe der Anti-Image-Korrelationsmatrix 42
Tabelle 6: SPSS-Tabelle der Kommunalitäten 43
Tabelle 7: SPSS-Ausgabe der erklärten Gesamtvarianz 43
Tabelle 8: Faktorwerte der extrahierten Faktoren 44
Tabelle 9: Reproduzierte Korrelationsmatrix 45
Tabelle 10: Ausgabe der rotierten Faktorladungen 45
Tabelle 11: Koeffizientenmatrix der Komponentenwerte 46
Tabelle 12: SPSS-Ausgabe der deskriptiven Statistiken 60
Tabelle 13: SPSS-Tabelle der Korrelationen 60
Tabelle 14: SPSS-Output der Modellbeurteilung 61
Tabelle 15: SPSS-Ausgabe der Varianzanalyse 61
Tabelle 16: SPSS-Output der Koeffizientenstatistik 62
Tabelle 17: SPSS-Ausgabe der Residuenstatistik 62
Tabelle 18: Häufigkeitstabelle Wohnungstypen 66
Tabelle 19: Kreuztabelle - Standardwohnung nach Baualter des Gebäudes und Lage 67
Tabelle 20: Mittlere Wohnflächen 68
Tabelle 21: Überblick Standardwohnungstypen 69
Tabelle 22: Vereinfachte Darstellung Tabellenmietspiegel 72
Tabelle 23: Ausschnitt SPSS-Output der Koeffizientenstatistik 73
Tabelle 24: Siedlungsstrukturelle Gemeindetypen nach der BBR-Systematik (eigene Darstellung) 81
Tabelle 25: Deskriptive Statistik der Variablen 83
Tabelle 26: Korrelationsanalyse 85
Tabelle 27: Kollinearitätsstatistik 88
Tabelle 28: SPSS-Output der Modellbeurteilung 91
Tabelle 29: SPSS-Ausgabe der Varianzanalyse 91
Tabelle 30: SPSS-Output der Koeffizientenstatistik 92
Tabelle 31: SPSS- Output der Korrelationsmatrix 96
Tabelle 32: Signifikanzniveau der Korrelation 97
Tabelle 33: KMO-Maß und Bartlett-Test 98

Tabelle 34: Anti-Image-Korrelationsmatrix 99
Tabelle 35: Anti-Image-Korrelationsmatrix 100
Tabelle 36: Anti-Image-Kovarianzmatrix 101
Tabelle 37: SPSS-Output der anfänglichen und endgültigen Kommunalitäten 102
Tabelle 38: SPSS-Ausgabe der erklärten Gesamtvarianz 102
Tabelle 39: SPSS-Ausgabe der reproduzierten Korrelationsmatrix und Residualmatrix. 105
Tabelle 40: Rotierte Komponentenmatrix 106
Tabelle 41: Koeffizientenmatrix der Komponentenwerte 109
Tabelle 42: KMO-Maß und Bartlett-Test der ersten 50%-Zufallsstichprobe 110
Tabelle 43: KMO-Maß der zweiten 50%-Zufallsstichprobe 110
Tabelle 44: Kollinearitätsstatistik 111
Tabelle 45: SPSS-Output der Modellbeurteilung 114
Tabelle 46: SPSS-Ausgabe der Varianzanalyse 114
Tabelle 47: SPSS-Output der Koeffizientenstatistik 115
Tabelle 48: Deskriptive Statistik der Variablen 122
Tabelle 49: Kollinearitätsstatistik 125
Tabelle 50: Kollinearitätsstatistik 127
Tabelle 51: SPSS-Output der Modellbeurteilung 130
Tabelle 52: SPSS-Ausgabe der Varianzanalyse 130
Tabelle 53: SPSS-Output der Koeffizientenstatistik 131
Tabelle 54: SPSS-Output der Korrelationsmatrix 135
Tabelle 55: Signifikanzniveau der Korrelation 137
Tabelle 56: KMO-Maß und Bartlett-Test 138
Tabelle 57: Anti-Image-Korrelationsmatrix 139
Tabelle 58: Anti-Image-Korrelationsmatrix 140
Tabelle 59: Anti-Image-Kovarianzmatrix 141
Tabelle 60: SPSS-Output der anfänglichen und endgültigen Kommunalitäten 142
Tabelle 61: SPSS-Ausgabe der erklärten Gesamtvarianz 142
Tabelle 62: Reproduzierte Korrelationsmatrix und Residualmatrix 145
Tabelle 63: Rotierte Komponentenmatrix 146
Tabelle 64: Koeffizientenmatrix der Komponentenwerte 148
Tabelle 65: KMO-Maß und Bartlett-Test der ersten 50%-Zufallsstichprobe 149
Tabelle 66: KMO-Maß und Bartlett-Test der zweiten 50%-Zufallsstichprobe 149
Tabelle 67: Kollinearitätsstatistik 150
Tabelle 68: SPSS-Output der Modellbeurteilung 153
Tabelle 69: SPSS-Ausgabe der Varianzanalyse 153
Tabelle 70: SPSS-Output der Koeffizientenstatistik 154

Abkürzungsverzeichnis

BA	Bundesagentur für Arbeit
BBR	Bundesamt für Bauwesen und Raumordnung
BGB	Bürgerliches Gesetzbuch
CFA	Konfirmatorische Faktorenanalyse
EFA	Explorative Faktorenanalyse
EFH	Einfamilienhaus
GeB	geringfügig Beschäftigte
GfK	Gesellschaft für Konsumforschung
HKA	Hauptkomponentenanalyse
IVD	Immobilienverband Deutschland Bundesverband
IWU	Institut für Wohnen und Umwelt
MFH	Mehrfamilienhaus
nmqm	monatliche Nettomiete in € pro m²
SVB	sozialversicherungspflichtig Beschäftigte
ZFH	Zweifamilienhaus

Symbolverzeichnis

Kapitel 2: Faktorenanalyse

r_{x_1,x_2}	Korrelationskoeffizient der Variablen x_1 und x_2
x_{t1}	Ausprägung der Variablen x_1 für Beobachtung t, $t = 1,...,T$
$\bar{x}_1$	Mittelwert der Ausprägung der Variablen x_1 über alle Beobachtungen
z_{tk}	Standardisierter Beobachtungswert, Element der standardisierten Ausgangsdatenmatrix
Z	Standardisierte Ausgangsdatenmatrix
s_{x_1,x_2}	Kovarianz der Variablen x_1 und x_2
s_{x_t}	Standardabweichung der Variablen x_t
T	Beobachtungsumfang
K	Anzahl der verwendeten Variablen
ϕ	Phi-Korrelationskoeffizient
χ^2	Chi-Quadrat-Verteilung
R	Korrelationsmatrix
PG	Prüfgröße des Bartlett-Tests auf Sphärizität
$\ln$	natürlicher Logarithmus
KMO	Kaiser-Meyer-Olkin-Maß
MSA_i	Kaiser-Meyer-Olkin-Maß der Variablen i, measure for sampling adequacy
a_{kq}	Faktorladung des q-ten Faktors der k-ten Variable, $k = 1,...,K$
Q	Anzahl der Faktoren
A	Faktorladungsmatrix
F	Matrix der Faktorwerte
A'	Transponierte der Matrix A
A^{-1}	Inverse der Matrix A
$\overline{OA}$	Länge der Strecke von Punkt O bis Punkt A
h_k^2	Kommunalität
U	Residualmatrix
M	Rotationsmatrix

$\tilde{A}$	rotierte Faktorladungsmatrix A
$\tilde{a}_{kq}$	Elemente der rotierten Faktorladungsmatrix
$\hat{R}$	reproduzierte Korrelationsmatrix
$r_{xy.z}$	partielle Korrelation zwischen den Variablen x und y ohne den Einfluss der Variablen z
$r^2_{ij.2...k}$	partielle Korrelation zwischen den Variablen i und j
f_q	Faktor q, $q = 1,...,Q$
$s^2_{f_q}$	Varianz des Faktors q
E	Einheitsmatrix
B	Matrix der Regressionskoeffizienten
e	Residuen
G	Kovarianzmatrix des Image
P	Kovarianzmatrix des Anti-Image
D^2	Matrix der Varianzen der Residuen
$\hat{Z}$	Schätzer von Z

Kapitel 3: Regressionsanalyse

y_t	abhängige Variable (Regressand, Prognosevariable)
x_t	unabhängige Variable (Regressor, Prädiktorvariable), $t = 1,...,T$
β_1	Konstante
β_k	Regressionskoeffizient der k-ten abhängigen Variablen, $k = 1,...,K$
u_t	Störgröße
T	Beobachtungsumfang
K	Anzahl erklärender Variablen
e_t	Residuum des Falles t
$\hat{y}_t$	Schätzer für y_t
E	Einheitsmatrix
$SST = \sum_t (y_t - \bar{y})^2$	Gesamtstreuung, Sum of Squares Total
$SSR = \sum_t (\hat{y}_t - \bar{y})^2$	erklärte Streuung, Sum of Squares Regression
$SSE = \sum_t (y_t - \hat{y}_t)^2$	nicht erklärte Streuung, Sum of Squares Error

R^2	Bestimmtheitsmaß
$\overline{R}^2$	korrigiertes Bestimmtheitsmaß
H_0	Nullhypothese
H_1	Alternativhypothese
F_{emp}	Prüfgröße des F-Tests, empirischer Wert
s_e	Standardschätzfehler
t_{emp}	Prüfgröße des t-Tests, empirischer Wert
s_{e_K}	Standardfehler des k-ten Regressionskoeffizienten
α	Signifikanzniveau
df	Freiheitsgrade
R_k^2	Bestimmtheitsmaß einer Regression von x_k auf die restlichen erklärenden Variablen

1 Einleitung

1.1 Problemstellung und Untersuchungsgegenstand

Dem Mietwohnungsmarkt in Deutschland kommt eine sehr große Bedeutung zu, da knapp 57% der Privathaushalte in einem Mietwohnverhältnis leben (vgl. DESTATIS2, S. 23). Die meisten dieser Mietwohnungen sind frei finanzierte Wohnungen, also Wohnungen, die nicht unter Einsatz öffentlicher Mittel erstellt wurden oder deren Mietpreisbindung in Folge einer öffentlichen Förderung inzwischen erloschen ist. Für diese Wohnungen gilt das Prinzip der ortsüblichen Vergleichsmiete.

Auf das Mietpreisniveau haben viele Faktoren Einfluss. So ist z.B. die Größe einer Wohnung eine wichtige Determinante der Miethöhe. Bezogen auf die Quadratmetermiete besteht hier eine negative Korrelation (je größer die Wohnung, desto niedriger ist die Miete pro m²)[1] und im Bezug auf die absolute Miethöhe eine positive Korrelation (je größer die Wohnung, desto höher fällt auch die monatliche Miete aus). Neben der Wohnungsgröße existieren noch andere Merkmale (z.B. das Baualter, die Dauer des Mietverhältnisses), die den Mietpreis beeinflussen.

Doch diese Faktoren erklären nicht die starke Variation des Mietpreisniveaus in Deutschland. Es bestehen nicht nur deutliche Unterschiede zwischen städtischen und ländlich geprägten Regionen, auch die Städte untereinander unterscheiden sich hinsichtlich der Kosten, die man für die Miete im Monat veranschlagen muss, erheblich. So lässt sich auf dem deutschen Mietwohnungsmarkt ein Süd-Nord-Gefälle und ein West-Ost-Gefälle feststellen: Wohnungen in München sind beispielsweise wesentlich teurer als in Hamburg und Wohnungen im Osten Deutschlands sind oft bedeutend günstiger als vergleichbare Wohnungen in Westdeutschland. Um diese Strukturen zu erfassen, braucht man andere Faktoren zur Erklärung.

In dieser Diplomarbeit soll nun der Einfluss verschiedener regionaler Faktoren (z.B. die Einwohnerzahl, die Kaufkraft und die Arbeitslosenquote) auf das Mietpreisniveau und die Mietpreissteigerung untersucht werden. Dies kann – ohne dass damit eine Kausalität nachgewiesen wird – Hinweise auf die Ursachen regionaler Mietpreisdifferenzierung geben.

Da die vorhandenen Preisdatenquellen nicht alle Städte und Gemeinden in Deutschland abdecken, können so mithilfe von Regressionsanalysen Datenbanken mit Schätzwerten auch für Städte und Gemeinden ohne Daten vervollständigt werden.

1.2 Zielsetzung und Vorgehensweise

Ziel dieser Arbeit ist es, einen Querschnitt der Mietpreise und deren Einflussfaktoren für Deutschland zu modellieren und darzustellen.

[1] Dies gilt zumindest für Wohnflächen bis ca. 120m², im darüber beginnenden Luxussegment sind auch höhere Quadratmeterpreise zu beobachten.

Dazu muss zunächst einmal ein umfangreicher Datenbestand erfasst werden, der für die Analyse herangezogen werden soll. Insgesamt lassen sich die folgenden Hauptbezugsquellen als Informationslieferanten identifizieren:

Das Institut für Wohnen und Umwelt (IWU) hat sich eine bundesweite Sammlung von Mietspiegeln (im weiteren Verlauf „Mietspiegel-Datenbank“ genannt) aufgebaut, aus denen Mietpreisinformationen gewonnen werden können. Eine weitere Datenquelle stellt eine detaillierte Preisdatensammlung des Immobilienverbands Deutschlands (IVD) dar, der IVD-Wohn-Preisspiegel. Des Weiteren fließen die Daten des Statistischen Bundesamtes und von anderen Institutionen als erklärende Variable in die Auswertungen mit ein.

Die Struktur dieser Arbeit gliedert sich wie folgt: Zunächst sollen die in dieser Untersuchung angewendeten Verfahren der Faktorenanalyse und der Regressionsanalyse theoretisch erläutert werden. Am Ende der jeweiligen Kapitel 2 und 3 findet sich zu jedem Verfahren ein kleines Beispiel, welches mit Hilfe der Statistik-Software SPSS® 14.0 berechnet wurde und dessen Output beschrieben und erklärt werden soll.

In einem nächsten Schritt werden dann die Datengrundlagen, auf die sich die Diplomarbeit stützen wird, vorgestellt und die zu untersuchenden, regionalen Faktoren bestimmt.

Im Anschluss daran folgen dann in den Kapiteln 5 und 6 die für diese Ausarbeitung relevanten Untersuchungen der Datensätze und die Interpretation der Ergebnisse. Diese werden anschließend in Kapitel 7 mit Hilfe des Geoinformationssystems RegioGraph® 10 grafisch dargestellt.

Das letzte Kapitel gibt dann noch einmal einen Überblick über die aus den Analysen gewonnenen Erkenntnisse und Ergebnisse.

Alle Analysen werden mit der Statistik-Software SPSS® 14.0 durchgeführt.

2 Faktorenanalyse

Multivariate Verfahren werden unter anderem dazu verwendet, in Datensätzen mit zwei oder mehr Variablen, Abhängigkeitsstrukturen zwischen den Variablen zu identifizieren. Man unterscheidet hierbei zwischen strukturprüfenden Verfahren (z.B. Varianz- oder Regressionsanalyse) und strukturentdeckenden Verfahren (z.B. Faktoren- oder Clusteranalyse).

Im folgenden Abschnitt soll nun das Verfahren der Faktorenanalyse vorgestellt werden. Dabei wird kurz darauf eingegangen, welche Ansätze bei dieser Methode zum Einsatz kommen und das allgemeine Modell beschrieben. Bevor dann die Ergebnisse der Faktorenanalyse ausgewertet und interpretiert werden können, muss zunächst noch die Überprüfung der Modellgüte erfolgen.

2.1 Einführung

Die Faktorenanalyse gehört zu den strukturentdeckenden Verfahren und hat als Ziel, die korrelativen Zusammenhänge zwischen einer Anzahl von gegebenen Variablen durch eine geringere Anzahl von nicht beobachtbaren (latenten) Faktoren möglichst vollständig zu erklären. Mit dem Verfahren lässt sich ohne entscheidenden Informationsverlust eine Datenreduktion erreichen und es kommt zu einer Vereinfachung bei der Auswertung, da stark miteinander korrelierende Variablen zu Faktoren zusammengefasst werden können und diese dann z.B. als neue unabhängige Variablen in ein Regressionsmodell einfließen.

Grundsätzlich lässt sich die Faktorenanalyse in zwei unterschiedliche Vorgehensweisen aufteilen:

- Explorative Faktorenanalyse (EFA)

  Die EFA wird eingesetzt, wenn a priori keinerlei Hypothesen über die Zusammenhänge zwischen den Variablen bzw. keine Annahmen über die Anzahl der benötigten Faktoren aufgestellt wurden. Dies geschieht erst nach Durchführung der Analyse. Es handelt sich also um ein hypothesengenerierendes Verfahren.

- Konfirmatorische Faktorenanalyse (CFA)

  Die CFA wird als hypothesenprüfendes Verfahren angewendet. In der Regel wird a priori ein (oder mehrere) Faktormodell(e) aufgestellt und mit Hilfe der CFA überprüft, ob die theoretischen Modellannahmen durch die Daten gestützt werden.

In Rahmen dieser Diplomarbeit wird der explorative Typ verwendet. Als Extraktionsmethode soll hier nur die Hauptkomponentenanalyse (engl. Principal Components Analysis, PCA) betrachtet werden, da nicht nach den hinter den Variablen stehenden Größen bzw. Ursachen gesucht wird, sondern nach einem Sammelbegriff (Komponente), der die zusammengefassten Variablen beschreibt[2].

[2] Vgl. BACK S. 293

2.2 Idee der Hauptkomponentenanalyse

Das Ziel einer Hauptkomponentenanalyse ist ebenfalls, eine Vielzahl von Variablen auf einige wenige Komponenten zu reduzieren, die jedoch möglichst viel Information der Originalvariablen enthalten (Dimensionsreduktion).

Dazu werden Linearkombinationen (Hauptkomponenten, HK) der beobachteten Variablen konstruiert, die sukzessive einen sinkenden Prozentsatz der Datenvariabilität erklären. Die erste Hauptkomponente ist die Linearkombination, die einen maximalen Anteil der Varianz erklärt. Jede folgende Hauptkomponente ist eine Linearkombination, die ein Maximum der Restvarianz erklärt und zu den bereits definierten Hauptkomponenten orthogonal, also nicht korreliert, ist. Die weniger relevanten Hauptkomponenten können oft ohne wesentlichen Verlust an Information weggelassen werden (sie beschreiben das Rauschen).

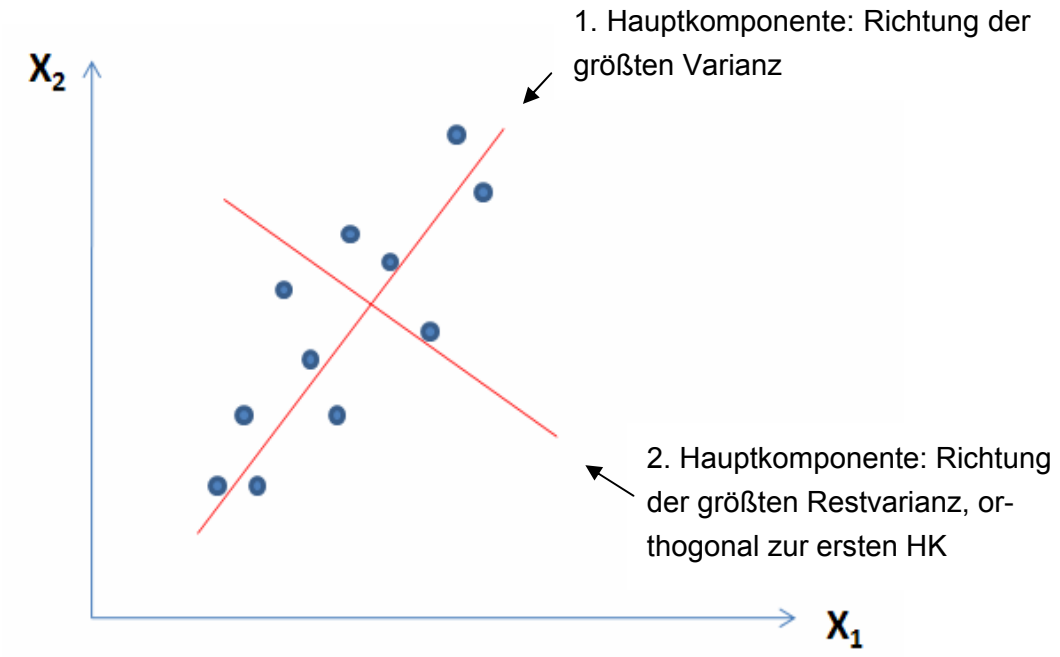


Abbildung 1: Idee der Hauptkomponentenanalyse[3]

Damit erfolgt eine „Umverteilung“ großer Teile der Gesamtvarianz auf wenige Hauptkomponenten.

Die Durchführung einer Faktorenanalyse lässt sich in sieben wesentliche Schritte unterteilen, die im Folgenden näher erläutert werden sollen[4]:

- Standardisierung der Ausgangsdatenmatrix
- Korrelationsanalyse
- Variablenauswahl

[3] Quelle: Eigene Darstellung

[4] Vgl. BACK S. 268

- Entwicklung des Modells
- Bestimmung der Anzahl der zu extrahierenden Faktoren
- Faktorinterpretation
- Bestimmung der Faktorwerte

2.3 Standardisierung der Ausgangsdatenmatrix

Als erster Schritt bei der Durchführung einer Faktorenanalyse empfiehlt es sich, die Ausgangsdatenmatrix zu standardisieren. Dadurch lassen sich die durchzuführenden Berechnungen vereinfachen und die Ergebnisse werden besser interpretierbar.

Dazu wird von dem jeweiligen Beobachtungswert einer Variablen das Stichprobenmittel (arithmetisches Mittel) des jeweiligen Merkmals abgezogen und die Differenz anschließend durch die Standardabweichung dividiert. Formal ausgedrückt bedeutet dies:

$$z_{tk} = \frac{x_{tk} - \bar{x}_k}{s_k}$$

Die standardisierte Datenmatrix hat dann die Form:

$$Z = \begin{pmatrix} z_{11} & \cdots & z_{1K} \\ \vdots & \ddots & \vdots \\ z_{T1} & \cdots & z_{TK} \end{pmatrix}$$

Die Anzahl an erklärenden Variablen beträgt K und es liegen Beobachtungen für T Zeitpunkte vor. Durch die Standardisierung werden die Mittelwerte und Standardabweichungen der Variablen auf Null und Eins normiert. Das hat zur Folge, dass die Kovarianzmatrix und die Korrelationsmatrix identisch sind:

$$r_{x_1,x_2} = \frac{s_{x_1,x_2}}{s_{x_1} \cdot s_{x_2}} \quad \text{mit } s_{x_1,x_2} = \frac{1}{T-1}\sum_{t=1}^{T}(x_{t1} - \bar{x}_1) \cdot (x_{t2} - \bar{x}_2)$$

Durch die Standardisierung gilt $s_{x_1} = 1, s_{x_2} = 1$ und somit $r_{x_1,x_2} = s_{x_1,x_2}$ [5]. Auch lässt sich der Korrelationskoeffizient einfach aus den standardisierten Variablen berechnen:

$$\begin{aligned} r_{x_1,x_2} &= \frac{\frac{1}{T-1}\sum_{t=1}^{T}(x_{t1} - \bar{x}_1) \cdot (x_{t2} - \bar{x}_2)}{s_{x_1} \cdot s_{x_2}} \\ &= \frac{1}{T-1}\sum_{t=1}^{T}\frac{(x_{t1} - \bar{x}_1) \cdot (x_{t2} - \bar{x}_2)}{s_{x_1} \cdot s_{x_2}} \\ &= \frac{1}{T-1}\sum_{t=1}^{T}\frac{(x_{t1} - \bar{x}_1)}{s_{x_1}} \cdot \frac{(x_{t2} - \bar{x}_2)}{s_{x_2}} \end{aligned}$$

[5] Vgl. BACK S. 271ff.

$$= \frac{1}{T-1} \sum_{t=1}^{T} z_{t1} \cdot z_{t2}$$

In Matrixnotation sieht die Gleichung folgendermaßen aus:

$$R = \frac{1}{T-1} \cdot Z' \cdot Z$$

2.4 Korrelationsanalyse[6]

Im ersten Schritt wird mit Hilfe der Korrelationsmatrix die Abhängigkeit der Vielzahl von Variablen untereinander untersucht. Das ist nötig, um herauszufinden, ob sich der Datensatz überhaupt für eine Faktorenanalyse eignet. Sollten die berechneten Korrelationskoeffizienten nur sehr geringe absolute Werte aufweisen, wäre eine Faktorenanalyse wenig sinnvoll, da sich gemeinsame Faktoren nur für solche Variablen finden lassen, die relativ stark miteinander korrelieren. Einzelne Variablen, die mit den übrigen Variablen nur sehr gering korrelieren, sollten möglicherweise unberücksichtigt bleiben. Für die Bestimmung des Zusammenhangs zwischen den Variablen existieren verschiedene Korrelations- und Assoziationsmaße in Abhängigkeit des Skalenniveaus der Merkmale. Hier soll sich auf die folgenden zwei Maße beschränkt werden:

- Bravais-Pearson-Korrelationskoeffizient r
- Phi-Koeffizient ϕ

2.4.1 Der Bravais-Pearson-Korrelationskoeffizient r

Ein Maß für die Stärke des linearen Zusammenhangs zwischen zwei Variablen ist der Bravais-Pearson-Korrelationskoeffizient r . Er wird bestimmt durch

$$r_{x_1,x_2} = \frac{s_{x_1,x_2}}{s_{x_1} \cdot s_{x_2}} = \frac{\sum_{t=1}^{T} (x_{t1} - \bar{x}_1) \cdot (x_{t2} - \bar{x}_2)}{\sqrt{\sum_{t=1}^{T} (x_{t1} - \bar{x}_1)^2 \cdot \sum_{t=1}^{T} (x_{t2} - \bar{x}_2)^2}}, \quad -1 \le r \le 1$$

wobei

$$s_{x_1,x_2} = \frac{1}{T-1} \sum_{t=1}^{T} (x_{t1} - \bar{x}_1) \cdot (x_{t2} - \bar{x}_2)$$

die empirische Kovarianz bezeichnet und

$$s_{x_1} = \sqrt{\frac{1}{T-1} \sum_{t=1}^{T} (x_{t1} - \bar{x}_1)^2} \qquad s_{x_2} = \sqrt{\frac{1}{T-1} \sum_{t=1}^{T} (x_{t2} - \bar{x}_2)^2}$$

für die Standardabweichungen der Merkmale x_1 und x_2 stehen.

[6] Vgl. BACK S. 270ff. und WIRTZ S. 155

Der Korrelationskoeffizient kann Werte zwischen -1 und 1 annehmen. Das Vorzeichen gibt dabei die Richtung des Zusammenhangs („+“: gleichläufig, „-“: gegenläufig) an und der Absolutwert die Stärke. Werte um +/- 1 drücken dabei einen perfekten linearen Zusammenhang aus. Im Falle von $r = 0$ sind die beiden Merkmale unkorreliert, d.h. zwischen ihnen besteht kein linearer Zusammenhang. Der Bravais-Pearson-Korrelationskoeffizient lässt sich alternativ auch darstellen als:

$$r_{x_1,x_2} = \frac{\sum_{t=1}^{T} x_{t1} \cdot x_{t2} - T \cdot \bar{x}_1 \cdot \bar{x}_2}{\sqrt{(\sum_{t=1}^{T} {x_{t1}}^2 - T \cdot \bar{x}_1^2) \cdot (\sum_{t=1}^{T} {x_{t2}}^2 - T \cdot \bar{x}_2^2)}}$$

Beweisskizze:

- Herleitung Zähler:

$$\sum_{t=1}^{T} (x_{t1} - \bar{x}_1) \cdot (x_{t2} - \bar{x}_2) = \sum_{t=1}^{T} (x_{t1} \cdot x_{t2} - \bar{x}_1 \cdot x_{t2} - \bar{x}_2 \cdot x_{t1} + \bar{x}_1 \cdot \bar{x}_2)$$

$$= \sum_{t=1}^{T} x_{t1} \cdot x_{t2} - \bar{x}_1 \sum_{t=1}^{T} x_{t2} - \bar{x}_2 \sum_{t=1}^{T} x_{t1} + \sum_{t=1}^{T} \bar{x}_1 \cdot \bar{x}_2$$

$$= \sum_{t=1}^{T} x_{t1} \cdot x_{t2} - T \cdot \bar{x}_1 \cdot \bar{x}_2 - T \cdot \bar{x}_1 \cdot \bar{x}_2 + T \cdot \bar{x}_1 \cdot \bar{x}_2$$

$$= \sum_{t=1}^{T} x_{t1} \cdot x_{t2} - T \cdot \bar{x}_1 \cdot \bar{x}_2$$

- Die Herleitung des Nenners erfolgt auf ähnliche Weise und soll hier nicht näher erläutert werden.

2.4.2 Der Phi-Koeffizient ϕ

Der Bravais-Pearson-Korrelationskoeffizient ist grundsätzlich nur für metrische Variablen geeignet. Eine Ausnahme stellen dichotome Variablen dar, die speziell auf die zwei Ausprägungen „0“ und „1“ normiert wurden, d.h. $x_1, x_2 \in \{0,1\}$. Die absoluten Häufigkeiten der vier Kombinationen von $x_1 -$ und $x_2 -$ Werten werden in einer Kontingenztafel dargestellt:

x_1 \ x_2	**1**	**0**	
1	a	b	a + b
0	c	d	c + d
	a + c	b + d	**T**

Tabelle 1: Kontingenztafel für zwei dichotome Merkmale[7]

Für die (0; 1)-kodierten Variablen soll nun der Bravais-Pearson-Koeffizient berechnet werden.

Es gilt dabei:

- $T = a + b + c + d$
- Da die beiden Merkmale dichotom sind, gilt $\sum_{t=1}^{T} x_{t1} = \sum_{t=1}^{T} x_{t1}^{\ 2} = (a+b)$ und analog $\sum_{t=1}^{T} x_{t2} = \sum_{t=1}^{T} x_{t2}^{\ 2} = (a+c)$
- Das Produkt $x_{t1} \cdot x_{t2}$ ist nur dann $\neq 0$, wenn $x_{t1} = 1$ und $x_{t2} = 1$ ist; daraus folgt also $\sum_{t=1}^{T} x_{t1} \cdot x_{t2} = a$
- $\bar{x}_{t1} = \frac{(a+b)}{T}$, $\bar{x}_{t2} = \frac{(a+c)}{T}$ und es folgt daraus $\bar{x}_{t1}^{\ 2} = \frac{(a+b)^2}{T^2}$, $\bar{x}_{t2}^{\ 2} = \frac{(a+c)^2}{T^2}$

Nun werden diese Größen in die Formel für r_{x_1,x_2} eingesetzt:

[7] Quelle: Eigene Darstellung

$$
\begin{aligned}
r_{x_1,x_2} &= \frac{\sum_{t=1}^{T} x_{t1} \cdot x_{t2} - T \cdot \bar{x}_1 \cdot \bar{x}_2}{\sqrt{(\sum_{t=1}^{T} {x_{t1}}^2 - T \cdot \bar{x}_1^{\,2}) \cdot (\sum_{t=1}^{T} {x_{t2}}^2 - T \cdot \bar{x}_2^{\,2})}} \\
&= \frac{a - \frac{(a+b)\cdot(a+c)}{a+b+c+d}}{\sqrt{\left[(a+b) - \frac{(a+b)^2}{a+b+c+d}\right] \cdot \left[(a+c) - \frac{(a+c)^2}{a+b+c+d}\right]}} \\
&= \frac{(a+b+c+d)\cdot a - (a+b)\cdot(a+c)}{\sqrt{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)}} \\
&= \frac{(a\cdot d) - (b\cdot c)}{\sqrt{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)}} \\
&= \phi_{x_1,x_2} \qquad , \quad -1 \le \phi_{x_1,x_2} \le 1
\end{aligned}
$$

Dieses Maß wird als Phi-Koeffizient bezeichnet und liegt, da es aus dem Korrelationskoeffizienten abgeleitet worden ist, ebenfalls zwischen -1 und 1. Es treten genau dann deutliche Abweichungen von Null auf, wenn eine der beiden Diagonalen der Vierfeldertafel deutlich stärker besetzt ist:

- Ist die Diagonale (a, d) stärker besetzt, so ist die Korrelation ϕ positiv.
- Ist die Diagonale (b, c) stärker besetzt, so ist die Korrelation ϕ negativ.

Der Phi-Koeffizient ist also gleich der Produkt-Moment-Korrelation von x_1 und x_2. Dies ist zunächst verwunderlich, da der Phi-Koeffizient nur bei kategorialen dichotomen Merkmalen berechnet werden darf und der Bravais-Pearson-Koeffizient erst ab Intervallskalenniveau definiert ist. Ein dichotomes nominal- oder ordinalskaliertes Merkmal darf aber wie ein intervallskaliertes Merkmal behandelt werden, da das kritische Merkmal für Intervallskalenqualität, nämlich die Gleichabständigkeit der Skalenpunkte, gewährleistet ist[8].

Im Vergleich mit anderen Maßen für Kontingenztabellen zeigt sich, dass der Phi-Koeffizient in einer engen Beziehung mit dem χ^2-Wert steht: er kann direkt aus diesem abgeleitet werden:

$$|\phi| = \sqrt{\frac{\chi^2}{T}} \Leftrightarrow \phi^2 \cdot T = \chi^2 \qquad \text{(ohne Beweis)}$$

Somit können dichotome Merkmale sowohl als intervall- als auch als nominalskalierte Merkmale aufgefasst werden: Die Zusammenhangsmaße für beide Skalenarten r_{x_1,x_2} und χ^2 lassen sich entsprechend ineinander überführen. Der Phi-Koeffizient ist allerdings nur für 2x2-Tabellen definiert.

[8] Vgl. WIRTZ S. 155

2.5 Variablenauswahl

Die für den Datensatz bestimmte Korrelationsmatrix der Form

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1K} \\ r_{21} & 1 & \cdots & \vdots \\ \vdots & \vdots & 1 & \vdots \\ r_{T1} & r_{T2} & \cdots & 1 \end{pmatrix}$$

kann nun zum einen daraufhin überprüft werden, ob sich Variablencluster mit relativ hohen bivariaten Korrelationen finden lassen. Ist dies nicht der Fall, so ist eine sinnvolle Anwendung der Faktorenanalyse in Frage gestellt. Häufig lässt sich aber nur mit der visuellen Überprüfung keine eindeutige Aussage treffen, da z.B. neben vielen kleinen Korrelationen auch einige sehr hohe Werte auftreten können. Daher ist es sinnvoll, noch weitere Kriterien zur Überprüfung Korrelationskoeffizienten auf Eignung zur Faktorenanalyse heranzuziehen[9].

Im weiteren Verlauf sollen die von SPSS bereitgestellten Kriterien betrachtet werden.

2.5.1 Signifikanzniveau der Korrelation[10]

SPSS berechnet die Signifikanzniveaus der Korrelationskoeffizienten automatisch und gibt diese in einer Matrix aus. Getestet wird die Nullhypothese

$$H_0 : \textit{Es besteht kein Zusammenhang zwischen den Variablen}.$$

Das Signifikanzniveau des Korrelationskoeffizienten gibt die Irrtumswahrscheinlichkeit an, mit der die Nullhypothese abgelehnt werden kann. Ein Signifikanzniveau von 0,45 bedeutet dabei z.B., dass sich die Korrelation nur mit einer Wahrscheinlichkeit von 55% von Null unterscheidet. Zu 45% wird sich der Anwender täuschen, wenn er von einem Zusammenhang ungleich Null zwischen den Variablen ausgeht.

2.5.2 Bartlett-Test (test of sphericity)[11]

Auch wenn mittels der Korrelationsmatrix Zusammenhänge zwischen den Variablen identifiziert werden konnten, ist es dennoch möglich, dass sich diese nur zufällig für die zu Grunde gelegte Stichprobe ergeben haben. Der Bartlett-Test auf Sphärizität überprüft deshalb die Nullhypothese, dass die Korrelationsmatix der Grundgesamtheit eine Einheitsmatrix ist. Damit eine Faktorenanalyse durchgeführt werden kann, ist es notwendig, dass die Variablen untereinander korrelieren. Sollte die Korrelationsmatrix R tatsächlich eine Einheitsmatrix darstellen, wären alle Nichtdiagonalelemente Null, d.h. es würde keine Korrelation zwischen den Variablen vorliegen. Folglich ist es erstrebenswert, dass der Bartlett-Test signifikant wird und somit die Nullhypothese verworfen werden kann.

[9] Vgl. BACK S. 273

[10] Vgl. BACK S. 273

[11] Vgl. BACK S. 274

Der Bartlett-Test setzt voraus, dass die Variablen in der Erhebungsgesamtheit annähernd normalverteilt sind und die entsprechende Prüfgröße annähernd χ^2-verteilt mit $\frac{K \cdot (K-1)}{2}$ -Freiheitsgraden ist. Als Prüfgröße wird folgender Wert herangezogen:

$$PG = -\left[(T-1) \cdot \frac{1}{6} \cdot (2 \cdot K + 5) \right] \cdot \ln|R|$$

Falls PG größer ist als das $(1-\alpha)$-Quantil der χ^2-Verteilung mit $\frac{K \cdot (K-1)}{2}$-Freiheitsgraden, also $PG > F^{-1}(1-\alpha)$ oder $1 - F(PG) < \alpha$ gilt, kann die Nullhypothese verworfen werden und die Korreliertheit der Variablen ist bestätigt.

Ein großer Nachteil des Bartlett-Tests ist seine Empfindlichkeit gegenüber Abweichungen von der Normalverteilung. Der Wert der Prüfgröße wird außerdem in hohem Maße von der Stichprobengröße beeinflusst. Deshalb ist es sinnvoll, noch andere Kriterien zu betrachten, bevor die Variablen für eine Faktorenanalyse verwendet werden.

2.5.3 Kaiser-Meyer-Olkin-Maß

Das Kaiser-Meyer-Olkin-Maß (KMO-Maß) ist ein Maß zur Überprüfung der Stichprobenangemessenheit und gilt als bestes Maß zur Beurteilung der Eignung einer Korrelationsmatrix für eine Faktorenanalyse. Es setzt die einfachen Korrelationskoeffizienten zu den partiellen Korrelationskoeffizienten in Beziehung und schwankt zwischen 0 und 1. Dabei wird der Wert umso größer, je kleiner die partiellen Korrelationen ausfallen. Das KMO-Maß wird wie folgt berechnet:

$$KMO = \frac{\sum\limits_{i \neq j} \sum r_{ij}^2}{\sum\limits_{i \neq j} \sum r_{ij}^2 + \sum\limits_{i \neq j} \sum r_{ij.2...k}^2}$$

Dabei bezeichnet r_{ij} den einfachen Korrelationskoeffizienten zwischen der Variablen i und der Variablen j. Weil Korrelationen einer Variablen mit sich selbst natürlich nicht mit berücksichtigt werden dürfen, werden diese bei der Summenbildung ausgeschlossen. $r_{ij.2...k}^2$ bezeichnet die partiellen Korrelationskoeffizienten. Diese beschreiben den Zusammenhang zwischen zwei Variablen, wenn der Einfluss der anderen Variablen ausgeschaltet ist:

$$r_{xy.z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2) \cdot (1 - r_{yz}^2)}}$$

wobei r_{xy}, r_{xz} und r_{yz} die Pearson-Korrelationskoeffizienten der Variablen x, y und z darstellen[12].

[12] Vgl. SACHS S. 571

Das KMO-Maß nimmt Werte in der Nähe von Eins an, wenn die partiellen Korrelationskoeffizienten klein sind und wird sehr klein, wenn die partiellen Korrelationen hoch sind. Um mehrere Variablen über einen Faktor beschreiben zu können, ist es notwendig, dass die Korrelationen zwischen zwei Variablen auf andere Variablen zurückzuführen sind. D.h. man braucht möglichst kleine partielle Korrelationskoeffizienten. Zur Beurteilung des KMO-Maßes wird folgende Beurteilung herangezogen:

Wert	Beurteilung
$\geq 0{,}9$	fabelhaft (marvelous)
$\geq 0{,}8$	verdienstvoll (meritorious)
$\geq 0{,}7$	mittelprächtig (middling)
$\geq 0{,}6$	mäßig (mediocre)
$\geq 0{,}5$	kläglich (miserable)
$< 0{,}5$	inakzeptabel (unacceptable)

Tabelle 1: Beurteilung des Ergebnis des KMO-Maß nach Kaiser[13]

Das KMO-Maß sollte mindestens einen Wert von 0,7 aufweisen, damit sich die Korrelationsmatrix für eine Faktorenanalyse eignet. Werte unter 0,5 sind dagegen für eine Faktorenanalyse nicht tragbar.

Das variablenspezifische KMO-Maß wird auch MSA-Maß (measure for sampling adequacy) bezeichnet und berechnet sich wie folgt:

$$MSA_i = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} r_{ij.2...k}^2}$$

Das MSA-Maß ist ein Maß für die Korrelation einer Variablen i mit allen anderen Variablen. Die Interpretation der angenommenen Werte erfolgt auf die gleiche Weise wie beim KMO-Maß[14].

2.5.4 Image-Analyse von Guttman[15]

Die Image-Analyse von Guttman beruht darauf, die Varianz in zwei Teile zu zerlegen: das Image und das Anti-Image. Der Varianzanteil, der von den übrigen Variablen anhand einer Regressionsanalyse erklärt werden kann, wird als Image bezeichnet. Das Anti-Image stellt

[13] Vgl. BACK S. 276

[14] Vgl. FAKTOR S. 20 und BACK S. 276

[15] Vgl. BACK S. 275f, Beweis aus [6] S. 113ff.

somit den von den übrigen Variablen unabhängigen Teil dar. Da es für die Faktorenanalyse wichtig ist, dass die Variablen hoch korrelieren, sollte das Anti-Image möglichst gering ausfallen.

Zur Vereinfachung der nachfolgenden Formeln wird die Anzahl der Fälle aus den Gleichungen ausgeschlossen und es gilt z.B. $R = Z' \cdot Z$. Die Matrix D^2 enthält die Varianzen der Residuen für jede Variable regressiert auf die k-1 anderen Variablen. Für jede der k Variablen abwechselnd regressiert auf die k-1 anderen Variablen, kann die Matrix der Regressionskoeffizienten B geschrieben werden als: $B = E - R^{-1} \cdot D^2$. Die Residuen $e = (Z - \hat{Z})$ für jede Variable regressiert auf die übrigen Variablen betragen $Z - \hat{Z} = Z \cdot R^{-1} \cdot D^2$ und für die Matrix des Image der Variablen gilt $\hat{Z} = Z \cdot B = Z \cdot (E - R^{-1} \cdot D^2)$. Die Kovarianzmatrix des Image G und des Anti-Image P berechnen sich nun wie folgt:

$$
\begin{aligned}
G = \hat{Z}' \cdot Z &= (E - R^{-1} \cdot D^2)' \cdot \underbrace{Z' \cdot Z}_{R} \cdot (E - R^{-1} \cdot D^2) \\
&= (E - D^2 \cdot R^{-1}) \cdot R \cdot (E - R^{-1} \cdot D^2) \\
&= R + D^2 \cdot R^{-1} \cdot D^2 - 2 \cdot D^2
\end{aligned}
$$

$$
\begin{aligned}
P = e' \cdot e \; &= (Z - \hat{Z})' \cdot (Z - \hat{Z}) \\
&= (Z \cdot R^{-1} \cdot D^2)' \cdot (Z \cdot R^{-1} \cdot D^2) \\
&= (D^2 \cdot R^{-1}) \cdot R \cdot (R^{-1} \cdot D^2) \\
&= D^2 \cdot R^{-1} \cdot D^2
\end{aligned}
$$

Dziuban und Shirkey[16] schlagen vor, eine Korrelationsmatrix dann als ungeeignet zu betrachten, wenn der Anteil der Nicht-Diagonal-Elemente der Anti-Image-Kovarianzmatrix, die einen kritischen Wert von 0,09 überschreiten, über 25% liegt.

In der Anti-Image-Korrelationsmatrix sind die MSA-Werte als Gütemaß für die einzelnen Variablen auf der Hauptdiagonalen zu finden.

2.6 Entwicklung des Modells

Nachdem die ausgewählten Merkmale überprüft wurden, soll nun die Vorgehensweise für die Entwicklung des Modells der Faktorenanalyse beschrieben werden. Dazu wird zunächst

- das Fundamentaltheorem der Faktorenanalyse eingeführt und anschließend
- die Interpretation der Faktoren grafisch dargestellt.

[16] Vgl. BACK S. 275f.

2.6.1 Das Fundamentaltheorem der Faktorenanalyse[17]

Das Fundamentaltheorem der Faktorenanalyse unterstellt, dass sich jede der (standardisierten) Variablen als Linearkombination aus den mit den Faktorladungen gewichteten Faktorwerten darstellen lässt. Mathematisch lässt sich dies wie folgt formulieren:

$$z_{tk} = a_{k1} \cdot f_{t1} + a_{k2} \cdot f_{t2} + ... + a_{KQ} \cdot f_{TQ} = \sum_{q=1}^{Q} a_{kq} \cdot f_{tq} \quad \text{mit} \sum_{q=1}^{Q} a_{kq}{}^{2} = 1$$

Die Faktorladungen beschreiben dabei, wie stark die Faktoren mit den Variablen zusammenhängen. Die Matrix der Faktorladungen A kann daher auch als Korrelationsmatrix der Faktoren mit den Ausgangsvariablen bezeichnet werden. In Matrixnotation sieht diese „Grundgleichung der Faktorenanalyse“ folgendermaßen aus:

$$Z = F \cdot A'$$

Nun kann die Korrelationsmatrix R mit dieser Gleichung in Beziehung gesetzt werden. Wie in Abschnitt 2.3. beschrieben, lässt sich die Korrelationsmatrix R folgendermaßen berechnen:

$$R = \frac{1}{T-1} \cdot Z' \cdot Z$$

Wird nun für Z die Grundgleichung der Faktorenanalyse eingesetzt, erhält man:

$$\begin{aligned} R &= \frac{1}{T-1} \cdot (F \cdot A')'(F \cdot A') \\ &= \frac{1}{T-1} \cdot A \cdot F' \cdot F \cdot A' \\ &= A \cdot \underbrace{\frac{1}{T-1} \cdot F' \cdot F}_{C} \cdot A' \\ &= A \cdot C \cdot A' \end{aligned}$$

Dabei stellt die Matrix C die Korrelationsmatrix der Faktoren dar. Sie entspricht einer Einheitsmatrix (Hauptdiagonalelemente gleich Eins, Nebendiagonalelemente gleich Null), da die Ermittlung der Faktoren darauf beruht, dass diese orthogonal zueinander und somit unkorreliert sind[18]. Damit vereinfacht sich das Fundamentaltheorem der Faktorenanalyse nach Thurstone zu:

$$R = A \cdot A'$$

Das gekürzte Fundamentaltheorem gilt aber nur bei der Voraussetzung einer Linearverknüpfung und der Unabhängigkeit der Faktoren.

[17] Vgl. BACK S. 278f.

[18] Vgl. Abschnitt 2.2

2.6.2 Grafische Interpretation der Faktoren[19]

Der Korrelationskoeffizient lässt sich grafisch als Winkel zwischen Vektoren (Variablen) darstellen. Stehen die Vektoren senkrecht (orthogonal) aufeinander, d.h. der Winkel zwischen ihnen beträgt 90°, so sind sie voneinander unabhängig. In diesem Fall beträgt der Korrelationskoeffizient $r = 0$. Nimmt der Korrelationskoeffizient einen Wert von Eins an, so sind die Vektoren deckungsgleich, d.h. der Winkel zwischen ihnen beträgt 0°. Nach der Standardisierung ist die Länge eines Vektors Eins. Der Zusammenhang zwischen Winkel und Korrelationskoeffizient leitet sich über den Cosinus ab:

$$\cos\alpha = r\ .$$

Die Korrelationsmatrix lässt sich daher auch mit Winkelausdrücken beschreiben:

$$R = \begin{pmatrix} 1 & & \\ 0{,}5 & 1 & \\ 0{,}9397 & 0{,}766 & 1 \end{pmatrix} \quad \rightarrow \quad R = \begin{pmatrix} 0 & & \\ \cos(60^\circ) & 0 & \\ \cos(20^\circ) & \cos(40^\circ) & 0 \end{pmatrix}$$

Die entsprechenden Werte sind einer Cosinus-Tabelle zu entnehmen.

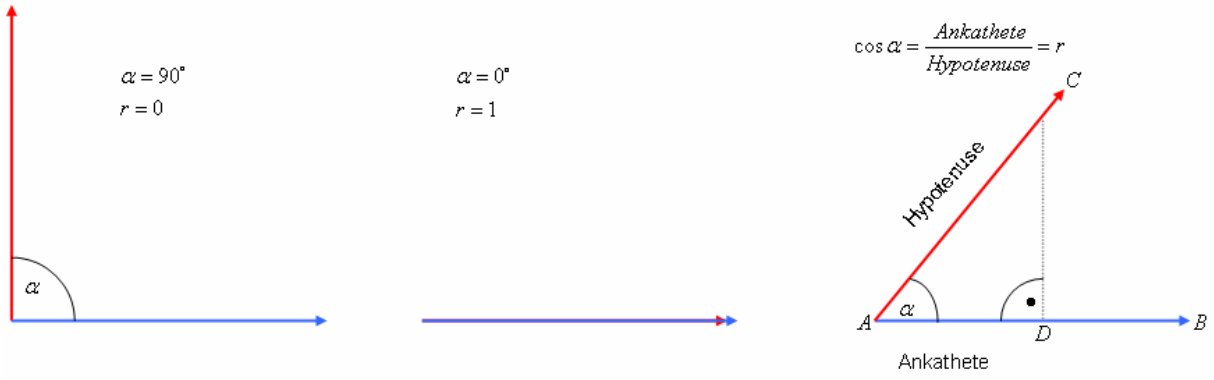


Abbildung 2: Vektordarstellung einer Korrelation[20]

Mit zunehmender Zahl der Variablen (und somit der Vektoren), steigt die Dimension der Darstellung. Ziel der Faktorenanalyse ist es, die Vektoren in einem möglichst klein dimensionierten Raum darzustellen. Die Zahl der benötigten Achsen gibt dabei die entsprechende Anzahl der Faktoren an.

Auf welche Art die Faktoren gewonnen werden (grafisch gesprochen), lässt sich am besten anhand eines kleinen Beispiels mit zwei Vektoren und zwei Faktoren beschreiben:

[19] Vgl. BACK S. 279ff.

[20] Quelle: Eigene Darstellung

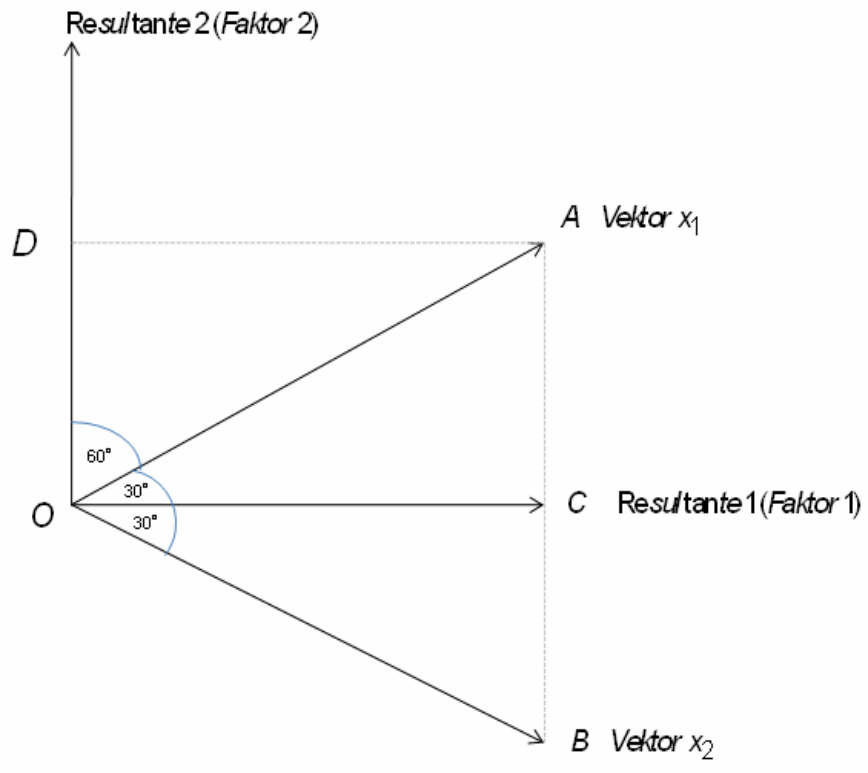


Abbildung 3: Zwei-Variablen-Zwei-Faktor-Lösung (vgl. BACK S.287)

Die zwei Variablen stehen als Vektoren (OA und OB) in einem 60°-Winkel zueinander.

Der Vektor OC (Resultante) ist die faktorielle Beschreibung der beiden Vektoren. Sie entspricht in diesem Beispiel der Winkelhalbierenden und die neu entstehenden 30°-Winkel repräsentieren die Korrelationskoeffizienten zwischen dem Faktor und den jeweiligen Variablen. Sie werden als Faktorladungen bezeichnet.

Der zweite Faktor lässt sich durch die orthogonale Errichtung eines Vektors in O auf den ersten Faktor finden, da die Faktoren unabhängig voneinander sein müssen.

Die Summe der quadrierten Faktorladungen ist bei vollständiger Erklärung der Variablen durch die gefundenen Faktoren gleich Eins. Dies bedeutet z.B. für die Variable x_1 [21]:

- Ladung des 1. Faktors: $\cos Winkel : COA = \frac{\overline{OC}}{\overline{OA}}$
- Ladung des 2. Faktors: $\cos Winkel : DOA = \frac{\overline{OD}}{\overline{OA}}$

Wenn die obige Behauptung stimmt, müsste gelten:

$$\left(\frac{\overline{OC}}{\overline{OA}}\right)^2 + \left(\frac{\overline{OD}}{\overline{OA}}\right)^2 = 1$$

[21] Vgl. BACK S. 288

Beweis:

$$\frac{\overline{OC}^2}{\overline{OA}^2} + \frac{\overline{OD}^2}{\overline{OA}^2} = \frac{\overline{OC}^2 + \overline{OD}^2}{\overline{OA}^2}$$

Laut dem Satz des Pythagoras gilt: $\overline{OA}^2 = \overline{OC}^2 + \overline{AC}^2$ und wie in Abb. 3 zu erkennen ist, gilt außerdem $\overline{AC} = \overline{OD}$. In obige Formel eingesetzt ergibt sich dann:

$$\frac{\overline{OC}^2 + \overline{OD}^2}{\overline{OC}^2 + \overline{OD}^2} = 1$$

Durch die Standardisierung der Ausgangsdaten ist die Standardabweichung der Variablen gleich Eins und somit auch die Varianz: $s_k^2 = 1$. Die Varianz steht auf der Diagonalen der Kovarianzmatrix. Da durch die Standardisierung die Kovarianzmatrix und die Korrelationsmatrix übereinstimmen, kann folgende Beziehung hergestellt werden: $s_k^2 = r_{kk} = 1$.

Daraus lässt sich mit der obigen Formel folgende wichtige Beziehung abgeleitet werden:

$$s_k^2 = r_{kk} = a_{k1}^2 + a_{k2}^2 + ... + a_{KQ}^2 = \sum_{q=1}^{Q} a_{kq}^2 = 1$$

Der Varianzerklärungsanteil der betrachteten Variablen lässt sich also durch Summation der quadrierten Faktorladungen darstellen.

$\sum_{q=1}^{Q} a_{kq}^2$ ist das Bestimmtheitsmaß und wenn alle möglichen Faktoren extrahiert werden ist sein Wert gleich Eins.

2.7 Bestimmung der Anzahl der zu extrahierenden Faktoren

Als informationshaltige Faktoren sollten nur diejenigen Faktoren akzeptiert werden, die einen genügend großen Anteil der Varianz im Datensatz erklären. Die Bestimmung des Varianzanteils, der erklärt werden soll, wird in diesem Kapitel in Abschnitt 2.7.1 beschrieben. Im Anschluss daran erfolgt die Berechnung der Faktorladungsmatrix A. Für die Bestimmung der optimalen Faktorenzahl lassen sich statistische Kriterien heranziehen, von denen insbesondere die folgenden als bedeutsam anzusehen sind und hier näher erläutert werden sollen[22]:

- Kaiser-Kriterium
- Scree-Test

[22] Vgl. BACK S. 295

2.7.1 Bestimmung der Kommunalitäten[23]

Als Kommunalität h_k^2 bezeichnet man den Teil der Gesamtvarianz einer Variablen, der durch die extrahierten Faktoren erklärt werden soll. Da die Faktoren i.d.R. nicht die gesamte Varianz erklären, sind die Kommunalitäten meistens kleiner als Eins. Daher muss das Fundamentaltheorem der Faktorenanalyse entsprechend angepasst werden:

$$R = A \cdot A' + U$$

U bezeichnet dabei die Residualmatrix, die die Einzelrestvarianz (=spezifische Varianz der Variablen und potentielle Messfehler) beschreibt.

Da die Kommunalitäten unbekannt sind, müssen sie vorab geschätzt werden. Dabei bedeutet eine festgelegte Kommunalität von 0,6, dass man vermutet, 60% der Ausgangsvarianz durch gemeinsame Faktoren erklären zu können. Die Kommunalitäten werden dann in die Haupt-Diagonale der Korrelationsmatrix eingesetzt und die Faktorenextraktion durchgeführt.

In dieser Untersuchung soll als Extraktionsmethode die Hauptkomponentenanalyse verwendet werden, bei der die Bestimmung der Kommunalitäten wie folgt durchgeführt wird:

Bei der Hauptkomponentenanalyse wird davon ausgegangen, dass die Varianz einer Ausgangsvariablen vollständig durch die extrahierten Faktoren erklärt werden kann. Man unterstellt also, dass keine Einzelrestvarianz in den Variablen existiert und setzt die Kommunalitäten auf Eins. Dieser Wert kann allerdings nur dann vollständig reproduziert werden, wenn ebenso viele Faktoren wie Variablen extrahiert werden. Da das Ziel der Faktorenanalyse eine Dimensionsreduktion ist, werden weniger Faktoren extrahiert wie Variablen, die z.B. 80% der Gesamtvarianz erläutern. Somit sind die Kommunalitätenwerte kleiner Eins und der nicht reproduzierte Varianzanteil beschreibt den Informationsverlust, der hierbei wissentlich in Kauf genommen wird.

2.7.2 Berechnung der Faktorladungsmatrix A[24]

In einem nächsten Schritt soll nun mit Hilfe des Fundamentaltheorems $R = A \cdot A'$ und der Grundgleichung der Faktorenanalyse $Z = F \cdot A'$ die Faktorladungsmatrix A berechnet werden. Ziel bei der Hauptkomponentenanalyse ist es, nacheinander unkorrelierte Faktoren f_q, $(q = 1,...,Q)$ zu ermitteln, deren Varianz jeweils maximal ist. Die Varianz der Faktoren berechnet sich wie folgt:

$$s_{f_q}^2 = a_{k1}^2 + a_{k2}^2 + ... + a_{KQ}^2 = \sum_{q=1}^{Q} a_{kq}^2 = 1$$

Für die Ermittlung des ersten Faktors f_1 wird ein sich aus den Faktorladungen $a_{11}, a_{21},..., a_{K1}$ zusammensetzender Vektor a_1 gesucht

[23] Vgl. BACK S. 289ff.

[24] Vgl. BACK S. 332ff.

$$a_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{K1} \end{pmatrix},$$

der die Gleichung $f_1 = Z \cdot a_1$ unter der Nebenbedingung $a_1^T \cdot a_1 = 1$ erfüllt. Die geforderte Varianzmaximierung lässt sich dabei wie folgt erreichen:

$$\begin{aligned} s_{f_1}^2 &= \frac{1}{T-1} \cdot f_1' \cdot f_1 \\ &= \frac{1}{T-1} \cdot (Z \cdot a_1)' \cdot (Z \cdot a_1) \\ &= \frac{1}{T-1} \cdot a_1' \cdot Z' \cdot Z \cdot a_1 \\ &= a_1' \cdot \underbrace{(\frac{1}{T-1} \cdot Z' \cdot Z)}_{=R \quad (s.S.23)} \cdot a_1 \\ &= a_1' \cdot R \cdot a_1 \\ &= \max! \end{aligned}$$

Dieses Optimierungsproblem lässt sich mit Hilfe des Lagrange-Ansatz lösen:

$$L(a_1, \lambda) = a_1' \cdot R \cdot a_1 - \lambda \cdot (a_1' \cdot a_1 - 1)$$

Zur Bestimmung des Maximums müssen die partiellen Ableitungen gebildet werden und es ergibt sich das folgende homogene Gleichungssystem:

$$(R - \lambda_1 E) a_1 = 0$$

Wegen $a_1' \cdot a_1 = 1$ gilt:

$$\lambda_1 = a_1' \cdot R \cdot a_1 = s_{f_1}^2$$

Da $s_{f_1}^2$ maximiert werden soll, muss λ_1 der größte Eigenwert von R sein. Der zugehörige Eigenvektor liefert die Ladungen von Faktor 1. Analog liefert der zweitgrößte Eigenwert die Ladungen von Faktor 2 usw.

Die Eigenwerte der Matrix R lassen sich demnach aus der charakteristischen Gleichung

$$\left| R - \lambda E \right| = 0$$

und die Eigenvektoren durch

$$(R - \lambda E) \vec{a} = 0$$

berechnen.

2.7.3 Reproduzierte Korrelationsmatrix[25] $\hat{R}$

Nachdem die Faktorladungsmatrix A ermittelt ist, kann nun die reproduzierte Korrelationsmatrix durch

$$\hat{R} = A \cdot A'$$

ermittelt werden. Auf der Hauptdiagonalen der reproduzierten Korrelationsmatrix können die endgültigen Kommunalitäten abgelesen werden. Die Nichtdiagonalelemente geben die durch die Faktorenstruktur reproduzierten Korrelationen wieder. Die Korrelationsmatrix kann dann vollständig reproduziert werden, wenn alle Faktoren extrahiert werden. Da dies in der Praxis nicht der Fall ist, treten mehr oder weniger große Abweichungen auf. Diese Abweichungen können in der von SPSS ausgegebenen Residualmatrix U betrachtet werden. Diese gibt die Differenzwerte zwischen den ursprünglichen und den reproduzierten Korrelationen aus. Dabei sollten besonders die Anteile > 0,05 betrachtet werden. Ist ihr Anteil hoch, müssen evtl. weitere Faktoren extrahiert werden, um eine bessere Abbildungsgenauigkeit zu erreichen.

2.7.4 Kaiser-Kriterium[26]

Das Kaiser-Kriterium nutzt den Eigenwert eines Faktors zur Bestimmung der optimalen Anzahl an Faktoren. Dieser gibt an, wie viel Varianz der Faktor q $(q = 1,..,Q)$ für alle Variablen insgesamt erklärt und berechnet sich wie folgt:

$$\lambda_q = \sum_{k=1}^{K} a_{kq}^2$$

Je höher der Eigenwert eines Faktors, desto größer ist sein Beitrag zur Erklärung der Varianz aller Variablen. Das Kaiser-Kriterium besagt nun, dass nur Faktoren mit einem Eigenwert > 1 extrahiert werden sollten. Die Begründung liegt darin, dass die Varianz einer standardisierten Variable Eins beträgt. Ein Faktor, dessen Varianzerklärungsanteil über alle Variablen kleiner als Eins ist, würde also weniger Varianz erklären, als eine einzelne Variable. Auch der prozentuale Anteil eines Faktors an der Erklärung der Gesamtvarianz lässt sich mit Hilfe der Eigenwerte bestimmen:

$$\textit{Anteil eines Faktors an der Gesamtvarianz in \%} = \frac{\lambda_q}{K} \cdot 100$$

2.7.5 Scree-Test[27]

Eine weitere Entscheidungshilfe für die Bestimmung der geeigneten Faktorenanzahl ist der so genannte Scree-Plot. Dieses Diagramm enthält ein Koordinatensystem, auf dessen y-Achse die nach ihrer Größe geordneten Eigenwerte abgetragen sind und auf der x-Achse die jeweils zugehörigen Faktornummern stehen. Mit steigender Faktornummer bekommt man

[25] Vgl. BACK S. 294

[26] Vgl. BACK S. 295 und S. 316

[27] Vgl. BACK S. 296

also immer kleiner werdende Eigenwerte und die Punkte nähern sich mehr und mehr der x-Achse an (s. Abbildung 4).

Typischerweise weist die Kurve des Scree-Plots einen ähnlichen Verlauf auf wie in Abbildung 4: Zunächst fällt die Kurve sehr steil ab, weist dann einen Knick auf und fällt im weiteren Verlauf nur noch sehr langsam ab. Der erste Punkt links des Knicks bestimmt dabei die Anzahl der zu extrahierenden Faktoren. Der Hintergrund dieser Vorgehensweise ist darin zu sehen, dass die Faktoren mit den kleinsten Eigenwerten für Erklärungszwecke als unbrauchbar (Scree = Geröll) angesehen werden und deshalb auch nicht extrahiert werden.

Ein Nachteil des Scree-Tests ist seine Unübersichtlichkeit, wenn viele Faktoren extrahiert werden. Daher ist er als Unterstützung von anderen Kriterien gedacht.

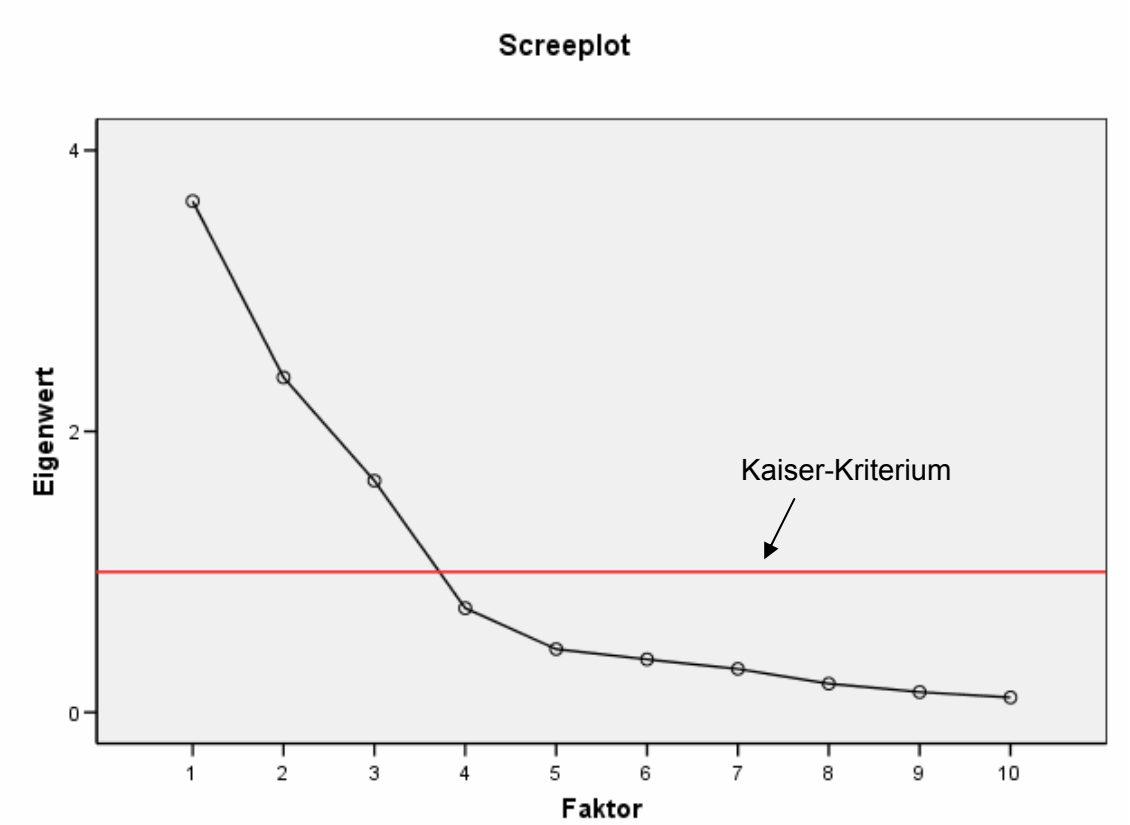


Abbildung 4: Beispiel eines Scree-Plots[28]

2.8 Faktorinterpretation

Nachdem die Zahl der Faktoren bestimmt wurde, muss geklärt werden, welche inhaltliche Bedeutung diese Faktoren besitzen. Da hier eine Hauptkomponentenanalyse durchgeführt wurde, muss nach einem Sammelbegriff gesucht werden, der den auf einen Faktor hochladenden Variablen entspricht. Als grobe Regel hat sich durchgesetzt, dass alle Variablen zur Interpretation der Bedeutung eines Faktors herangezogen werden sollten, die im Betrag ≥

[28] Quelle: Eigene Darstellung

0,5 auf einen Faktor laden. Sollte eine Variable auf mehreren Faktoren Ladungen von 0,5 oder mehr aufweisen, so muss sie bei jedem dieser Faktoren zur Interpretation mit herangezogen werden[29].

Um die einzelnen Faktoren möglichst gut interpretieren zu können, wäre es wünschenswert, dass eine Einfachstruktur vorliegt. Dies ist der Fall, wenn die einzelnen Variablen immer nur auf einen Faktor hoch und auf alle anderen Faktoren möglichst gering laden. In der Praxis ist es aber meistens so, dass mehrere Variablen auf mehrere Faktoren relativ gleich hoch laden, was eine sinnvolle Interpretation der Faktoren schwierig macht.

Um dieses Problem zu lösen, wendet man so genannte Rotationsverfahren an. Hierbei kann zwischen Methoden der obliquen und der orthogonalen Rotation unterschieden werden. Die in dieser Diplomarbeit ermittelten Faktoren sollen hinterher in eine Regressionsanalyse einfließen und müssen deshalb unabhängig voneinander sein. Daher wird hier nur die Methode der orthogonalen Rotation näher erläutert.

2.8.1 Orthogonale Rotation

Bei der orthogonalen (rechtwinkligen) Rotation wird das Koordinatensystem in seinem Ursprung so gedreht, dass die Achsen weiterhin senkrecht aufeinander stehen. Der erklärte Varianzanteil ergibt sich damit weiterhin als Summe der quadrierten Ladungen. Als Algorithmus wird in dieser Diplomarbeit die Varimax-Rotation verwendet.

2.8.1.1 Varimax-Rotation

Bei der Varimax-Rotation wird das Koordinatensystem im Ursprung so gedreht, dass die Varianz der Ladungsquadrate maximal ist. Der Rotationswinkel α wird dabei entgegengesetzt dem Uhrzeigersinn gemessen. Die rotierte Ladungsmatrix ergibt sich durch Multiplikation von A mit der Transformationsmatrix M :

$$\tilde{A} = A \cdot M$$

Bei der Betrachtung von zwei Faktoren hat M die Form

$$M = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$

Damit alle Variablen mit dem gleichen Gewicht in die Varimax-Rotation eingehen, müssen die Ladungen normiert werden. Hierfür werden die Ladungen der einzelnen Variablen durch die Wurzel aus den erklärten Varianzanteilen dividiert:

$$\tilde{a}_{k,q} = \frac{a_{kq}}{\sqrt{\sum_{q=1}^{Q} a_{kq}^2}}$$

Das Varimax-Kriterium lautet dann[30]

[29] Vgl. BACK S. 298f und WIRTZ S. 211

$$V = \frac{1}{K}\sum_{q=1}^{Q}(\sum_{k=1}^{K}\tilde{a}_{kq}^{4} - \frac{1}{K}\sum_{k=1}^{K}\tilde{a}_{kq}^{2}) \rightarrow Max$$

wobei $\tilde{a}_{kq}$ die Elemente der transformierten Faktorladungsmatrix $\tilde{A}$ sind.

Die Berechnung der Rotationsmatrix M erfolgt iterativ über die Maximierung von V. Die Bedingungen für das Finden eines Maximums werden hier nicht näher ausgeführt. Nach der Rotation werden die Ladungen wieder mit der Wurzel aus den erklärten Varianzanteilen multipliziert, um die ursprünglichen Kommunalitäten wiederherzustellen, die sich durch die Rotation nicht verändern. Lediglich die durch die einzelnen Faktoren aufgeklärten Varianzanteile – und somit auch deren Eigenwerte – verändern sich durch die Rotation.

2.9 Bestimmung der Faktorwerte[31]

Da die extrahierten Faktoren hinterher in eine Regressionsanalyse mit einbezogen werden sollen, müssen nun die Faktorwerte berechnet werden. Diese geben die Ausprägungen der Faktoren bei den jeweiligen Beobachtungen an und werden in der Matrix F zusammengefasst.

Ausgangsgleichung für die Bestimmung der Faktorwerte ist

$$Z = F \cdot A'$$

Da Z gegeben ist und die Faktorladungen der Matrix A bestimmt wurden, kann die Gleichung nach der Matrix der gesuchten Faktorwerte F aufgelöst werden:

$$Z \cdot (A')^{-1} = F \cdot \underbrace{A' \cdot (A')^{-1}}_{E}$$

$$Z \cdot (A')^{-1} = \underbrace{F \cdot E}_{F \cdot E = F}$$

$$F = Z \cdot (A')^{-1}$$

Da die Matrix A i.d.R. nicht quadratisch ist (es sollen ja gerade weniger Faktoren als Variablen extrahiert werden), ist eine Inversion nicht möglich. In diesem Fall wird von rechts mit A multipliziert:

$$Z \cdot A = F \cdot A' \cdot A$$

Die Matrix $(A' \cdot A)$ ist quadratisch und damit invertierbar:

$$Z \cdot A \cdot (A' \cdot A)^{-1} = F \cdot (A' \cdot A) \cdot (A' \cdot A)^{-1}$$

Da $(A' \cdot A) \cdot (A' \cdot A)^{-1}$ definitionsgemäß eine Einheitsmatrix ergibt, gilt:

$$F = Z \cdot A \cdot (A' \cdot A)^{-1}$$

Zur Lösung dieser Gleichung sind evtl. Schätzverfahren anzuwenden.

---

[30] Vgl. STIER, S. 294

[31] Vgl. BACK S. 302ff.

2.10 Beispiel einer Faktorenanalyse mit SPSS

Im Folgenden soll auf Basis des fiktiven Datensatzes Faktor.sav[32] eine Faktorenanalyse durchgeführt und erläutert werden. Die Stichprobe enthält sozioökonomische Daten von 12 Regionen. Die Faktorenanalyse wird mit Hilfe der Statistik-Software SPSS® 14.0 durchgeführt. Die wichtigsten von SPSS erzeugten Ergebnisse sollen kurz erläutert werden.

Pfad: *Analysieren → Dimensionsreduktion → Faktorenanalyse...*

```
FACTOR
  /VARIABLES Dichte BIP Besch_LW WBIP Geburt Wanderung  /MISSING LISTWISE
  /ANALYSIS Dichte BIP Besch_LW WBIP Geburt Wanderung
  /PRINT UNIVARIATE INITIAL CORRELATION SIG KMO INV REPR AIC EXTRACTION
  ROTATION FSCORE
  /FORMAT BLANK(.3)
  /PLOT EIGEN
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PC
  /CRITERIA ITERATE(25)
  /ROTATION VARIMAX
  /SAVE REG(ALL)
  /METHOD=CORRELATION .
```

Deskriptive Statistiken

	Mittelwert	Standard-abweichung	N
Dichte Einwohner je km²	326,558	173,8854	12
BIP Bruttoinlandsprodukt je Einwohner	22532,08	1940,683	12
Besch_LW Anteil Erwerbstätiger in der Landwirtschaft	10,200	5,4734	12
WBIP Wachstumsrate BIP (der letzten 10 Jahre)	57,542	10,0916	12
Geburt Lebendgeborene je 1000 Einwohner	9,608	2,6872	12
Wanderung Wanderungssaldo je 1000 Einwohner	,742	3,0684	12

Tabelle 2: SPSS-Ausgabe der deskriptiven Statistiken

Die erste Tabelle „Deskriptive Statistiken“ zeigt für die gültigen Fälle (jew. N=12) die Mittelwerte und Standardabweichungen an. Anhand dieser Werte kann die Standardisierung der Ausgangsdatenmatrix vorgenommen werden.

[32] Daten aus VOSS, S. 538

Korrelationsmatrix

		Dichte	BIP	Besch_LW	WBIP	Geburt	Wanderung
Korrelation	Dichte Einwohner je km²	1,000	,907	-,834	-,161	-,787	-,309
	BIP Bruttoinlandsprodukt je Einwohner	,907	1,000	-,845	-,054	-,711	-,220
	Besch_LW Anteil Erwerbstätiger in der Landwirtschaft	-,834	-,845	1,000	-,067	,719	,039
	WBIP Wachstumsrate BIP (der letzten 10 Jahre)	-,161	-,054	-,067	1,000	,226	,832
	Geburt Lebendgeborene je 1000 Einwohner	-,787	-,711	,719	,226	1,000	,454
	Wanderung Wanderungssaldo je 1000 Einwohner	-,309	-,220	,039	,832	,454	1,000
Signifikanz (1-seitig)	Dichte Einwohner je km²		,000	,000	,309	,001	,164
	BIP Bruttoinlandsprodukt je Einwohner	,000		,000	,434	,005	,246
	Besch_LW Anteil Erwerbstätiger in der Landwirtschaft	,000	,000		,418	,004	,452
	WBIP Wachstumsrate BIP (der letzten 10 Jahre)	,309	,434	,418		,240	,000
	Geburt Lebendgeborene je 1000 Einwohner	,001	,005	,004	,240		,069
	Wanderung Wanderungssaldo je 1000 Einwohner	,164	,246	,452	,000	,069	

Tabelle 3: SPSS-Tabelle der Korrelationen und ihrer Signifikanzniveaus

Der Korrelationsmatrix kann entnommen werden, dass zwischen einem Großteil der untersuchten Variablen ein Zusammenhang besteht. Bei ca. der Hälfte der Variablen ist der Betrag des Korrelationskoeffizienten größer als 0,7, was auf einen starken Zusammenhang hinweist.

KMO- und Bartlett-Test

Maß der Stichprobeneignung nach Kaiser-Meyer-Olkin.		,693
Bartlett-Test auf Sphärizität	Ungefähres Chi-Quadrat	48,881
	df	15
	Signifikanz nach Bartlett	,000

Tabelle 4: SPSS-Ausgabe des KMO-Maßes und des Bartlett-Tests

Nach dem KMO-Maß ist die Korrelationsmatrix nur mäßig für eine Hauptkomponentenanalyse geeignet. Die Nullhypothese des Bartlett-Test besagt, dass die Variablen in der Grundgesamtheit unkorreliert sind. Diese kann auf einem Signifikanzniveau von 1% zurückgewiesen werden, so dass davon auszugehen ist, dass die Variablen in einem Maße korrelieren, das sie für eine Faktorenanalyse geeignet erscheinen lässt.

Anti-Image-Matrizen

Anti-Image-Korrelation

	Dichte	BIP	Besch_LW	WBIP	Geburt	Wanderung
Dichte Einwohner je km²	,798(a)	-0,642	0,22	0,149	0,3	-0,013
BIP Bruttoinlandsprodukt je Einwohner	-0,642	,757(a)	0,385	-0,211	-0,162	0,207
Besch_LW Anteil Erwerbstätiger in der Landwirtschaft	0,22	0,385	,774(a)	-0,142	-0,423	0,41
WBIP Wachstumsrate BIP (der letzten 10 Jahre)	0,149	-0,211	-0,142	,482(a)	0,293	-0,814
Geburt Lebendgeborene je 1000 Einwohner	0,3	-0,162	-0,423	0,293	,747(a)	-0,51
Wanderung Wanderungssaldo je 1000 Einwohner	-0,013	0,207	0,41	-0,814	-0,51	,479(a)

a Maß der Stichprobeneignung

Tabelle 5: Ausgabe der Anti-Image-Korrelationsmatrix

Das variablenspezifische MSA-Maß ist auf der Hauptdiagonalen der Anti-Image-Korrelationsmatrix abgetragen. In diesem Beispiel weisen die Variablen „Wachstumsrate des BIP“ und „Wanderungssaldo je 1000 Einwohner“ inakzeptable Werte auf und müssten sukzessive aus dem Modell entfernt werden, bis alle MSA-Werte >0,5 betragen. Darauf wird in diesem Beispiel verzichtet.

Kommunalitäten

	Anfänglich	Extraktion
Dichte Einwohner je km²	1,000	,920
BIP Bruttoinlandsprodukt je Einwohner	1,000	,891
Besch_LW Anteil Erwerbstätiger in der Landwirtschaft	1,000	,897
WBIP Wachstumsrate BIP (der letzten 10 Jahre)	1,000	,905
Geburt Lebendgeborene je 1000 Einwohner	1,000	,798
Wanderung Wanderungssaldo je 1000 Einwohner	1,000	,933

Extraktionsmethode: Hauptkomponentenanalyse.

Tabelle 6: SPSS-Tabelle der Kommunalitäten

Die sechste Tabelle gibt Auskunft über die anfänglichen Kommunalitäten, die bei der Hauptkomponentenanalyse immer auf Eins gesetzt werden. Die berechneten Kommunalitäten sind in der Spalte „Extraktion“ aufgeführt. So wird z.B. durch alle extrahierten Faktoren 92% der Varianz der Variablen „Dichte“ erklärt und 89,1% der Varianz der Variablen „Bruttoinlandsprodukt“.

Erklärte Gesamtvarianz

Komponente	Anfängliche Eigenwerte		
	Gesamt	% der Varianz	Kumulierte %
1	3,562	59,366	59,366
2	1,782	29,704	89,070
3	,301	5,021	94,090
4	,182	3,027	97,117
5	,102	1,698	98,815
6	,071	1,185	100,000

Extraktionsmethode: Hauptkomponentenanalyse.

Tabelle 7: SPSS-Ausgabe der erklärten Gesamtvarianz

Unter „erklärte Gesamtvarianz“ sind die Eigenwerte der einzelnen Komponenten und der von ihnen erklärte Varianzanteil angegeben. So hat die erste Komponente einen Eigenwert von 3,562 bei einem erklärten Varianzanteil von 59,366%. Die zweite Komponente bestimmt 29,704% der Restvarianz und erklärt zusammen mit der ersten Komponente 89,07% der Varianz.

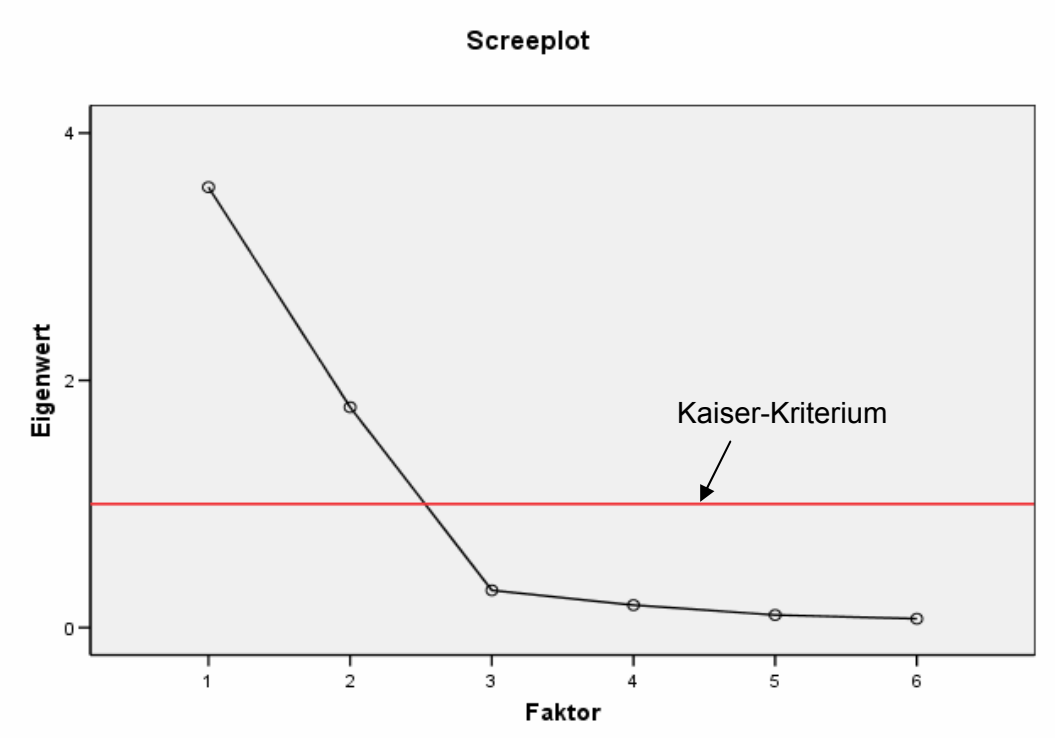


Abbildung 5: Auswertung des Scree-Tests

Wie in Tabelle 7 zu sehen ist, liegen nur die Eigenwerte der ersten zwei Faktoren über Eins. Nach dem Kaiser-Kriterium sind damit zwei Faktoren zu extrahieren. Der Scree-Plot zeigt in diesem Beispiel beim dritten Eigenwert einen deutlichen Knick. Somit legt auch der Scree-Plot eine Extraktion von zwei Faktoren nahe.

Komponentenmatrix(a)

	Komponente	
	1	2
Dichte Einwohner je km²	-,950	
BIP Bruttoinlandsprodukt je Einwohner	-,913	
Besch_LW Anteil Erwerbstätiger in der Landwirtschaft	,860	-,395
WBIP Wachstumsrate BIP (der letzten 10 Jahre)		,908
Geburt Lebendgeborene je 1000 Einwohner	,892	
Wanderung Wanderungssaldo je 1000 Einwohner	,456	,851

Extraktionsmethode: Hauptkomponentenanalyse.
a 2 Komponenten extrahiert

Tabelle 8: Faktorwerte der extrahierten Faktoren

In Tabelle 8 wird die unrotierte Faktorladungsmatrix ausgegeben. Zur bessern Übersicht wurde die Ausgabe aller Ladungsgrößen < 0,3 unterdrückt. Um die Faktoren besser interpretieren zu können, werden diese noch einer Varimax-Rotation unterzogen.

Reproduzierte Korrelationen

	Dichte	BIP	Besch_LW	WBIP	Geburt	Wanderung
Dichte Einwohner je km²	,920(b)	,899	-,870	-,149	-,842	-,321
BIP Bruttoinlandsprodukt je Einwohner	,899	,891(b)	-,880	-,041	-,805	-,213
Besch_LW Anteil Erwerbstätiger in der Landwirtschaft	-,870	-,880	,897(b)	-,115	,751	,056
WBIP Wachstumsrate BIP (der letzten 10 Jahre)	-,149	-,041	-,115	,905(b)	,292	,902
Geburt Lebendgeborene je 1000 Einwohner	-,842	-,805	,751	,292	,798(b)	,444
Wanderung Wanderungssaldo je 1000 Einwohner	-,321	-,213	,056	,902	,444	,933(b)

Extraktionsmethode: Hauptkomponentenanalyse.
b Reproduzierte Kommunalitäten

Tabelle 9: Reproduzierte Korrelationsmatrix

Auf der Hauptdiagonalen der reproduzierten Korrelationsmatrix lassen sich die endgültigen Kommunalitäten ablesen (vgl. auch Tabelle 6). Durch die Extraktion der zwei Faktoren wird 92% der Varianz der Variablen Dichte erklärt. Der Vergleich mit der Korrelationsmatrix lässt erkennen, dass die Werte durch die Faktorladungsmatrix recht gut reproduziert werden konnten.

Rotierte Komponentenmatrix(a)

	Komponente	
	1	2
Dichte Einwohner je km²	-,946	
BIP Bruttoinlandsprodukt je Einwohner	-,943	
Besch_LW Anteil Erwerbstätiger in der Landwirtschaft	,940	
WBIP Wachstumsrate BIP (der letzten 10 Jahre)		,951
Geburt Lebendgeborene je 1000 Einwohner	,838	,311
Wanderung Wanderungssaldo je 1000 Einwohner		,950

Extraktionsmethode: Hauptkomponentenanalyse.

Tabelle 10: Ausgabe der rotierten Faktorladungen

Anhand der rotierten Komponentenmatrix können nun die extrahierten Faktoren interpretiert werden. Man erkennt, dass auf den ersten Faktor die Variablen „Anteil Erwerbstätiger in der Landwirtschaft“ und „Lebendgeborene je 1000 Einwohner“ stark positiv laden, die Variablen „Bevölkerungsdichte“ und „Bruttoinlandsprodukt“ dagegen stark negativ laden. Dieser Faktor nimmt also in Regionen mit

- geringer Bevölkerungsdichte
- geringem BIP
- hoher Beschäftigung in der Landwirtschaft und
- hohen Geburtenraten

tendenziell hohe Werte an. Hierbei handelt es sich also um Regionen in ländlichen Gebieten. Dieser Faktor könnte als „Gegenteil von Agglomeration“ interpretiert werden.

Der zweite Faktor weist hohe positive Ladungen bei den Variablen „Wachstumsrate des BIP“ und „Wanderungssaldo“ auf. Er gibt demnach die „Attraktivität“ eines Standortes an.

Koeffizientenmatrix der Komponentenwerte

	Komponente	
	1	2
Dichte Einwohner je km²	-,277	-,010
BIP Bruttoinlandsprodukt je Einwohner	-,285	,050
Besch_LW Anteil Erwerbstätiger in der Landwirtschaft	,297	-,138
WBIP Wachstumsrate BIP (der letzten 10 Jahre)	-,078	,510
Geburt Lebendgeborene je 1000 Einwohner	,232	,099
Wanderung Wanderungssaldo je 1000 Einwohner	-,022	,494

Extraktionsmethode: Hauptkomponentenanalyse.
Rotationsmethode: Varimax mit Kaiser-Normalisierung.
Komponentenwerte.

Tabelle 11: Koeffizientenmatrix der Komponentenwerte

Mit Hilfe der Koeffizientenmatrix der Komponentenwerte können die Faktorwerte berechnet werden, die von SPSS in Dateneditor ausgegeben werden. Da zu ihrer Bestimmung die standardisierten Variablen verwendet wurden, liegen auch die Faktorwerte in standardisierter Form vor.

3 Lineare Regressionsanalyse

3.1 Einführung

Die Regressionsanalyse ist vermutlich eines der am häufigsten eingesetzten statistischen Analyseverfahren[33] . Sie gehört zu den strukturprüfenden Verfahren und wird verwendet, um die Beziehung zwischen einer Anzahl von unabhängigen Variablen (x_1, x_2,...) und einer abhängigen Variablen (y) zu untersuchen. Ziel dabei ist es, eine funktionale Beziehung zwischen den Größen zu finden, die es erlaubt, aus vorgegebenen bzw. zu beliebigen Werten von X die jeweils abhängige Zielgröße Y zu schätzen.

Der Begriff der „Regression“ geht auf den englischen Wissenschaftler Sir Francis Galton (1822 – 1911) zurück, der die Abhängigkeit der Körpergröße von Söhnen in Abhängigkeit der Körpergröße ihrer Väter untersuchte und dabei die Tendenz einer Rückkehr (regress) zur durchschnittlichen Körpergröße feststellte. D.h. z.B., dass die Söhne von extrem großen Vätern tendenziell weniger groß und die von extrem kleinen Vätern tendenziell weniger klein sind[34]. Das Ziel der Regressionsanalyse besteht somit darin, den Einfluss der erklärenden Variablen auf den Mittelwert der Zielgröße zu untersuchen.

In diesem Kapitel soll nun ein kurzer Überblick über die Theorie und Anwendung des linearen Regressionsmodells gegeben werden und der Prozess zur Bestimmung der Regressionsgeraden beschrieben werden. Anschließend folgt die Erklärung des zugehörigen SPSS-Outputs anhand eines kleinen Beispiels zur Veranschaulichung bevor (in Kapitel 5 und 6) die für diese Diplomarbeit relevanten Datensätze analysiert und ausgewertet werden.

3.2 Das lineare Regressionsmodell

Wie bereits in der Einleitung erwähnt, soll der Einfluss einer Reihe erklärender Variablen X auf eine Variable y untersucht werden. Das Ziel der multiplen Regressionsanalyse ist die Ermittlung einer Modellgleichung, in der ein Regressionskoeffizient als ein Maß für die Änderung der abhängigen Variable interpretiert werden kann, wenn der entsprechende Prädiktor um eine Einheit ansteigt und alle anderen Prädiktoren konstant gehalten werden können.

Die Idee hinter der linearen Regressionsanalyse ist nun, die abhängige Variable y als eine Funktion der unabhängigen Variablen auszudrücken, die linear in den Parametern ist. Ob eine lineare Beziehung unterstellt werden kann, lässt sich evtl. anhand eines Streudiagramms erkennen, in dem die Beobachtungen als Punkte eingetragen werden. Streuen die Punkte bandförmig auf oder um eine gedachte Linie, dann ist die Annahme eines linearen Verlaufs gerechtfertigt, und die Punktwolke kann durch eine lineare Funktion beschrieben werden. Dieser Zusammenhang ist im Allgemeinen aber nicht exakt, sondern wird durch zufällige Störungen u_t überlagert.

[33] Vgl. BACK S. 46

[34] Vgl. BACK S. 49

Für diese Diplomarbeit werden additive Störgrößen zu Grunde gelegt, so dass das Regressionsmodell folgendermaßen aussieht:

$$y = \beta_1 + \beta_2 \cdot x_{t2} + ... + u_t \quad (t = 1,...,T)$$

Die Parameter β_t sind dabei unbekannt und müssen geschätzt werden. Der Parameter β_1 ist der Achsenabschnitt (Intercept) bzw. der Schnittpunkt mit der y-Achse. Die Modellparameter und ihre Schätzungen werden im weiteren Verlauf durch ein „Dach“ unterschieden (d.h. der Schätzer von β wird mit $\hat{\beta}$ bezeichnet usw.).

Für die Störgrößen werden die Annahmen getroffen $E(u_t) = 0$ und $Var(u_t) = \sigma^2$. Die Regressionsgerade wird mit Hilfe der Parameterschätzer berechnet, so dass folgendes Schätzmodell gilt:

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 \cdot x_{t2} + ... + \hat{\beta}_K \cdot x_{TK} \quad (t = 1,...,T)$$

Die Differenz zwischen den wahren Werten y_t und den geschätzten Werten $\hat{y}_t$ wird Residualgröße genannt und berechnet sich wie folgt:

$$e_t = y_t - \hat{y}_t$$

In Matrixnotation ergibt sich folgende Modellformulierung:

$$y = X \cdot \beta + u \; mit \; y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} \quad X = \begin{pmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \ddots & \vdots \\ x_{T1} & \cdots & x_{TK} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} \quad u = \begin{pmatrix} u_1 \\ u_2 \\ \cdots \\ u_T \end{pmatrix}$$

Hier sind K erklärende Variablen aufgenommen (von denen eine i.d.R. das Absolutglied sein wird, d.h. i.d.R. gilt $x_{t1} = 1 (t = 1,...,T)$) und es liegen Beobachtungen für T Zeitpunkte vor.

In diesem Modell wird mit den folgenden Annahmen gearbeitet[35]:

- X ist eine deterministische (TxK)-Matrix mit $rg(X) = K$.
- β ist ein reeller Parametervektor
- u ist ein $(Tx1)$-Zufallsvektor mit $E(u) = 0$ und $Var(u) = \sigma^2 \cdot E$
- Häufig fügt man ergänzend hinzu: u ist normalverteilt, d.h. $u \sim N(0, \sigma^2 \cdot E)$

Die Schätzfunktion besitzt die Form

$$\hat{y} = X'\hat{\beta}$$

[35] Vgl. BACH

3.3 Parameterschätzungen

Die Residualgröße e_t bezeichnet den Schätzfehler, also die Abweichung des wahren Wertes y_t vom Schätzwert $\hat{y}_t$ (vgl. Abb. 5). Die Residuen beinhalten also die nicht erklärte Restreuung der Daten. Das Ziel der Regressionsanalyse ist nun, eine Funktion zu finden, für die diese Abweichungen möglichst klein sind.

Dazu wird im nächsten Abschnitt die Methode der kleinsten Quadrate (engl. Ordinary Least Squares, OLS) vorgestellt und die unbekannten Parameter β und σ^2 geschätzt.

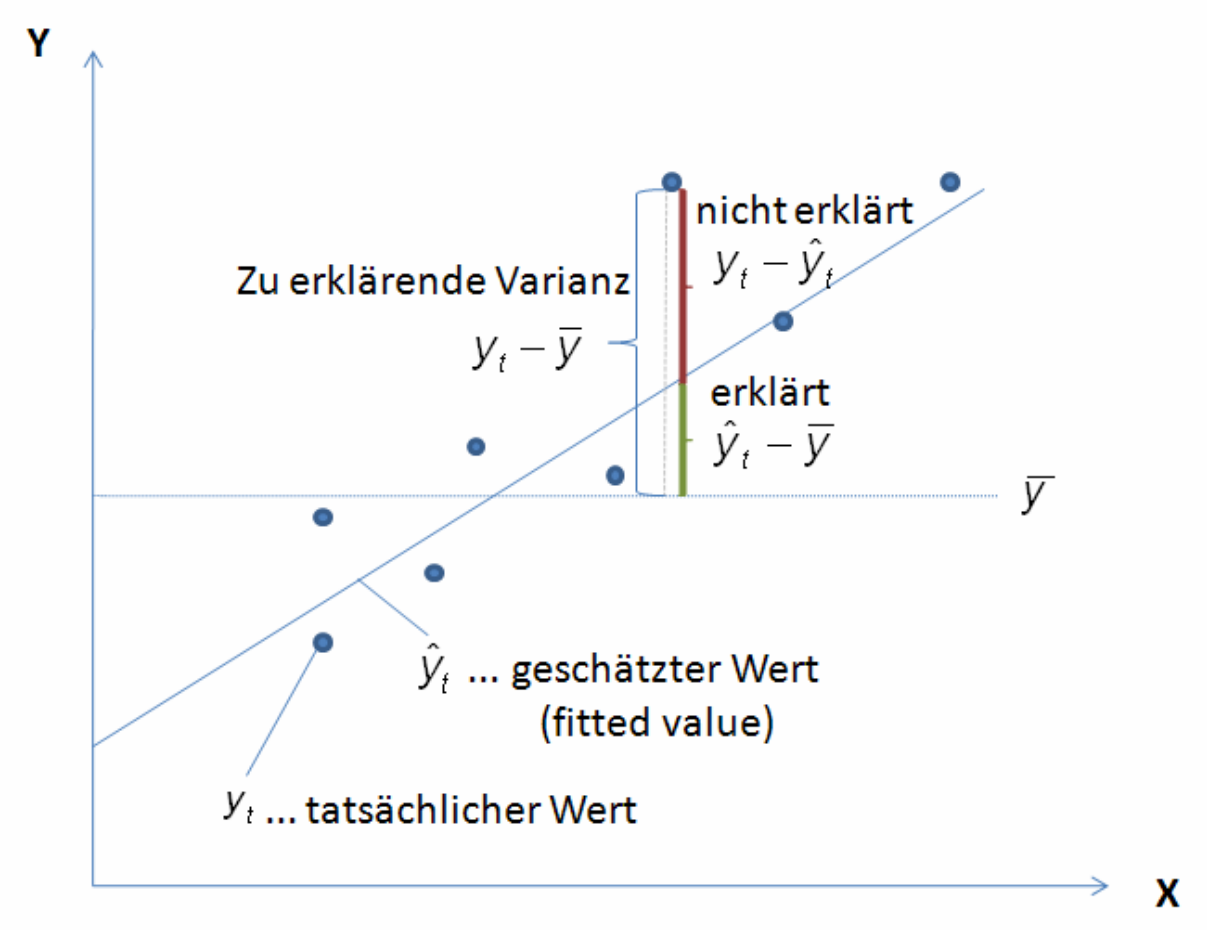


Abbildung 6: Streudiagramm mit Regressionsgerade[36]

[36] Quelle: Eigene Darstellung

3.3.1 Schätzung der Regressionskoeffizienten

3.3.1.1 Methode der kleinsten Quadrate[37]

Der unbekannte Parameter β soll nach der Methode der kleinsten Quadrate (KQ-Methode) geschätzt werden. Dazu wird $\hat{\beta}$ so bestimmt, dass die Summe der quadrierten Residuen minimal wird:

$$\sum_t e_t^2 = \sum_t (y_t - \hat{y}_t)^2 \to \min_{\hat{\beta}}$$

Die Minimierungsbedingung kann auch in Matrixnotation formuliert werden:

$$\langle y - \hat{y}; y - \hat{y} \rangle \to \min_{\hat{\beta}} \text{ bzw. } (y - X \cdot \hat{\beta})' \cdot (y - X \cdot \hat{\beta}) \to \min_{\hat{\beta}} \qquad \text{mit } \hat{y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_T \end{pmatrix}$$

Zur Bestimmung des Minimums erfolgt zunächst die Umformung

$$(y - X \cdot \hat{\beta})'(y - X \cdot \hat{\beta}) = y' \cdot y - \underbrace{\hat{\beta}' \cdot X' \cdot y}_{=(\hat{\beta}' \cdot X' \cdot y)' = y' \cdot X \cdot \hat{\beta}} - y' \cdot X \cdot \hat{\beta} + \hat{\beta}' \cdot (X' \cdot X) \cdot \hat{\beta}$$

$$= y' \cdot y - 2 \cdot y' \cdot X \cdot \hat{\beta} + \hat{\beta}' \cdot (X' \cdot X) \cdot \hat{\beta}$$

und die Bezeichnung

$$f(\hat{\beta}) := y' \cdot y - 2 \cdot y' \cdot X \cdot \hat{\beta} + \hat{\beta}' \cdot (X' \cdot X) \cdot \hat{\beta} .$$

Um diese Funktion zu minimieren, wird zunächst die erste Ableitung gebildet:

$$\frac{\partial f}{\partial \hat{\beta}} \overset{!}{=} 0 \quad \Leftrightarrow \quad -2 \cdot X' \cdot y + \left[(X' \cdot X) + (X' \cdot X)' \right] \cdot \hat{\beta} \overset{!}{=} 0$$

$$\Leftrightarrow \quad -2 \cdot X' \cdot y + 2 \cdot (X' \cdot X) \cdot \hat{\beta} \overset{!}{=} 0$$

$$\Leftrightarrow \quad (X' \cdot X) \cdot \hat{\beta} \overset{!}{=} 0 \qquad (\textit{Normalengleichung})$$

$$\Leftrightarrow \quad \hat{\beta} = (X' \cdot X)^{-1} \cdot X' \cdot y$$

Nochmaliges Differenzieren liefert:

$$\frac{\partial^2 f}{\partial^2 \hat{\beta}} = 2 \cdot (X' \cdot X)$$

Gemäß Annahme sind die Spalten von X linear unabhängig, d.h. $rg(X) = K$. Die Matrix $X' \cdot X$ ist also positiv definit, so dass $\hat{\beta}$ tatsächlich ein Minimum darstellt.

[37] Vgl. BACH

Ein Nachteil bei der KQ-Methode ist die Anfälligkeit für Ausreißer, da durch die Quadrierung der Abweichungen der Beobachtungswerte von den Schätzwerten, größere Abweichungen stärker gewichtet werden.

3.3.1.2 Eigenschaften des KQ-Schätzers[38]

- $\hat{\beta}$ ist im klassischen Modell ein erwartungstreuer Schätzer für β, da $\mathrm{E}(\hat{\beta}) = \beta$
- Die Varianz des Schätzers ergibt sich zu $Var(\hat{\beta}) = \sigma^2 \cdot (X' \cdot X)^{-1}$
- Sind die Störgrößen normalverteilt mit Erwartungswert Null und Varianz-Kovarianz-Matrix $\sigma^2 \mathrm{I}$, so ist auch der KQ-Schätzer normalverteilt, d.h. es gilt dann $\hat{\beta} \sim N(\beta, \sigma^2 \cdot (X' \cdot X)^{-1})$.
- Die Residuen sind im Mittel Null, d.h. $\sum_t e_t = 0 \quad bzw. \quad \bar{e} = \frac{1}{T}\sum_t e_t = 0.$
- Der Mittelwert der geschätzten Werte $\hat{y}_t$ ist gleich dem Mittelwert der beobachteten Werte y_t, d.h. $\bar{\hat{y}} = \frac{1}{T}\sum_t \hat{y}_t = \bar{y}$.

(ohne Beweis)

3.3.2 Schätzung der Störgrößenvarianz[39]

Ein erwartungstreuer Schätzer für die Störgrößenvarianz σ^2 ist gegeben durch $\hat{\sigma}^2 = \frac{e' \cdot e}{T - K}$.

(ohne Beweis)

3.4 Überprüfung der Modellgüte

Nach der Schätzung der Regressionskoeffizienten muss das Regressionsmodell auf die Güte der Anpassung an die Daten untersucht werden. Dazu werden die folgenden Größen zur Beurteilung herangezogen:

- Bestimmtheitsmaß und korrigiertes Bestimmtheitsmaß
- F-Test
- Standardfehler der Schätzung

[38] Vgl. BACH

[39] Vgl. BACH

3.4.1 Bestimmtheitsmaß und korrigiertes Bestimmtheitsmaß[40]

Wie in Abbildung 6 zu sehen ist, lässt sich die Gesamtabweichung der Beobachtung y_t vom Mittelwert $\bar{y}$ in zwei Teile zerlegen und zwar in die erklärte und die nicht erklärte Abweichung.

Die Abweichung der Schätzgröße $\hat{y}$ vom Mittelwert $\bar{y}$ kann durch die Regressionsfunktion erklärt werden. Die Abweichung des beobachteten Wertes y_t von dem geschätzten Wert $\hat{y}_t$ (Residualgröße) hängt allerdings von nicht erfassbaren Größen ab (z.B. Messfehler) und kann daher nicht durch die Regressionsgerade erklärt werden.

Das Bestimmtheitsmaß wird nun ermittelt, um die Güte der Anpassung der Regressionsgeraden an die empirischen Daten zu beurteilen („goodness of fit“). Dazu werden die Summen der quadrierten Abweichungen analysiert, um zu verhindern, dass sich erklärte und nicht erklärte Abweichungen zum Teil kompensieren. Mit den Bezeichnungen

$SST = \sum_t (y_t - \bar{y})^2$ Sum of Squares Total

$SSR = \sum_t (\hat{y}_t - \bar{y})^2$ Sum of Squares Regression

$SSE = \sum_t (y_t - \hat{y}_t)^2 = \sum_t e_t^2$ Sum of Squares Error

gilt die Streuungszerlegungsformel: $SST = SSR + SSE$.

Das Bestimmtheitsmaß R² gibt den Anteil der erklärten Streuung an, der durch das Modell beschrieben wird:

$$R^2 = \frac{SSR}{SST} = \frac{erklärte\ Streuung}{Gesamtstreuung}$$

So bedeutet ein Wert von $R^2 = 0{,}785$, dass 78,5% der (senkrechten) Streuung in den y-Werten durch die Regressionsgerade erklärt werden können.

Ein Nachteil des R² ist, dass sein Wert immer ansteigt, wenn eine weitere unabhängige Variable im Modell aufgenommen wird, egal wie gering ihr Einfluss auf die Zielgröße auch sein mag. Da jede aufgenommene Variable einen Freiheitsgrad kostet, wäre es wünschenswert die Verbesserung in der Anpassung und den Verlust an Freiheitsgraden abzuwägen.

Genau das ist die Idee hinter dem korrigierten Bestimmtheitsmaß (engl. adjusted R²). Es ist definiert als

$$\bar{R}^2 = 1 - \frac{SSE/(T-K)}{SST/(T-1)} = 1 - (1-R^2)\frac{(T-1)}{(T-K)}$$

Das Einfügen des Terms (T-K) bestraft die Aufnahme neuer Variablen, so dass das korrigierte Bestimmtheitsmaß nur dann wächst, wenn die Anpassung deutlich besser wird. Bei einer

[40] Vgl. BACH

schlechten Anpassung des Modells kann das korrigierte Bestimmtheitsmaß abnehmen, evtl. auch negativ werden.

3.4.2 Test auf Signifikanz des Erklärungsansatzes[41]

Nachdem mit dem korrigierten Bestimmtheitsmaß überprüft wurde wie gut die Anpassung des Modells an die Daten ist, stellt sich die Frage nach der Signifikanz des Modells. Es muss ausgeschlossen werden, dass die Ergebnisse zufällig entstanden sind und in der Grundgesamtheit eigentlich kein linearer Zusammenhang existiert.

Als Prüfgröße wird dafür der F-Test verwendet. Als Nullhypothese fungiert hierbei die Annahme, dass in der Grundgesamtheit kein linearer Zusammenhang zwischen X und Y besteht, d.h. dass alle Regressionskoeffizienten β_K in der Grundgesamtheit Null sind.

$$H_0: \quad \beta_2 = \beta_3 = ... = \beta_K = 0$$
$$H_1: \quad \min d.ein\ \beta_i \neq 0 \ \ für\ 1 < i \leq K$$

Die Prüfgröße ergibt sich als gewichtetes Verhältnis aus SSR und SSE:

$$F_{emp} = \frac{SSR/(K-1)}{SSE/(T-K)}$$

Als kritischer Wert zur Prüfung der Nullhypothese dient ein theoretischer F-Wert, der aus einer F-Tabelle entnommen werden kann und mit dem empirischen F-Wert zu vergleichen ist. Wie bei allen statistischen Tests ist ein Signifikanzniveau festzulegen.

Es gilt:

$F_{emp} > F_{theor} \rightarrow H_0\ wird\ verworfen$

$F_{emp} \leq F_{theor} \rightarrow H_0\ kann\ nicht\ verworfen\ werden$

3.4.3 Standardfehler der Schätzung

Die Standardabweichung der Residuen wird als Standardschätzfehler bezeichnet. Er gibt die Streuung der y-Werte um die Regressionsgerade an:

$$s_e = \sqrt{\frac{\sum_t e_t^2}{T-K}}$$

Die Genauigkeit der Regressionsvorhersage wächst dabei mit kleiner werdendem Standardschätzfehler, d.h. je kleiner die Residuen in der Stichprobe ausfallen.

3.5 Überprüfung der Regressionskoeffizienten

Die eben betrachteten Größen erlauben eine Aussage zur Güte des gesamten Modells. Entscheidend sind jedoch auch die Regressionskoeffizienten der einzelnen erklärenden Variablen. Üblicherweise werden dafür folgende Kriterien betrachtet:

- T-Statistik

[41] Vgl. BACK S. 70ff. und BACH

- Konfidenzintervalle
- Plausibilität der ausgegebenen Werte

3.5.1 T-Test[42]

Eine signifikante F-Statistik bedeutet nicht notwendigerweise, dass alle Modellparameter von Null verschieden sind. Daher wird für jeden Regressionskoeffizienten ein Test auf Signifikanz (t-Test) durchgeführt mit den Hypothesen

$$H_0: \quad \beta_k = 0$$
$$H_1: \quad \beta_k \neq 0$$

Die Prüfgröße ergibt sich als Quotient aus dem jeweiligen geschätzten Regressionskoeffizienten und seinem Standardfehler:

$$t_{emp} = \frac{\hat{\beta}_k}{s_{e_K}}$$

Vor der Durchführung des Tests wird das Signifikanzniveau α bestimmt. Die Teststatistik t_{emp} ist t-verteilt mit $T - K$ Freiheitsgraden. Als kritischer Wert zur Prüfung der Nullhypothese dient ein theoretischer t-Wert, der aus einer t-Tabelle abgelesen werden kann. Dieser ist mit dem Absolutbetrag des empirischen t-Wertes zu vergleichen, da der t-Wert auch negativ werden kann.

Es gilt:

$|t_{emp}| > t_{theor} \rightarrow H_0$ *wird verworfen*

$|t_{emp}| \leq t_{theor} \rightarrow H_0$ *kann nicht verworfen werden*

3.5.2 Konfidenzintervalle[43]

Die Parameterschätzungen weichen mehr oder weniger vom wahren Wert des zu schätzenden Parameters ab. Um derartige Schwankungen einschätzen zu können, werden Bereiche bestimmt, in denen die Koeffizienten mit einer bestimmten Wahrscheinlichkeit liegen werden. Die Schätzgenauigkeit erhöht sich dabei mit kleiner werdender Intervallbreite.

Für die Modellparameter berechnen sich diese Konfidenzintervalle wie folgt:

$$\left[\hat{\beta}_k - t_{T-K;1-\alpha/2} \cdot s_{e_i}; \quad \hat{\beta}_k + t_{T-K;1-\alpha/2} \cdot s_{e_i} \right]$$

3.5.3 Plausibilitätskontrolle

Nachdem für jeden einzelnen Regressor der Einfluss auf die abhängige Variable untersucht wurde, muss nun noch überprüft werden, ob aus inhaltlicher Sicht die Höhe und die Vorzei-

[42] Vgl. BACH und BACK S. 74ff.

[43] Vgl. BACH

chen der Koeffizienten stimmig sind. Dabei kann sich z.B. ein methodisch gesehen signifikanter Einfluss als unplausibel erweisen.

3.6 Überprüfung der Modellannahmen

Bevor ein Regressionsmodell aufgrund seiner Modellgüte oder der statistischen Signifikanz seiner Parameter angenommen werden kann, muss es sorgfältig daraufhin überprüft werden, ob alle Modellprämissen eingehalten wurden. Ist das gefundene Modell valide, kann es inhaltlich interpretiert werden. Auf folgende Probleme muss das Modell untersucht werden:

- Multikollinearität
- Normalverteilung der Residuen
- Heteroskedastizität
- Autokorrelation

3.6.1 Multikollinearität[44]

Die Regressoren des linearen Regressionsmodell sollten nach Möglichkeit untereinander nicht linear abhängig sind. In der praktischen Anwendung kann ein gewisser Grad an linearer Abhängigkeit zwischen den unabhängigen Variablen nicht ganz ausgeschlossen werden. Allerdings werden mit zunehmender Multikollinearität die Schätzungen der Regressionsparameter unzuverlässiger und deshalb muss das Modell auf lineare Zusammenhänge unter den unabhängigen Variablen überprüft werden.

Um Multikollinearität aufzuspüren, empfiehlt es sich eine Regression jeder unabhängigen Variablen x_K auf die übrigen Regressoren durchzuführen. Das dabei berechnete Bestimmtheitsmaß dient als Grundlage für die Ermittlung des Toleranzmaßes Tol :

$$Tol = 1 - R_k^2$$

Ein Wert nahe Eins deutet auf geringe Multikollinearität hin. Der Kehrwert der Toleranz ist der Variance Inflation Factor VIF und ermittelt sich wie folgt:

$$VIF(\hat{\beta}_k) = \frac{1}{1 - R_k^2}$$

Existiert nun ein linearer Zusammenhang zwischen den erklärenden Variablen, so vergrößern sich die Varianzen der Regressionskoeffizienten um den Faktor VIF [45].

Ein hoher VIF -Wert deutet auf Multikollinearität hin, da z.B. ein VIF von 5 ein R² von 80% bedeutet.

[44] Vgl. BACK S. 89ff.

[45] Vgl. BACH

3.6.2 Normalverteilung der Residuen

Eine weitere Modellannahme bei der linearen Regressionsanalyse ist die Normalverteilung der Residuen, also $e_t \sim N(0;\sigma^2)$. Grafisch kann diese Annahme zum einen mit einem Histogramm, das die Häufigkeiten der standardisierten Residuen darstellt und eine Normalverteilungskurve eingezeichnet hat, überprüft werden. Das Histogramm muss einer Standardnormalverteilung mit Mittelwert Null und einer Standardabweichung von Eins entsprechen. Zum anderen kann mit Hilfe eines QQ-Plots die Normalverteilungsannahme überprüft werden. Hierbei plottet man die standardisierten Residuen gegen die entsprechenden Quantile der Standardnormalverteilung.

3.6.3 Heteroskedastizität[46]

Heteroskedastizität liegt dann vor, wenn die Streuung der Residuen (Abweichung der Schätzwerte vom tatsächlichen Wert) nicht konstant ist. Liegt Heteroskedastizität vor, wird die Schätzung ineffizient und verfälscht den Standardfehler des Regressionskoeffizienten. Dadurch wird auch die Schätzung des Konfidenzintervalls ungenau.

Grafisch kann Heteroskedastizität mit Hilfe eines Streudiagramms identifiziert werden. Hierbei werden die standardisierten Residuen gegen die standardisierten geschätzten Werte geplottet.

Ein bekanntes Testverfahren, um die Nullhypothese Homoskedastizität zu prüfen, ist der Goldfeld/Quandt-Test[47]. Dabei wird die Stichprobe vom Umfang T in zwei Unterstichproben T_1 und T_2 aufgeteilt (z.B. durch vorherige Inspektion der Residuen). Für beide Stichproben werden dann separate Regressionen geschätzt und die Varianzen der Residuen verglichen und ins Verhältnis gesetzt. Sind die Residuen normalverteilt und die Annahme der Homoskedastizität trifft zu, so folgt das Verhältnis der geschätzten Varianzen (größere im Zähler) einer F-Verteilung mit (T_1-K) und (T_2-K) Freiheitsgraden. Die F-Statistik berechnet sich wie folgt (es sei angenommen $\hat{\sigma}_2^2 > \hat{\sigma}_1^2$):

$$F_{emp} = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}$$

Fällt der berechnete F-Wert größer aus als der theoretische F-Wert gemäß F-Tabelle, so ist die Nullhypothese Homoskedastizität zugunsten der Alternativhypothese Heteroskedastizität zu verwerfen.

3.6.4 Autokorrelation

Residuen sind autokorreliert, wenn es eine Beziehung zwischen den Residuen für aufeinander folgende Beobachtungen gibt. Tritt Autokorrelation auf, kann die Fehlervarianz nicht mehr zuverlässig geschätzt werden. Bei positiver Autokorrelation wird die Varianz beispiels-

[46] Vgl. BACK S. 85ff.

[47] Siehe BACK S. 86

weise unterschätzt. Die t-Teststatistik wird damit vergrößert und die Konfidenzintervalle verbreitern sich. Mit Hilfe des Durbin-Watson-Tests kann die serielle Unkorreliertheit der Störgrößen überprüft werden. Für die geschätzten Residuen werden folgende Prüfgrößen berechnet:

$$d = \frac{\sum_{t=2}^{T} (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{T} \hat{e}_t^2}$$

Unter der Nullhypothese, dass keine Autokorrelation vorliegt, sollte sich d in der Nähe von 2 befinden. Die Residuen sind positiv autokorreliert, wenn $d \approx 0$ gilt.

3.7 Beispiel einer Regressionsanalyse mit SPSS

Im Folgenden soll auf Basis des fiktiven Datensatzes Miete.sav eine Regressionsanalyse durchgeführt und erläutert werden. Die Stichprobe enthält die Wohnungs- und Mietdaten von 555 Wohnungen. Um das Beispiel einfach zu halten, wird der Einfluss einer unabhängigen Variablen auf die abhängige Variable Miete mit Hilfe der Statistik-Software SPSS® 14.0 untersucht.

Pfad: *Analysieren → Regression → Linear...*

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS BCOV R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Miete
  /METHOD=ENTER Größe
  /PARTIALPLOT ALL
  /SCATTERPLOT=(Miete,Größe)
  /SCATTERPLOT=(*ZRESID,Größe)
  /SCATTERPLOT=(*ZRESID ,*ZPRED )
  /RESIDUALS DURBIN HIST(ZRESID) NORM(ZRESID)
  /SAVE ZPRED COOK LEVER ZRESID MAHAL .
```

Um die Residuen einer grafischen Analyse unterziehen zu können, werden drei Streudiagramme angefordert: $x \cdot y$, $x \cdot y_{\text{Re siduen}}$ und $\hat{y} \cdot y_{\text{Re siduen}}$:

```
GRAPH
  /SCATTERPLOT (BIVAR) = Größe WITH Miete
  /MISSING LISTWISE
  /TITLE= "Linearitätstest"
  /FOOTNOTE "Datenbasis: Beobachtungen".
GRAPH
  /SCATTERPLOT (BIVAR) = Größe WITH ZRE_1
  /MISSING LISTWISE
  /TITLE= "Test auf Varianzungleichheit und Ausreißer (1)"
  /FOOTNOTE "Datenbasis: UV-Beobachtungen, AV-Residuen".
GRAPH
  /SCATTERPLOT (BIVAR) = ZPR_1 WITH ZRE_1
  /MISSING LISTWISE
  /TITLE= "Test auf Varianzungleichheit und Ausreißer (2)"
  /FOOTNOTE "Datenbasis: UV-Schätzer, AV-Residuen".
```

Als erstes erfolgt die grafische Residuenanalyse anhand der drei angeforderten Streudiagramme. Die /SCATTERPLOT-Grafiken in der REGRESSION-Syntax sollten mit den Grafiken des GRAPH-Aufrufes übereinstimmen, da sie auf den Daten desselben Modells basieren. Anschließend werden nur die GRAPH-Diagramme dargestellt, obwohl die Analysen auch uneingeschränkt für die /SCATTERPLOT-Diagramme gelten. Das erste Diagramm enthält die Werte der unabhängigen Variablen auf der x-Achse und die Werte der abhängigen Variablen sind auf der y-Achse aufgetragen.

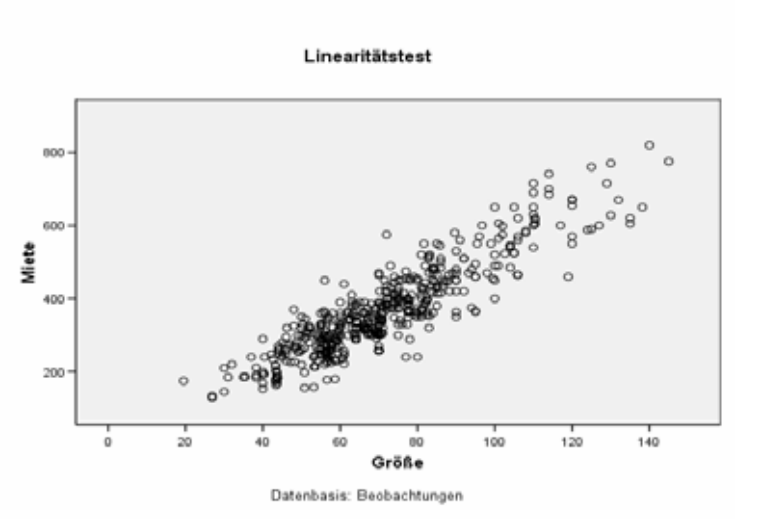


Abbildung 7: Streudiagramm für Linearitätstest

Der Zusammenhang zwischen den beiden Variablen ist eindeutig als linear zu identifizieren. Somit ist die Grundannahme der Linearität gegeben und der Pearson-Korrelationskoeffizient kann zur Berechnung des Zusammenhangs als geeignet angesehen werden.

Des Weiteren kann dem Diagramm entnommen werden, dass die Linie eindeutig nicht als Parallele zur x-Achse verläuft (die Steigung ist somit ungleich NULL); daraus kann der Schluss gezogen werden, dass die Variable „Größe“ eindeutig einen Einfluss auf die Variable „Miete“ ausübt.

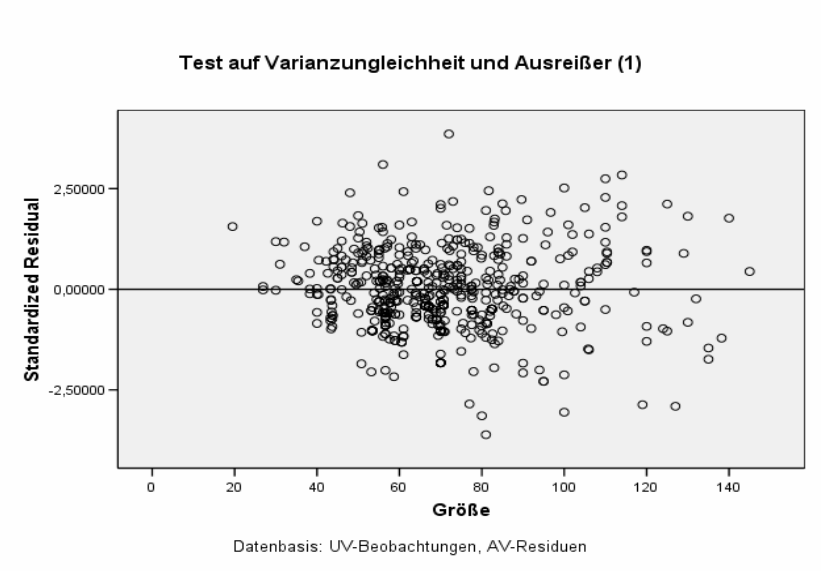


Abbildung 8: Streudigramm für Test auf Varianzungleichheit und Ausreißer (1)

Ein Hinweis auf Heteroskedastizität läge vor, wenn die Streuung der standardisierten Residuen mit zunehmenden x-Werten immer größer werden würde. In dem obigen Diagramm ist zu erkennen, dass die Streuung der Residuen von „Miete“ bei zunehmenden Werten der Variable „Größe“ nicht konstant bleiben. Bei Zunahme der Variablen „Größe“ streuen die Residuen breiter um die Nulllinie. Dagegen scheinen die Anteile der Werte oberhalb und unterhalb der Nulllinie ausgewogen verteilt zu sein und es ist keine Systematik in der Streuung erkennbar.

Auch das dritte Streudiagramm dient der Überprüfung der Varianzheterogenität und untersucht den Datensatz auf Ausreißer. Diesmal erfolgt die Diagnose anhand der geschätzten (vorhergesagten) Werte und der standardisierten Residuen der abhängigen Variablen „Miete“. Wie in dem Diagramm zuvor, scheint auch hier die Streuung um die Nulllinie ausgewogen und zufällig zu sein. Es lässt sich aber auch hier eine verstärkte Streuung der Residuen bei Zunahme der geschätzten Werte feststellen, so dass Heteroskedastizität nicht ausgeschlossen werden kann.

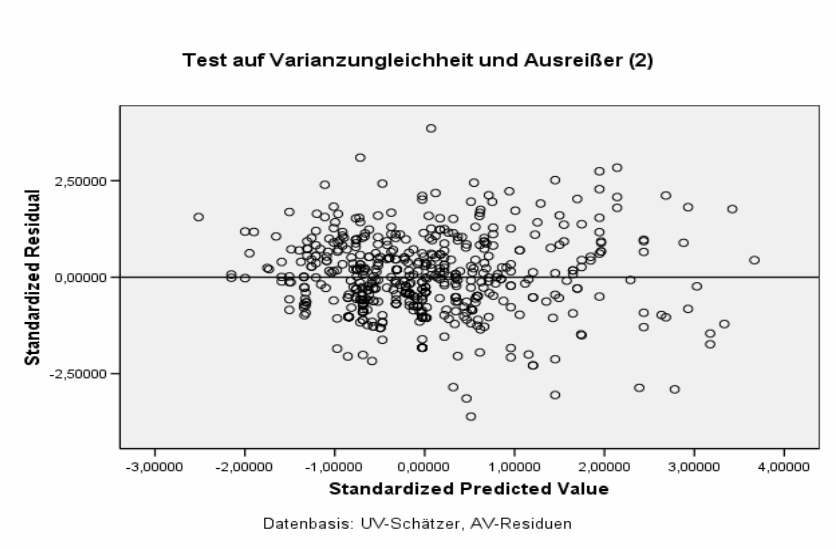


Abbildung 9: Streudigramm für Test auf Varianzungleichheit und Ausreißer (2)

Die Ergebnisse der nachfolgenden Regressionsanalyse können nach der grafischen Residuenanalyse, abgesehen von einer möglichen Heteroskedastizität, als vermutlich auf einem gut angepassten, linearen Modell basierend bewertet und interpretiert werden. Ob Homoskedastizität vorliegt, wäre ggf. noch genauer zu prüfen.

Deskriptive Statistiken

	Mittelwert	Standard-abweichung	N
Miete	360,23	119,742	555
Größe	70,55	20,294	555

Tabelle 12: SPSS-Ausgabe der deskriptiven Statistiken

Die erste Tabelle „Deskriptive Statistiken“ zeigt für die gültigen Fälle (jew. N=555) Mittelwerte und Standardabweichungen an.

Korrelationen

		Miete	Größe
Korrelation nach Pearson	Miete	1,000	,899
	Größe	,899	1,000
Signifikanz (einseitig)	Miete	.	,000
	Größe	,000	.
N	Miete	555	555
	Größe	555	555

Tabelle 13: SPSS-Tabelle der Korrelationen

Der Tabelle „Korrelationen“ kann entnommen werden, dass zwischen den untersuchten Variablen eine relativ hoher Zusammenhang besteht (0,899, p=0,000). Der Korrelationskoeffizient kann Werte zwischen -1 und 1 annehmen und je höher sein Betrag, desto stärker ist der Zusammenhang zwischen den Variablen.

Aufgenommene/Entfernte Variablen(b)

Modell	Aufgenommene Variablen	Entfernte Variablen	Methode
1	Größe(a)	.	Eingeben

a Alle gewünschten Variablen wurden aufgenommen.
b Abhängige Variable: Miete

In diesem SPSS-Output-Teil werden für die entsprechend ausgewählte Methode die Statistiken für die bei jedem Schritt aufgenommenen bzw. wieder entfernten Variablen angezeigt. Den Angaben unterhalb der Tabelle kann entnommen werden, dass alle Variablen aufgenommen wurden (hier: Größe) und dass „Miete“ die abhängige Variable im Modell ist.

Modellzusammenfassung(b)

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Durbin-Watson-Statistik
1	,899(a)	,808	,808	52,480	1,232

a Einflussvariablen : (Konstante), Größe
b Abhängige Variable: Miete

Tabelle 14: SPSS-Output der Modellbeurteilung

Die Tabelle „Modellzusammenfassung“ gibt u.a. die Werte des Bestimmtheitsmaßes (R^2) und des korrigierten Bestimmtheitsmaßes an. Des Weiteren werden noch der Standardfehler des Schätzers und die Durbin-Watson-Statistik mit ausgegeben. Der Index-Wert zur Prüfung auf Autokorrelation liegt in diesem Modellansatz bei 1,232, was auf Autokorrelation hindeuten kann.[48] In diesem Fall hat das Modell ein R^2 von 80,8%, was mit gut zu bewerten ist. D.h. fast 81% der Variation in der abhängigen Variablen können durch das Modell erklärt werden.

ANOVA(b)

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	6420304,9	1	6420304,9	2331,165	,000(a)
	Residuen	1523027,4	553	2754,118		
	Gesamt	7943332,3	554			

a Einflussvariablen : (Konstante), Größe
b Abhängige Variable: Miete

Tabelle 15: SPSS-Ausgabe der Varianzanalyse

In den Spalten der Tabelle des fünften SPSS-Outputteils ist neben den Freiheitsgraden (df) auch die Quadratsummenzerlegung angegeben. Außerdem kann die Schätzung der Restvarianz $s^2 = 2754{,}118$ abgelesen werden.

In der vierten Spalte wird ein „globaler F-Test auf Regression“ (Signifikanztest) durchgeführt. Wie in der fünften Spalte abzulesen ist, liegt Signifikanz vor, d.h. zwischen der unabhängigen

[48] In diesem Fall liegt der niedrige Wert an der Sortierung der Fälle im Datensample. Bei Querschnittsanalysen ist Autokorrelation im Regelfall nicht gegeben.

Variablen „Größe“ und der abhängigen Variablen „Miete“ besteht ein linearer Zusammenhang.

Koeffizienten(a)

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	T	Signifikanz
1	(Konstante)	-14,007	8,065		-1,737	,083
	Größe	5,305	,110	,899	48,282	,000

a Abhängige Variable: Miete

Tabelle 16: SPSS-Output der Koeffizientenstatistik

Die Tabelle „Koeffizienten“ gibt für die Prädiktoren des geschätzten Modells jeweils den nicht standardisierten Regressionskoeffizienten B (5,305), den dazugehörigen Standardfehler (0,110) und den standardisierten Regressionskoeffizient Beta (0,899) aus. Des Weiteren werden die T-Statistik und ihre Signifikanz angegeben. In diesem Beispiel ist das Ergebnis des t-Tests zum Niveau $\alpha = 0{,}1$ signifikant, d.h. der jeweils geprüfte Koeffizient ist signifikant von Null verschieden

Residuenstatistik(a)

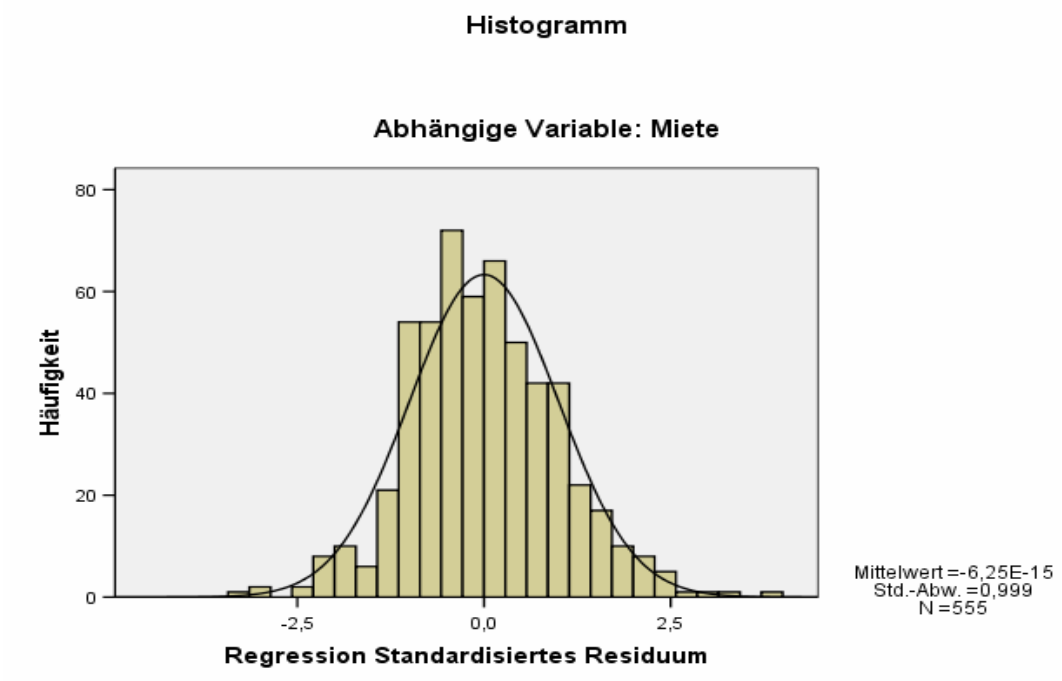
	Minimum	Maximum	Mittelwert	Standard-abweichung	N
Nicht standardisierter vorhergesagter Wert	89,43	755,15	360,23	107,652	555
Standardisierter vorhergesagter Wert	-2,515	3,668	,000	1,000	555
Standardfehler des Vorhersagewerts	2,228	8,477	2,983	1,013	555
Korrigierter Vorhersagewert	88,28	754,62	360,21	107,658	555
Nicht standardisierte Residuen	-170,355	207,081	,000	52,432	555
Standardisierte Residuen	-3,246	3,946	,000	,999	555
Studentisierte Residuen	-3,250	3,950	,000	1,001	555
Gelöschtes Residuum	-170,729	207,457	,014	52,661	555
Studentisierte ausgeschlossene Residuen	-3,278	4,003	,000	1,004	555
Mahalanobis-Abstand	,000	13,458	,998	1,730	555
Cook-Distanz	,000	,056	,002	,005	555
Zentrierter Hebelwert	,000	,024	,002	,003	555

a Abhängige Variable: Miete

Tabelle 17: SPSS-Ausgabe der Residuenstatistik

In der Tabelle „Residuenstatistik“ werden deskriptive Statistiken für die vorhergesagten Werte, die Residuen und für die Hebel- und Einflusswerte ausgegeben. Hinweise auf eine nicht optimale Modellangemessenheit können hier z.B. betragsmäßig hohe Werte der Residuen

geben (vgl. Minimum, Maximum). Der Tabelle kann aber nicht entnommen werden, ob die Residuen einer Normalverteilung folgen. Diese Information kann den nachfolgenden Diagrammen entnommen werden, die in der REGRESSION-Syntax angefordert wurden.



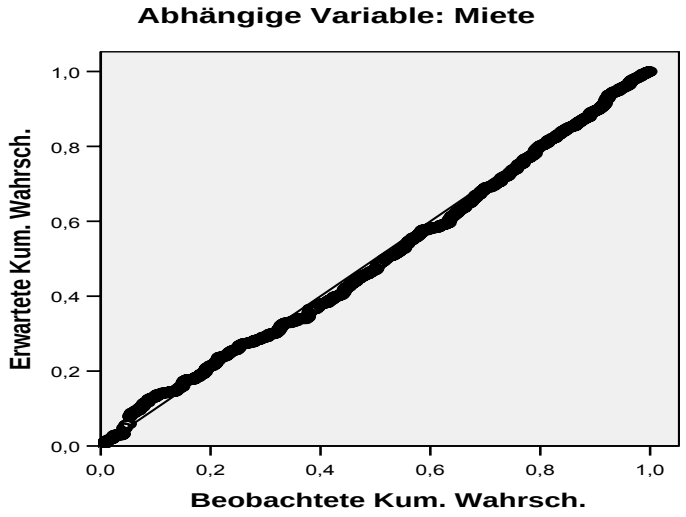


Abbildung 10: Histogramm und P-P-Plot der standardisierten Residuen

Dem Histogramm kann entnommen werden, dass die standardisierten Residuen der abhängigen Variablen ungefähr einer Normalverteilung folgen. Das P-P-Diagramm zeigt, dass die standardisierten Residuen der abhängigen Variablen in etwa auf der eingezeichneten Linie der Referenzverteilung (Normalverteilung) liegen. Die Annahme einer Normalverteilung der standardisierten Residuen kann als erfüllt angesehen werden.

Die nachfolgenden drei Streudiagramme wurden über den Unterbefehl /SCATTERPLOT= angefordert. Die Auswertung wurde bereits am Anfang bei der Residuenanalyse erläutert und soll nicht noch einmal erklärt werden.

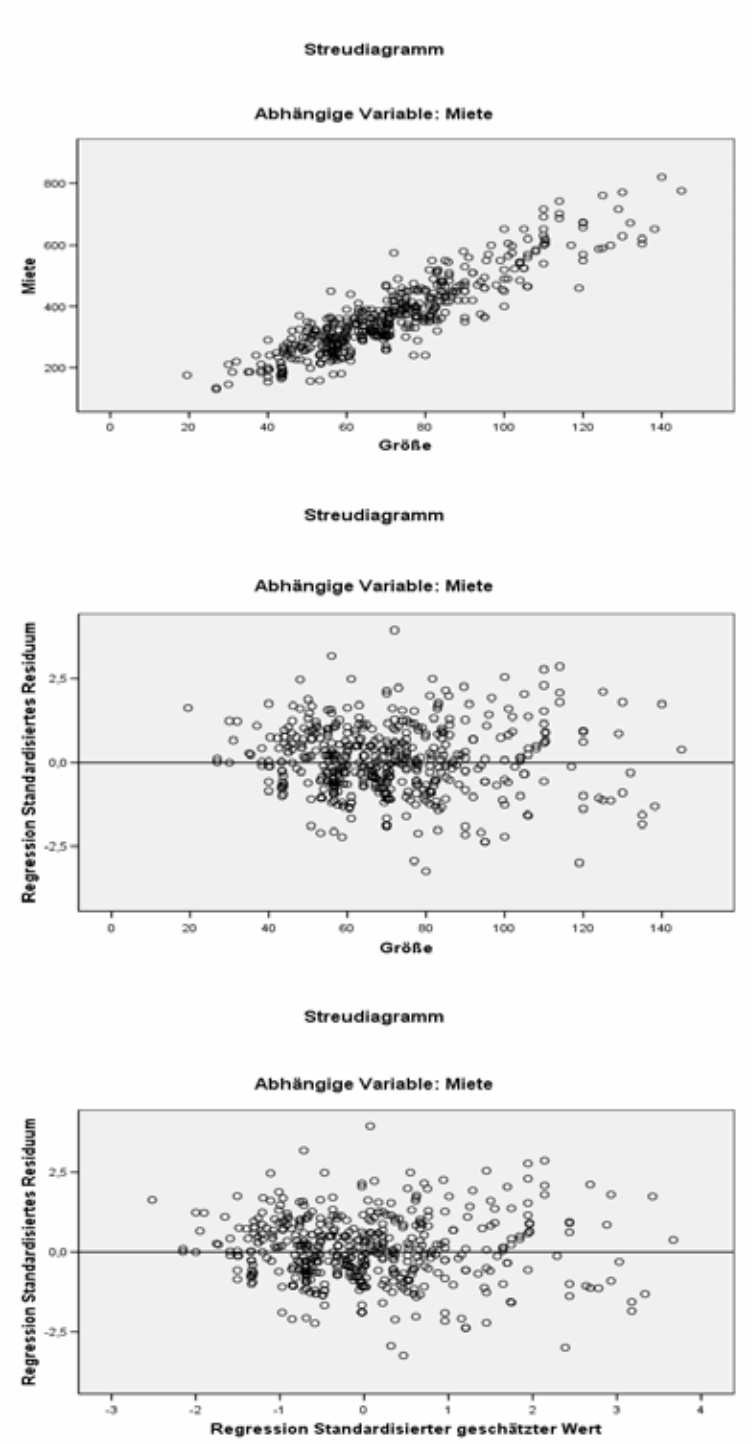


Abbildung 11: Streudiagramme

4 Datenerhebung

In Deutschland gibt es bisher keine Quelle, die bundesweit über das Niveau und die Entwicklung von Mietpreisen auf Ebene von Städten und Gemeinden berichtet. Inzwischen werden zwar Zeitungsannoncen deutschlandweit erfasst und ausgewertet, allerdings enthalten diese Angebotsmieten und keine Vertragsmieten. Insbesondere in entspannten Wohnungsmärkten kann es zwischen diesen beiden Größen erhebliche Unterschiede geben.

Auch die regionale Vergleichbarkeit der Mietpreise durch die Auswertung von Mietspiegeln gestaltet sich als schwierig, da diese zum einen nicht für alle Städte und Gemeinden Deutschlands vorliegen und sich zum anderen in der Datenqualität (je nach Erhebungsmethode) sowie in der Aufbereitung und Darstellung deutlich unterscheiden.

Die Grundlage für die in dieser Diplomarbeit durchgeführten Untersuchungen bilden daher zwei unterschiedliche Datensätze, die sich auf die o.g. Datengrundlagen stützen:

- Erhebung von Bestandsmieten durch Mietspiegelauswertungen
- Mietpreisinformationen des Immobilienverband Deutschlands (Neuvertragsmieten)

Im Folgenden wird zunächst die Erhebung von Bestandsmieten durch die Auswertung von Mietspiegeln erläutert. Danach folgt ein kurzer Exkurs über die Bedeutung und Arten von Mietspiegeln bevor dann der Datensatz des Immobilienverband Deutschlands (IVD) vorgestellt wird. Zum Schluss dieses Kapitels werden die Variablen beschrieben, die in die Analyse des Einflusses regionaler Faktoren auf den Wohnungsmietpreis einfließen sollen.

4.1 Mietpreisinformationen durch Auswertung von Mietspiegeln

Um an die für die Untersuchung benötigten Daten zu gelangen, wurde von der Autorin im Rahmen eines Praktikums am IWU eine Mietspiegel-Datenbank erstellt, in der die deutschlandweit vorhandenen Mietspiegel gesammelt wurden. Diese konnten dann für die nachfolgenden Analysen ausgewertet werden. Um die daraus gewonnenen Mietpreisinformationen vergleichbar zu machen, mussten zunächst typische Referenzwohnungen definiert werden, für welche Mietpreise berechnet werden sollen.

4.1.1 Standardwohnungstypen

Für die Definition der Referenzwohnungen wurde ein Datensatz des IWU verwendet, der Wohnungsdaten aus Darmstadt und Frankfurt a. Main enthält. Diese Datenbasis wurde im Rahmen von Befragungen aus den Jahr 2007 für die Mietspiegelerstellung in den beiden Städten gewonnen.

Die Bestimmung der Referenzwohnungstypen erfolgt mit Hilfe der Statistik-Software SPSS® 14.0. Dazu wurden aus den insgesamt 3609 Beobachtungen des Basisdatensatzes die vier am häufigsten vorkommenden Wohnungstypen und ein Einfamilienhaus ausgewählt (im weiteren Verlauf werden diese als „Standardwohnungstypen“ bezeichnet).

StdWhg3 Einfamilienhaus / Etagenwohnung mit Küche - StdWhg 3

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	100 Einzelzimmer	3	,1	,1	,1
	110 1-Z-Whg + Küche	153	4,2	4,3	4,3
	120 Appart + IntKü	20	,6	,6	4,9
	130 Appart + Kochnische	72	2,0	2,0	6,9
	140 Appart + Kochgeleg.	23	,6	,6	7,5
	210 2-Z-Whg + Küche	1172	32,5	32,6	40,1
	230 2-Z-Whg + Kochnische	60	1,7	1,7	41,8
	240 2-Z-Whg + Kochgeleg.	17	,5	,5	42,2
	310 3-Z-Whg + Küche	1455	40,3	40,4	82,7
	330 3-Z-Whg + Kochnische	22	,6	,6	83,3
	410 4-Z-Whg + Küche	410	11,4	11,4	94,7
	510 5+Z-Whg + Küche	93	2,6	2,6	97,3
	1010 EFH + Küche	98	2,7	2,7	100,0
	Gesamt	3598	99,7	100,0	
Fehlend	-6 Unplaus. Whg-Typen	11	,3		
Gesamt		3609	100,0		

Tabelle 18: Häufigkeitstabelle Wohnungstypen[49]

Im Anschluss daran wurden die fünf Standardwohnungstypen nach den unterschiedlichen Baualterklassen, die in Deutschland typischerweise vertreten sind, aufgeteilt. Des Weiteren wurde eine Unterteilung in unterschiedliche Lagekategorien (sehr gute und gute Wohnlage, gehobene und mittlere Wohnlage, einfache und sehr einfache Wohnlage) vorgenommen, womit sich dann eine Aufteilung in neun Baualtersklassen und drei Lagen ergab (vgl. Tab. 20).

Die 2-Z-Whg + Küche wurde bewusst aus der Baualtersklasse 1995 – 2001 ausgewählt, da auch eine neuere Baualtersklasse (Baujahr zwischen 1995 – 2007) mit untersucht werden soll. Die Gruppe 2002 – 2007 war mit zu wenigen Fällen besetzt, um diese zu berücksichtigen. Aus diesem Grund wurde die Baualtersklasse 1995 – 2001 mit der 2-Z-Whg als stärksten Repräsentanten ausgewählt. Alle anderen Standardwohnungstypen wurden der Baualtersklasse zugeordnet, in der sie im Datensatz am häufigsten auftraten. Die 3-Z-Whg + Küche wurde aus der Klasse gewählt, in der sie am zweitstärksten repräsentiert ist, da die Baualtersklasse 1958 – 1968 bereits durch die 1-Z-Whg + Küche in die Untersuchung mit einbezogen ist und keine Baualtersklasse doppelt vertreten sein soll.

[49] Quelle: Eigene Berechnung auf Basis des Datensatzes „Kombi.sav“.

StdWhg3 Einfamilienhaus / Etagenwohnung mit Küche - StdWhg 3 * BA Baualter des Gebäudes, gruppiert * Lage Kreuztabelle

Lage			BA Baualter des Gebäudes, gruppiert (nach BAS3)									Gesamt
			Bis 1918	1919 - 1948	1949 - 1957	1958 - 1968	1969 - 1977	1978 - 1984	1985 - 1994	1995 - 2001	2002 - 2007	
Sehr gute(Ffm+Da) und gute(Ffm) Wohnlage (GA-Ausschuss)	Einfamilienhaus / Etagenwohnung mit Küche – StdWhg 3	1-Z-Whg + Küche	0	0	3	0	0	0	0	0	0	3
		2-Z-Whg + Küche	5	8	14	13	1	3	0	3	0	47
		3-Z-Whg + Küche	12	12	19	21	10	2	1	8	1	86
		4-Z-Whg + Küche	6	3	13	8	5	0	1	1	0	37
		EFH + Küche	0	2	1	2	1	2	0	0	1	9
	Gesamt		23	25	50	44	17	7	2	12	2	182
Gehobene Wohnlage(Ffm) und mittlere(Ffm+Da) Wohnlage (GA-Ausschuss)	Einfamilienhaus / Etagenwohnung mit Küche – StdWhg 3	1-Z-Whg + Küche	10	7	33	46	10	7	2	2	0	117
		2-Z-Whg + Küche	96	135	237	224	102	40	32	33	7	906
		3-Z-Whg + Küche	171	122	298	309	92	46	25	27	24	1114
		4-Z-Whg + Küche	75	32	65	55	34	21	6	12	8	308
		EFH + Küche	6	19	12	15	10	5	5	5	3	80
	Gesamt		358	315	645	649	248	119	70	79	42	2525
Einfache(Ffm+Da), sehr einfache(Ffm) bzw. Industriegebiet(Da) Wohnlage (GA-Ausschuss)	Einfamilienhaus / Etagenwohnung mit Küche – StdWhg 3	1-Z-Whg + Küche	2	3	4	16	1	1	1	0	0	28
		2-Z-Whg + Küche	15	22	31	69	12	16	8	2	1	176
		3-Z-Whg + Küche	17	42	39	85	21	8	4	3	2	221
		4-Z-Whg + Küche	14	8	9	21	6	2	0	0	0	60
		EFH + Küche	1	3	1	0	0	0	0	0	1	6
	Gesamt		49	78	84	191	40	27	13	5	4	491

Tabelle 19: Kreuztabelle - Standardwohnung nach Baualter des Gebäudes und Lage[50]

[50] Quelle: Eigene Berechnung auf Basis des Datensatzes „Kombi.sav“.

Danach wurden für jeden Wohnungstyp die mittleren Wohnflächen berechnet. Dabei ergaben sich die folgenden Wohnungsgrößen:

	Mittelwert	Standardabweichung
1-Z-Whg + Küche	34,85	6,518
2-Z-Whg + Küche	57,84	7,958
3-Z-Whg + Küche	69,04	11,202
4-Z-Whg + Küche	101,87	18,516
EFH + Küche	96,88	45,692

Tabelle 20: Mittlere Wohnflächen[51]

Um die Struktur des Wohnungsmarktes in Deutschland möglichst realitätsnah darzustellen, wurde die 4-Z-Whg als sanierter Altbau und die 3-Z-Whg als sanierter Nachkriegsbau definiert. Um auch unsanierte Wohnungen mit zu berücksichtigen, wurden dementsprechend die 1-Z-Whg und das Einfamilienhaus als unsanierte Gebäude angenommen. Außerdem wurden der 3-Z-Whg und der 4-Z-Whg ein Balkon zugeteilt und das Einfamilienhaus als Reihenhaus mit Gartennutzung definiert. In der neueren Baualtersklasse 1995 – 2001 ist eine Zentralheizung Standard. Da in modernisierten Alt- und Nachkriegsbauten inzwischen ebenfalls häufig Zentralheizungen zu finden sind, werden diese als Beheizungsart für die 3-Z-Whg und die 4-Z-Whg bestimmt. Gaseinzelöfen und Fernwärme kommen in der Praxis in älteren Wohngebäuden vor und sollen bei der Auswertung der Mietspiegel ebenfalls berücksichtigt werden. Die Gaseinzelöfen werden deshalb dem Reihenhaus (Baualtersklasse 1919 – 1948) zugeordnet, die Fernwärme der 1-Z-Whg aus der Baualtersklasse 1958 – 1968.

Für die weiteren Untersuchungen werden demnach folgende Standardwohnungstypen in mittlerer Lage zu Grunde gelegt:

[51] Quelle: Eigene Berechnung auf Basis des Datensatzes „Kombi.sav“.

	Baualter	*m²*	*Zustand*	*Balkon / Garten*	*Heizung*	*Bsp. Gebäudetyp*
1-Z-Whg + Küche	1958 – 1968	34	Unsaniert	-	Fernwärme (Zentralheizung)	
2-Z-Whg + Küche	1995 – 2001	58	-	Balkon	Zentralheizung	
3-Z-Whg + Küche	1949 – 1957	69	Saniert	Balkon	Zentralheizung	
4-Z-Whg + Küche	bis 1918	102	Saniert	-	Zentralheizung	
EFH + Küche (Reihenhaus)	1919 - 1948	97	Unsaniert	Garten	Gaseinzelöfen	

Tabelle 21: Überblick Standardwohnungstypen[52]

4.1.2 Mietspiegelauswertung

Die Mietspiegeldatenbank des IWU enthält 359 aktuelle Mietspiegel aus Städten und Gemeinden in Deutschland[53]. Im Rahmen der vorliegenden Untersuchungen wurden diese von der Autorin im Rahmen eines Praktikums am IWU einzeln ausgewertet und für jeden Standardwohnungstyp die entsprechende Nettokaltmiete bestimmt.

Der Vergleich der Mietspiegel wurde vor allem durch die unterschiedlichen Merkmalsdefinitionen und Merkmalsausprägungen erschwert. Deshalb sind folgende vereinfachenden Annahmen getroffen worden:

- unter saniert wird **vollmodernisiert** verstanden
- alle Wohnungen sind mit einem **innen liegendem WC** ausgestattet
- alle Wohnungen sind mit einem **Bad mit Wanne / Dusche** ausgestattet.

Eine weitere Merkmalsdifferenzierung für die Wohnungstypen (z.B. Fußbodenbelag, Gäste-WC, Gegensprechanlage,…) wurde nicht vorgenommen, um die Vergleichbarkeit zu gewährleisten. Da Wohnungen in Einfamilienhäusern häufig nicht in den Anwendungsbereich der

[52] Grafiken: IWU

[53] Stand: 25.05.2009

Mietspiegel fallen, muss später untersucht werden, ob eine ausreichende Datenmenge vorhanden ist oder ob die Einfamilienhäuser von den weiteren Analysen ausgeschlossen werden müssen.

Die Auswertung der Mietspiegel konnte nicht ohne weiteres durch ein VBA – Programm o.ä. automatisiert werden, da sich die Mietspiegel untereinander sehr stark unterscheiden. Dies ist zum einen dadurch bedingt, dass die Erstellungsmethoden differieren, zum anderen wirken sich die betrachteten Merkmale unterschiedlich stark auf den Mietpreis aus. Die Mietspiegel wurden deshalb per Hand ausgewertet und die monatlichen Nettokaltmieten in Euro und in Euro/m² in eine Excel – Tabelle übertragen.

Ein großer Unterschied existiert auch bei der Ausweisung der Mietpreise: Einige Mietspiegel geben Durchschnittspreise für die Wohnungsmieten an, andere weisen dagegen Mietpreisspannen aus. Für die nachfolgenden Analysen werden die Durchschnittspreise herangezogen. Bei fehlender Angabe des Durchschnittspreises wurde dieser durch das arithmetische Mittel aus oberer und unterer Grenze der Preisspanne berechnet.

4.2 Exkurs: Bedeutung und Arten von Mietspiegeln

4.2.1 Aufgaben des Mietspiegels[54]

Ein Mietspiegel ist eine Übersicht über die ortsübliche Vergleichsmiete, die von der Gemeinde oder von Interessenvertretern der Vermieter und der Mieter gemeinsam erstellt oder anerkannt worden ist. Die ortsübliche Vergleichsmiete wird nach der gesetzlichen Definition aus den üblichen Entgelten gebildet, die in der Gemeinde oder einer vergleichbaren Gemeinde für Wohnraum vergleichbarer Art, Größe, Ausstattung, Beschaffenheit und Lage in den letzten vier Jahren vereinbart oder geändert worden sind. Mietspiegel gelten für den freien Wohnungsmietmarkt und haben folgende wesentliche Aufgaben:

- Sie sorgen für Markttransparenz und helfen so, Streitigkeiten zwischen Mietern und Vermietern außergerichtlich zu regeln.
- Sie geben Anhaltspunkte für die ortsübliche Vergleichsmiete in einem Gebiet und begrenzen damit Mieterhöhungen für Bestandsmietverträge.

4.2.2 Arten und Verbreitung von Mietspiegeln

Seit Mitte der 70er Jahre haben Mietspiegel in Deutschlands Städten und Gemeinden eine zunehmende Verbreitung gefunden und im Jahr 2006 führten knapp 6% der Kommunen einen Mietspiegel. Dabei hat Nordrhein-Westfalen als bevölkerungsreichstes Bundesland auch den höchsten Anteil an Mietspiegeln (vgl. Abbildung 45 im Anhang). Außerdem ist zu beobachten, dass sich das Vorhandensein eines Mietspiegels vor allem auf die Kernstädte und

[54] Die Ausführungen in den Abschnitten 4.2.1, 4.2.2 und 4.2.3 sind in einigen Teilen identisch mit dem Text der Veröffentlichung: Hinweise zur Erstellung von Mietspiegeln, Bundesministerium für Verkehr, Bau- und Wohnungswesen (Hrsg.)

zentralen Orte in Agglomerationsräumen konzentriert[55]. In über 70% der Städte bis 100.000 Einwohner und in 40% der Kommunen bis 50.000 Einwohner sind Mietspiegel vorhanden.

Das Gesetz unterscheidet Mietpreisübersichten nach ihren Rechtsfolgen in einfache und qualifizierte Mietspiegel: Zunächst ist nach der gesetzlichen Definition jede Übersicht über die ortsübliche Vergleichsmiete, die von der Gemeinde oder von Interessenvertretern der Vermieter und der Mieter gemeinsam erstellt oder anerkannt worden ist, ein Mietspiegel. An Mietspiegel, die bestimmte Anforderungen erfüllen, werden besondere Rechtsfolgen geknüpft. Solche Mietspiegel werden als qualifizierte Mietspiegel bezeichnet. Mietspiegel, die diese Anforderungen nicht erfüllen, nennt man einfache Mietspiegel.

Ein qualifizierter Mietspiegel muss gemäß § 558d BGB folgende Anforderungen erfüllen:

- Er muss nach anerkannten wissenschaftlichen Grundsätzen ermittelt worden sein **und**
- er muss von der Gemeinde oder von Interessenvertretern der Vermieter und Mieter anerkannt worden sein.

Außerdem muss ein qualifizierter Mietspiegel im Abstand von zwei Jahren an die Marktentwicklung angepasst und nach vier Jahren neu erstellt werden.

Für die Modellierung der ortsüblichen Vergleichsmiete bzw. der Preisübersichten in Mietspiegeln haben sich nach anerkannten wissenschaftlichen Grundsätzen (§ 558d II BGB) zwei gängige statistische Verfahren herausgebildet: die Tabellenmethode und die Regressionsmethode.

4.2.3 Tabellen- und Regressionsmietspiegel

Die Entscheidung, ob ein Tabellen- oder Regressionsmietspiegel erstellt werden soll, ist grundsätzlich freigestellt. Beide Verfahren werden in der Praxis für sich allein oder auch in Kombination angewandt und bieten verschiedene Vor- und Nachteile, auf die hier nicht näher eingegangen wird. Der Erstellungsprozess ist bei beiden Methoden ähnlich: Beide basieren auf einer repräsentativen empirischen Datenerhebung und bei beiden wird der Einfluss einzelner Wohnwertmerkmale auf die Miethöhe mit statistischen Verfahren untersucht.

4.2.3.1 Tabellenmietspiegel

Bei der Tabellenmethode werden alle Mietwerte einer Kategorie von Wohnungen zu einer Gruppe zusammengefasst. Die Kategorien werden durch Kombinationen von Wohnwertmerkmalen (z.B. Altbau, mit Bad, Größe unter 40 m², einfache Wohnlage) bestimmt und in einem Mietspiegelfeld abgebildet. Aus den in der Erhebung ermittelten Mietwerten jedes Mietspiegelfeldes wird dann ein Mittelwert berechnet, der die Vergleichsmiete dieser Kategorie von Wohnungen repräsentiert. Bei der Bildung der Kategorien müssen diejenigen Wohnwertmerkmale identifiziert werden, die den statistisch bedeutsamsten Einfluss auf die Miethöhe haben.

[55] Vgl. BBR S. 99 und [5]

Die Stichprobengröße wird über die Zahl der zu besetzenden Tabellenfelder bestimmt. Dabei sollte eine minimale Feldbesetzung von 30 Wohnungen je Mietspiegelfeld berücksichtigt werden. Reicht die Besetzung einzelner Tabellenfelder nicht aus, um bestimmte Teilmärkte repräsentativ abzubilden, kann die ortsübliche Vergleichsmiete dennoch ausgewiesen werden. Allerdings ist durch eine deutlich sichtbare Kennzeichnung unter Angabe der Fallzahl darauf hinzuweisen, dass diese Tabellenfelder nicht die Anforderungen eines qualifizierten Mietspiegels erfüllen.

Angaben in €/m²

Lage	Wohnfläche, 6 Größenklassen	Baualter des Gebäudes, gruppiert									
		Bis 1918	1919 - 1948	1949 - 1957	1958 - 1968	1969 - 1977	1978 - 1984	1985 - 1994	1995 - 2001	2002 - 2007	Insgesamt
Insgesamt	1 - 40m²	* 9,27	8,84	9,16	8,90	9,05	** 9,02	** 9,26	** 10,01	** 11,32	9,04
	41 - 65m²	7,50	6,87	7,11	7,19	7,31	7,71	7,41	8,32	** 9,42	7,23
	66 - 85m²	7,05	6,43	6,50	6,57	6,59	7,35	* 7,32	8,45	* 8,93	6,78
	86 - 105m²	7,03	6,05	6,75	6,62	7,21	* 6,63	** 7,21	* 8,20	** 8,41	6,88
	106 - 125m²	7,56	* 6,30	* 5,94	* 6,73	* 7,10	** 7,21	** 6,74	** 8,44	** 10,22	7,02
	über 126m²	7,34	** 6,86	** 5,77	** 8,29	** 8,55	** 7,42	** 6,63	** 11,32	** 9,83	7,59
	Insgesamt	7,34	6,77	7,02	7,13	7,26	7,45	7,43	8,46	9,15	7,20

** bis zu 14 Fälle * 15 - 29 Fälle

Tabelle 22: Vereinfachte Darstellung Tabellenmietspiegel[56]

4.2.3.2 Regressionsmietspiegel

Die Regressionsanalyse ist ein anerkanntes Verfahren, die Abhängigkeit eines bestimmten Merkmales (z.B. die Nettomiete) von anderen, unabhängigen Merkmalen (z.B. Größe der Wohnung, Ausstattung, usw.) zahlenmäßig zu erfassen. Die Grundlage hierfür bildet das Regressionsmodell, mit dessen Hilfe die Werte des abhängigen Merkmals aus den Werten der unabhängigen Merkmale berechnet werden können.

Bei dieser Methode fällt der Modellbildung, also der Entwicklung der Regressionsgleichung, eine zentrale Bedeutung zu. Im folgenden Beispiel wird die Quadratmetermiete als erklärende Variable gewählt. Bei jeder Neuerstellung des Mietspiegels ist eine Anpassung des Regressionsmodells an die konkrete Marktsituation erforderlich, da kein allgemeines und dauerhaft gültiges Modell zur Wohnungsmarktbeschreibung bekannt ist. Die Stichprobengröße liegt i.d.R. unter der eines Tabellenmietspiegels, da die Informationen der gesamten Stichprobe ausgenutzt werden können und eine Teilmengenbildung, wie beim Tabellenmietspiegel, entfällt.

Das Ergebnis der Regressionsanalyse wird üblicherweise in folgender Form dargestellt:

[56] Quelle: Eigene Berechnungen auf Basis des Datensatzes „Kombi.sav“.

- Die Konstante und wichtige (insbesondere metrische) Größen werden umgerechnet und in einer sog. Basistabelle dargestellt.
- Die Koeffizienten werden üblicherweise als Zu- und Abschläge ausgewiesen.

SPSS-Output:

	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	F	Signifikanz
	B	Standardfehler	Beta		
(Konstante)	10,548	0,353		894,624	0,000
Wfl Wohnfläche in m² (20%-Limit)	-0,082	0,007	-1,002	140,094	0,000
Wfl_sq	0,000	0,000	0,614	77,128	0,000
Appartement mit integrierter Küche oder Kochnische/-gelegenheit	0,476	0,271	0,041	3,084	0,079
1-Z-Wohnung mit Küche (Koch-/Wohnküche)	-0,150	0,232	-0,014	0,420	0,517
2-Z-Wohnung mit Küche (Koch-/Wohnküche, integrierte Küche)	-0,207	0,108	-0,046	3,675	0,055
4-Z-Wohnung mit Küche (Koch-/Wohnküche, integrierte Küche)	0,184	0,128	0,029	2,079	0,149
Baualter: 1918 und früher	0,448	0,135	0,072	11,051	0,001
Baualter: 1919 bis 1948	-0,307	0,132	-0,049	5,415	0,020
Baualter: 1958 bis 1968	-0,049	0,107	-0,010	0,209	0,648
Baualter: 1969 bis 1977	-0,167	0,141	-0,025	1,408	0,236
Baualter: 1978 bis 1984	-0,164	0,172	-0,019	0,910	0,340
Baualter: 1985 bis 1994	0,176	0,214	0,016	0,679	0,410
Baualter: 1995 bis 2001	0,911	0,205	0,088	19,786	0,000
Baualter: 2002 bis 2007	1,657	0,314	0,100	27,902	0,000
Einzelöfen aller Art zusammengefasst	-0,648	0,120	-0,098	29,133	0,000
Überhaupt(!) keine Heizung vom Vermieter gestellt	-1,475	0,268	-0,099	30,269	0,000
Sehr gute und gute Wohnlage	0,974	0,152	0,118	41,038	0,000
Einfache, sehr einfache bzw. Industriegebiet Wohnlage	-0,158	0,115	-0,025	1,891	0,169
Kürzere Seite des vom Vermieter gestellten Badezimmers ist mindestens 2,00 Meter breit	0,196	0,080	0,047	6,027	0,014
Kürzere Seite des vom Vermieter gestellten Badezimmers ist mindestens 3,00 Meter breit	0,325	0,196	0,031	2,763	0,097
Badezimmer ist an allen Wänden mindestens halbhoch gekachelt	0,392	0,083	0,086	22,323	0,000
Badewanne und separate Duschwanne im Badezimmer vom Vermieter gestellt	0,650	0,141	0,094	21,365	0,000
Mindestens zwei räumlich getrennte Toiletten innerhalb der Wohnung vorhanden	0,388	0,136	0,060	8,186	0,004
Kleinster Wohnraum größer 15 m² bei 2 oder mehr Wohn-/Schlafräume	0,290	0,089	0,064	10,702	0,001
Größter Wohnraum 25 m² oder größer bei 2 oder mehr Wohn-/Schlafräume	0,491	0,099	0,103	24,480	0,000
In über 50% der Wohn-/Schlafräume Fußboden mit Echtholzparkett, Massivholzdielen, Marmor oder gleichwertigen Natursteinen oder keramischen Bodenfliesen / Platten	0,562	0,103	0,103	29,525	0,000
In mindestens 1 Wohn-, Schlafraum liegen unverkleidete Wasser-, Gas- oder Heizungsleitungen auf Putz	-0,365	0,107	-0,064	11,649	0,001
Umfangreiche Einbauküchenmöbel in der Küche	0,647	0,175	0,067	13,640	0,000
Aufzug in Gebäuden mit bis zu 4 Geschossen	0,625	0,215	0,053	8,466	0,004

a. Abhängige Variable: nmqm Mon. Nettomiete in Euro/m² im Stichmonat

Tabelle 23: Ausschnitt SPSS-Output der Koeffizientenstatistik[57]

Dieser Tabellenausschnitt zeigt die Koeffizienten des Regressionsmodells. Die nicht signifikanten Variablen sind orange markiert. Auf eine nähere Erläuterung des Outputs an dieser Stelle wird verzichtet und auf das Kapitel 3.7 verwiesen.

[57] Quelle: Eigene Berechnungen auf Basis des Datensatzes „Kombi.sav“.

Für dieses Beispiel ergibt sich die folgende Regressionsgleichung, mit welcher dann die Basistabelle errechnet wird:

$$nmqm = 10{,}548 - 0{,}082 \cdot Wfl + 0{,}00026351 \cdot Wfl^2 + 0{,}448 \cdot BA0018 - 0{,}307 \cdot BA1948 + 0{,}911 \cdot BA9501 + 1{,}657 \cdot BA0207$$

Die Angaben *BA0018* bis *BA0207* bezeichnen dabei die unterschiedlichen Baualtersklassen und nehmen die Werte 0 (d.h. trifft nicht zu) oder 1 (d.h. trifft zu) an.

Basistabelle:

Angaben in €/m²

		Baualter				
		vor 1918	1919 bis 1948	1949 bis 1994	1995 bis 2001	2002 bis 2007
Wohnfläche in m²	20	9,46	8,71	9,01	9,92	10,67
	25	9,11	8,36	8,66	9,57	10,32
	30	8,77	8,02	8,33	9,24	9,98
	35	8,45	7,69	8,00	8,91	9,66
	40	8,14	7,38	7,69	8,60	9,35
	45	7,84	7,08	7,39	8,30	9,05
	50	7,55	6,80	7,11	8,02	8,76
	55	7,28	6,53	6,84	7,75	8,49
	60	7,02	6,27	6,58	7,49	8,23
	65	6,78	6,02	6,33	7,24	7,99
	70	6,55	5,79	6,10	7,01	7,76
	75	6,33	5,57	5,88	6,79	7,54
	80	6,12	5,37	5,67	6,59	7,33
	85	5,93	5,17	5,48	6,39	7,14
	90	5,75	5,00	5,30	6,21	6,96
	95	5,58	4,83	5,14	6,05	6,79
	100	5,43	4,68	4,98	5,89	6,64
	105	5,29	4,54	4,84	5,75	6,50
	110	5,16	4,41	4,72	5,63	6,37
	115	5,05	4,30	4,60	5,51	6,26
	120	4,95	4,20	4,50	5,41	6,16
	125	4,86	4,11	4,42	5,33	6,07
	130	4,79	4,03	4,34	5,25	6,00
	135	4,73	3,97	4,28	5,19	5,94
	140	4,68	3,93	4,23	5,14	5,89
	145	4,65	3,89	4,20	5,11	5,86
	150	4,62	3,87	4,18	5,09	5,83
	155	4,62	3,86	4,17	5,08	5,83
	160	4,62	3,87	4,17	5,08	5,83

Abbildung 12: Vereinfachte Darstellung Regressionsmietspiegel - Basistabelle[58]

Ausweisung der Koeffizienten:

Die in Tabelle 23 ermittelten signifikanten, nichtstandardisierten Koeffizienten, die nicht für die Berechnung der Basistabelle herangezogen worden sind, werden in einer Extra-Tabelle als Zu- bzw. Abschlagsmerkmale auf die Basismiete ausgewiesen. Die Beträge werden in Euro angegeben und auf zwei Nachkommastellen gerundet.

[58] Quelle: Eigene Berechnungen auf Basis des Datensatzes „Kombi.sav“.

Sehr gute und gute Wohnlage:	**0,97**
Appartement mit integrierter Küche oder Kochnische/-gelegenheit:	**0,48**
2-Z-Wohnung mit Küche (Koch-/Wohnküche, integrierte Küche):	**-0,21**
Einzelöfen aller Art zusammengefasst:	**-0,65**
Überhaupt(!) keine Heizung vom Vermieter gestellt:	**-1,48**
Kürzere Seite des vom Vermieter gestellten Badezimmers ist mindestens 2,00 Meter breit:	**0,20**
Kürzere Seite des vom Vermieter gestellten Badezimmers ist mindestens 3,00 Meter breit:	**0,33**
Badezimmer ist an allen Wänden mindestens halbhoch gekachelt:	**0,39**
Badewanne und separate Duschwanne im Badezimmer vom Vermieter gestellt:	**0,65**
Mindestens zwei räumlich getrennte Toiletten innerhalb der Wohnung vorhanden:	**0,39**
Kleinster Wohnraum größer 15 m² bei 2 oder mehr Wohn-/Schlafräume:	**0,29**
Größter Wohnraum 25 m² oder größer bei 2 oder mehr Wohn-/Schlafräume:	**0,49**
In über 50% der Wohn-/Schlafräume Fußboden mit Echtholzparkett, Massivholzdielen, Marmor oder gleichwertigen Natursteinen oder keramischen Bodenfliesen / Platten:	**0,56**
In mindestens 1 Wohn-,Schlafraum liegen unverkleidete Wasser-, Gas- oder Heizungsleitungen auf Putz:	**-0,37**
Umfangreiche Einbauküchenmöbel in der Küche:	**0,65**
Aufzug in Gebäuden mit bis zu 4 Geschossen:	**0,63**

Abbildung 13: Zu- und Abschlagsmerkmale[59]

Die Tabelle in Abbildung 12 enthält die durchschnittliche *Basis-Nettomiete* in Euro pro Quadratmeter nach Wohnfläche und Baualter. Zur Ermittlung der Basis-Nettomiete wählt man die Tabellen-Zeile aus, die der zutreffenden auf volle Quadratmeter gerundeten Wohnfläche entspricht. Anschließend wird in der Kopf-Zeile die entsprechende Baualtersklassen-Spalte herausgesucht. Im Schnittpunkt der gewählten „Wohnflächen-Zeile“ und „Baualters-Spalte“ kann man die Basis-Nettomiete ablesen. Danach muss geprüft werden, welche der Merkmale in Abbildung 13 zutreffen. Die ortsübliche Vergleichsmiete (Nettomiete) pro Quadratmeter ergibt sich als Summe aus Basis-Nettomiete und allen zutreffenden Zu- und Abschlägen. Zur Berechnung der ortsüblichen Vergleichsmiete (Nettomiete) für eine konkrete Wohnung muss der Quadratmeterwert mit der Wohnfläche multipliziert werden.

4.2.4 Mietdatenbank[60]

Neben den Tabellen- und Regressionsmietspiegeln wird im Gesetz als weitere Option die Pflege einer Mietdatenbank zur Ermittlung der ortsüblichen Vergleichsmiete genannt. Die einzige Kommune, die eine solche Datenbank betreibt ist Hannover (einschließlich Langenhagen und Laatzen).

In die Mietdatenbank tragen Mieter und Vermieter die Veränderungen ihrer Miethöhe ein. Diese Daten können dann gegen eine Gebühr abgefragt werden, um Vergleichsmieten für Wohnungen zu ermitteln. Allerdings können hierbei keine Aussagen zu der Aktualität der Daten und zu der Vergleichbarkeit der Wohnungen gemacht werden (keine Angaben von

[59] Quelle: Eigene Berechnungen auf Basis des Datensatzes „Kombi.sav“.

[60] Vgl. BBR, S. 101

berücksichtigten Zu- oder Abschlägen). Auch der Stichprobenumfang, der zum Vergleich herangezogen wird, ist nicht bekannt.

4.3 IVD – Datensatz

Der „Immobilienverband Deutschland“ (IVD) ist der größte deutsche Unternehmensverband in der Wohnungswirtschaft. Er entstand im Jahr 2004 als Zusammenschluss aus dem Ring deutscher Makler (RDM) und dem Verband deutscher Makler (VDM). Einmal jährlich erscheint der „IVD Wohn–Preisspiegel“, eine Preisdatensammlung mit Wohnungsmieten für 352 deutsche Städte und Gemeinden. Grundlage für die Preisangaben bilden die jeweils bei Neuvermietung erzielten Mietpreise. Bestandsmieten werden hierbei nicht erfasst (vgl. IVD, S.5f). Die angegebenen Netto-Kaltmieten beziehen sich auf eine ca. 70m² große Wohnung mit ca. 3 Zimmern ohne den öffentlich geförderten Wohnungsbau. Unterteilt wird in einfachen, mittleren und guten Wohnwert und in drei unterschiedliche Baualtersklassen:

- Fertigstellung bis 1948 (Wiedervermietung / Neuvertragsmiete)
- Fertigstellung ab 1949 (Wiedervermietung / Neuvertragsmiete)
- Neubau – Erstbezug (Erstvermietung im Berichtsjahr)

Für diese Diplomarbeit werden die Gebäude mit Fertigstellung ab 1949 mit gutem Wohnwert aus dem IVD-Wohn-Preisspiegel 2006/2007 betrachtet. Der zur Verfügung stehende Datensatz enthält 352 Beobachtungen.

Stadt	Bundesland	Wohnungsmieten – Nettokaltmieten, EUR je m² Wohnfläche, monatlich bezogen auf ca. 3 Zimmer, ca. 70 m², ohne öffentl. geförd. Wohnungsbau								
		Fertigstellung bis 1948 (Wiedervermietung/Neuvertragsmiete)			Fertigstellung ab 1949 (Wiedervermietung/Neuvertragsmiete)			Neubau – Erstbezug (Erstvermietung im Berichtsjahr)		Spitzen- bzw. Höchstmiete für Spitzenobjekte in Toplagen; ca.
		einfacher Wohnwert	mittlerer Wohnwert	guter Wohnwert	einfacher Wohnwert	mittlerer Wohnwert	guter Wohnwert	mittlerer Wohnwert	guter Wohnwert	
Aachen	Nordrhein-Westfalen	4,20	5,00	5,40	5,20	5,50	6,10	7,50	7,90	k.A.
Ahaus	Nordrhein-Westfalen	3,40	3,65	3,80	3,90	4,20	4,50	4,87	5,65	k.A.
Ahlen	Nordrhein-Westfalen	2,50	3,00	3,50	3,60	4,30	5,00	5,60	6,00	k.A.
Albstadt	Baden-Württemberg	2,50	3,70	4,30	3,20	4,10	4,80	5,50	6,00	6,50
Altena	Nordrhein-Westfalen	2,00	3,00	4,00	2,50	3,50	4,50	4,50	5,00	k.A.
Altenburg	Thüringen	3,00	4,00	5,00	3,00	4,00	5,00	4,80	5,25	k.A.
Alzey	Rheinland-Pfalz	k.A.	k.A.	k.A.	4,60	5,20	5,30	5,50	5,80	k.A.
Andernach	Rheinland-Pfalz	3,20	3,70	4,70	3,50	4,20	5,00	5,50	6,00	k.A.
Apolda	Thüringen	3,10	3,30	4,20	3,00	3,80	4,50	4,50	5,10	5,30
Arnsberg	Nordrhein-Westfalen	2,80	3,20	3,80	3,50	3,80	4,50	5,30	5,70	k.A.
Aschaffenburg	Bayern	4,50	5,23	6,23	4,63	5,13	6,60	7,63	7,95	k.A.

Abbildung 14: Ausschnitt aus dem IVD-Wohn-Preisspiegel 2006/2007[61]

[61] Quelle: IVD

4.4 Erklärende Variablen

Nachfolgend sind alle potentielle Variablen aufgelistet, die in die Analyse des Einflusses regionaler Faktoren auf den Wohnungsmietpreis eingeflossen sind:

- Einwohnerzahl
- Einwohner pro Haushalt
- Bevölkerungsdichte
- Einwohner in 30km Umkreis
- Sozialversicherungspflichtig Beschäftigte (SVB)
- SVB in 30km Umkreis
- SVB pro Einwohner
- SVB in 30km Umkreis pro Einwohner in 30km Umkreis
- Anteil Wohnungen in Ein-, Zwei- und Mehrfamilienhäusern
- Kaufkraftindex
- Arbeitslosenquote
- Anzahl der Übernachtungsgäste je Einwohner
- Siedlungsstrukturelle Raumtypen

4.4.1 Kurzbeschreibung der Variablen

4.4.1.1 Einwohnerzahl

Die Einwohnerzahl ist die absolute Zahl aller Einwohner einer Stadt oder Gemeinde (vgl. Abbildung 46 im Anhang). Die Zahlen basieren auf der Erhebung des Statistischen Bundesamtes und spiegeln den Stand vom 31.12.2006 wieder.[62]

4.4.1.2 Einwohner pro Haushalt

Mit dieser Kennzahl wird die durchschnittliche Haushaltsgröße je Stadt oder Gemeinde dargestellt. Diese lag im Jahr 2006 bei 2,08 Personen.[63] Es ist zu beobachten, dass Haushalte mit bis zu fünf Personen eher in den ländlich geprägten Regionen anzutreffen sind, während in den Großstädten überwiegend Ein- und Zweipersonenhaushalte vertreten sind (vgl. Abbildung 47 im Anhang). Die Angaben zu der Anzahl Haushalte auf Gemeindeebene konnten aus dem Geoinformationssystem Regiograph® gewonnen werden. Diese sind dort auf Gemeindeebene implementiert und basieren auf den Erhebungen der GfK - Gesellschaft für Konsumforschung (Stand vom 31.12.2006).

[62] Quelle: DESTATIS

[63] Quelle: [1]

4.4.1.3 Bevölkerungsdichte

Die Bevölkerungsdichte wird in Einwohner je km² wiedergegeben. In den Großstädten und Ballungsräumen ist eine sehr hohe Anzahl an Einwohnern pro Fläche zu beobachten, während z.B. weite Teile Ost-Deutschlands oder auch Schleswig-Holsteins eine relativ geringe Dichte aufweisen (vgl. Abbildung 48 im Anhang). So ist etwa in München mit 4.171 Einwohnern/km² die höchste Bevölkerungsdichte und in Wiedenborstel (Schleswig-Holstein) mit rund 1,55 Einwohnern/km² die niedrigste Bevölkerungsdichte zu beobachten.[64]

4.4.1.4 Einwohner in 30km Umkreis

Mit Hilfe von Regiograph® wurde eine 30km Radienbildung um jede Gemeinde vorgenommen. Diese soll die typische Pendlerdistanz repräsentieren.

Die Bildung der Umkreise war in Regiograph® nicht ohne weiters möglich, da die Städte und Gemeinden als sog. „Flächenlayer" dargestellt sind und für solche keine Radienbildung möglich ist. Um diese vorzunehmen müssen zuerst sog. „Punktelayer" erzeugt werden, für die dann Umkreise mit variablem Radius gebildet werden können.

Um diese Punktelayer zu erstellen, mussten zunächst die Geokoordinaten der deutschen Städte und Gemeinden aus der freien Online-Datenbank OpenGeoDB importiert werden und die Längen- und Breitengrade in Regiograph® in einem neuen Punktelayer abgelegt werden. Diesen Vorgang nennt man Geokodieren. Für jeden Datensatz wird dabei auf der Karte ein neuer Punkt erzeugt, der dann bearbeitet werden kann. Es können nun die gewünschten Umkreise von 30km gebildet werden und die Einwohnerzahl anteilig nach Flächen berechnet werden. Die so erzeugten Daten werden in einer neuen Spalte im Datensatz abgelegt. Mit einem sog. „Spatial-Join" ist es dann möglich, die Daten des neu erzeugten Punktelayers auf den vorhandenen Flächenlayer zu transferieren und die Daten grafisch darzustellen.

4.4.1.5 Sozialversicherungspflichtig Beschäftigte

Zu dem von der Sozialversicherungspflicht erfassten Personenkreis zählen alle Arbeitnehmer einschließlich der zu ihrer Berufsausbildung Beschäftigten, die kranken-, renten-, pflegeversicherungspflichtig und/oder beitragspflichtig sind oder für die von den Arbeitgebern Beitragsanteile zu entrichten sind. Daneben besteht in wenigen Fällen auch für Selbständige Versicherungspflicht in der Sozialversicherung.

Wehr- und Zivildienstleistende gelten nur dann als sozialversicherungspflichtig Beschäftigte, wenn sie ihren Dienst aus einem weiterhin bestehenden Beschäftigungsverhältnis heraus angetreten haben und nur wegen der Ableistung dieser Dienstzeiten kein Entgelt erhalten.

Nicht zu den sozialversicherungspflichtig Beschäftigten zählen dagegen der weitaus überwiegende Teil der Selbständigen, die mithelfenden Familienangehörigen sowie die Beamten.[65]

---

[64] Quelle: DESTATIS

[65] Quelle: DESTATIS3

Für die Analysen wurde die Kennzahl „sozialversicherungspflichtig Beschäftigte am Wohnort“ herangezogen (Stand: 30.06.2006).[66]

4.4.1.6 SVB in 30km Umkreis

Die 30km wurden als typische Pendlerdistanz ausgewählt. Für die Berechnung dieser Werte wurde das Geoinformationssytem Regiograph® verwendet.

Dazu mussten, wie bereits bei den Einwohnern in 30km Umkreis beschrieben, einige Bearbeitungsschritte in Regiograph® durchgeführt werden, die hier aber nicht noch einmal explizit erklärt werden sollen, sondern unter *4.4.1.4* nachgelesen werden können.

4.4.1.7 SVB pro Einwohner

Mit dieser Kennzahl wird der Anteil der sozialversicherungspflichtig Beschäftigten im Verhältnis zur Einwohnerzahl je Stadt oder Gemeinde dargestellt. Diese Variable wurde erst im weiteren Verlauf der Untersuchungen gebildet, da sie sich als relevant dargestellt hatte.

4.4.1.8 SVB in 30km Umkreis pro Einwohner in 30km Umkreis

Diese Kennzahl beschreibt das Verhältnis der sozialversicherungspflichtig Beschäftigten und der Einwohner in der 30km Pendeldistanz. Auch diese Variable stellte sich erst im Verlauf der Analysen als relevant heraus und wurde daher erst später gebildet.

4.4.1.9 Anteil Wohnungen in EFH / ZFH / MFH

Ein Einfamilienhaus (EFH) ist ein Gebäude mit nur einer Nutzungseinheit, wohingegen das Zweifamilienhaus (ZFH) für zwei Nutzer oder Mietparteien konzipiert ist. Zu den Mehrfamilienhäusern (MFH) zählen Gebäude mit drei oder mehr Wohnungen, die sich in der Regel auf mehrere Geschosse verteilen (z.B. Wohnblocks, Apartmenthäuser und Hochhäuser).

Die Verteilungsstruktur der Gebäudetypen in Deutschland ist regional unterschiedlich. Die deutlichsten Unterschiede sind erwartungsgemäß zwischen den Städten und den ländlich geprägten Regionen zu beobachten. In den Großstädten überwiegt der Anteil der Mehrfamilienhäuser, die ländlichen Gebiete sind eher durch Ein- und Zweifamilienhäuser geprägt (vgl. Abbildung 51, 52 und 53 im Anhang).

4.4.1.10 Kaufkraftindex

Der Kaufkraftindex spiegelt das Kaufkraftniveau einer Region pro Einwohner im Vergleich zum nationalen Durchschnitt wieder. Dieser wird auf 100 normiert. Liegt jetzt z.B. in einer Gemeinde A der Kaufkraftindex bei 88, so kann das dortige Kaufkraftniveau als leicht unterdurchschnittlich bezeichnet werden. Die Einwohner der Gemeinde A verfügen im Mittel nur über 88% der durchschnittlichen Kaufkraft. In Gemeinde B dagegen liegt der Kaufkraftindex

[66] Quelle: Beschäftigungsstatistik der Bundesagentur für Arbeit (BA)

bei beispielsweise 185. Damit verfügt Gemeinde B im nationalen Vergleich über eine überdurchschnittlich hohe Kaufkraft.[67]

Es ist zu beobachten, dass die meisten Regionen Ost-Deutschlands über eine sehr niedrige Kaufkraft verfügen. In Westdeutschland sind es gerade die Ballungsräume, die über die höchste Kaufkraft in Deutschland verfügen (vgl. Abbildung 54 im Anhang). Die Angaben zum Kaufkraftindex konnten Regiograph® entnommen werden und spiegeln den Stand vom 31.12.2006 wieder.

4.4.1.11 Arbeitslosenquote

Die hier angegebene Arbeitslosenquote entspricht nicht der Arbeitslosenquote, die durch die Bundesagentur für Arbeit ausgegeben wird, da diese nicht auf Gemeindeebene vorliegt. Für die Analysen in dieser Diplomarbeit wurde die Arbeitslosenquote wie folgt berechnet:

$$Arbeitslosenquote = \frac{Arbeitslose\ insgesamt}{\sum(SVB,\ GeB,\ Arbeitslose\ insgesamt)}$$

Hierbei ist zu beachten, dass die so ermittelte Arbeitslosenquote höher ausfällt, da mit einem kleineren Nenner gerechnet wird. Die Angaben zu den Arbeitslosen insgesamt, den sozialversicherungspflichtig Beschäftigten (SVB) und den geringfügig Beschäftigten (GeB) basieren auf der Erhebung der Bundesagentur für Arbeit und geben den Stand auf Gemeindeebene zum 30.06.2006 wieder.[68]

4.4.1.12 Anzahl der Übernachtungsgäste je Einwohner

Diese Variable wurde erst im weiteren Verlauf der Untersuchungen gebildet, da sie sich erst bei der Analyse des IVD-Datensatzes (Kapitel 6) als relevant dargestellt hatte. Die Anzahl der Gästeübernachtungen pro Jahr in einer Stadt oder Gemeinde basieren auf den Erhebungen des Statistischen Bundesamtes und spiegeln den Stand vom 31.12.2006 wieder.[69] Diese Kennzahl wurde durch die Anzahl der Einwohner in der jeweiligen Stadt oder Gemeinde geteilt, um reine Touristenorte spezifizieren zu können.

4.4.1.13 Siedlungsstrukturelle Raumtypen

Als weitere Größe wird der siedlungsstrukturelle Gemeindetyp nach der Klassifikation des BBR mit einbezogen (im Folgenden als BBR-Gemeindetyp bezeichnet). Die Gemeinden werden je nach ihrer Lage nach dem siedlungsstrukturellen Regions- und Kreistyp klassifiziert. Auf Gemeindeebene wird dann danach unterschieden, ob die Gemeinden aus raumordnerischer Perspektive eindeutig städtisch (Ober- oder Mittelzentrum) sind bzw. entsprechende Funktionen wahrnehmen oder nicht (sonstige Gemeinden). Insgesamt werden 17 Gemeindetypen unterschieden:

---

[67] Quelle: aus [2]

[68] Quelle: Arbeitsmarktstatistik der BA

[69] Quelle: DESTATIS

Kreistyp	Gemeindetyp	Code	Beispiel
Agglomerationsräume			
Kernstädte	Kernstädte > 500.000 Einwohner	1	Frankfurt am Main
	Kernstädte < 500.000 Einwohner	2	Darmstadt
Hochverdichtete Kreise	Ober- / Mittelzentrum	3	Dieburg
	Sonstige Gemeinden	4	Alsbach-Hähnlein
Verdichtete Kreise	Ober- / Mittelzentrum	5	Hanau
	Sonstige Gemeinden	6	Nidderau
Ländliche Kreise	Ober- / Mittelzentrum	7	Erding
	Sonstige Gemeinden	8	Zwingenberg (Neckar-Odw.-Kreis)
Verstädterte Räume			
Kernstädte	Kernstädte	9	Kassel
Verdichtete Kreise	Ober- / Mittelzentrum	10	Gießen
	Sonstige Gemeinden	11	Naumburg (Kreis Kassel)
Ländliche Kreise	Ober- / Mittelzentrum	12	Melsungen
	Sonstige Gemeinden	13	Oberaula
Ländliche Räume			
Verdichtete Kreise	Ober- / Mittelzentrum	14	Fulda
	Sonstige Gemeinden	15	Niederaula
Ländliche Kreise	Ober- / Mittelzentrum	16	Bad Kissingen
	Sonstige Gemeinden	17	Oberammergau

Tabelle 24: Siedlungsstrukturelle Gemeindetypen nach der BBR-Systematik (eigene Darstellung)[70]

[70] Quelle: [4]

5 Untersuchung des Mietspiegel-Datensatzes

Die Basis für die in dieser Diplomarbeit durchgeführten Untersuchungen bildet der Datensatz „Mietspiegel.sav“. Er enthält für 359 deutsche Städte und Gemeinden Mietpreisinformationen zu den in Kapitel 4.1.1 beschriebenen Standardwohnungstypen. Des Weiteren enthält dieser Datensatz die von der Autorin im Rahmen eines Praktikums beim IWU erhobenen Daten zu den potentiellen erklärenden Variablen.[71]

Die nachfolgenden Untersuchungen werden anhand der ermittelten Mieten für eine 3-Zimmer-Wohnung (nmqm_3Z) durchgeführt. Dies dient der Vergleichbarkeit der Ergebnisse mit den Ergebnissen aus der Untersuchung des IVD-Datensatzes (s. Kapitel 6), da dieser nur Mietpreisinformationen über 3-Zimmer-Wohnungen enthält.

Im Anhang sind alle Variablen mit Namen und ihrer Bedeutung in einer Tabelle zusammengefasst.

5.1 Übersicht über den Datensatz

Bevor die Daten für die geplanten Analysen verwendet werden können, steht eine Überprüfung der zu verwendenden Variablen an. Anhand der Tabelle „Statistiken“ kann sich ein erster Überblick über den Datensatz verschafft werden.

Statistiken

	N		Mittelwert	Standardabweichung	Minimum	Maximum
	Gültig	Fehlend				
nmqm_3Z Nettomiete in Euro pro m² für 3-Z-Wohnung	357	2	5,3710	1,08054	3,22	10,08
EW Anzahl Einwohner zum 31.12.2006	359	0	84370,28	234754,946	199	3404037
EW_HH Einwohner pro Haushalt	359	0	2,165991	,2045043	1,7242	3,4064
SVB_ins_Wohn Sozialverspflichtig Besch. am Wohnort zum 30.06.2006 lt. BA	359	0	25934,83	69261,826	98	933649
ALQ Arbeitslosenquote	357	2	,119306	,0523425	,0285	,2989
KKI Kaufkraftindex je EW	359	0	100,284859	13,0237041	68,3236	138,5142
km30_SVBinsg_antlg Soz.vers.pflichtig Besch. in 30km Umkreis	358	1	456649,39	338122,658	16236	1402801
km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	358	1	1467881,31	1118559,047	48604	4797792

[71] Siehe Kapitel 4

Dichte Bevölkerungs-dichte	359	0	757,80	671,308	20	4171
Anteil_Whg_efh Anteil Wohnungen in Einfa-milienhäusern	359	0	,296281	,1450053	,0710	,7116
Anteil_Whg_zfh Anteil Wohnungen in Zwei-familienhäusern	359	0	,197878	,0970152	,0228	,4267
Anteil_Whg_mfh An-teil Wohnungen in Mehrfamilienhäusern	359	0	,505841	,2194593	,0458	,9007
siedl_gemtyp sied-lungsstruktureller Ge-meindetyp	359	0	6,82	4,279	1	17

Tabelle 25: Deskriptive Statistik der Variablen

Die erhobenen Nettomieten aus den Mietspiegel-Analysen schwanken pro m² zwischen 3,22€ und 10,08€; die Durchschnittsmiete beträgt 5,37€.

Wichtig ist, dass alle Fälle, für die keine Nettomiete ausgewiesen ist (Missing-Values), von den weiteren Analysen ausgeschlossen werden. In dieser Diplomarbeit erfolgt der listenweise Ausschluss (der Fall wird ausgeschlossen, falls eine der Variablen einen fehlenden Wert aufweist[72]).

5.2 Abgeleitete Variablen

5.2.1 Bildung von Dummy-Variablen

Die Variable „Siedlungsstruktureller Gemeindetyp“ (siedl_gemtyp) kann Werte von 1 bis 17 annehmen (s. Tabelle 25). Dabei bedeutet z.B. „1“ Kern- und Großstädte >= 500.000 E. in Agglomerationsräumen, „2“ steht für Kern- und Großstädte unter 500.000 E. in Agglomerationsräumen usw. Um diese Variable mit in die Untersuchung einfließen zu lassen, muss sie in 17 Dummy-Variablen umprogrammiert werden. Diese erhalten die Namen „gtyp1“ bis „gtyp17“ und sind wie folgt definiert:

Die neue Variable „gtyp3“ erhält den Wert „1“, wenn die Variable „siedl_gemtyp“ für eine Stadt oder Gemeinde den Wert „3“ (OZ/MZ in hochverdichteten Kreisen von Agglomerationsräumen) annimmt, ansonsten bekommt die Variable den Wert „0“ zugewiesen. Genauso werden die anderen 16 Variablen definiert.

Dieses Vorgehen liegt darin begründet, dass in den nachfolgenden Untersuchungen überprüft werden soll, ob eine bestimmte Gemeindetypisierung Einfluss auf den Wohnungsmietpreis hat. Würde sich die Variable „siedl_gemtyp“ als signifikant erweisen, könnte nur erklärt werden, dass der Gemeindetyp einen Einfluss auf den Wohnungsmietpreis hat. Mit Hilfe der

[72] SPSS-Befehl: /MISSING LISTWISE

Umcodierung der Variablen kann nun aber der genaue Gemeindetyp identifiziert werden, der diesen Einfluss ausübt. Dies erleichtert die inhaltliche Interpretation der Ergebnisse.

5.2.2 Bildung von Referenzkategorien

Nachdem die Variable „siedl_gemtyp“ in 1/0-codierte Dummy-Variablen aufgelöst wurde, muss nun entschieden werden welcher Gemeindetyp als Referenzkategorie dienen soll. Die Bildung dieser Referenzkategorie ist notwendig, da sonst zwischen den neu gebildeten Variablen „gtyp1“ bis „gtyp17“ eine perfekte lineare Abhängigkeit besteht. Die Multikollinearität zwischen den Variablen beträgt dann „1“ und es ist keine Koeffizientenschätzung in der Regressionsanalyse möglich.

Die Referenzkategorie sollte möglichst so gewählt werden, dass sie nicht unverhältnismäßig wenige Fälle betrifft. Dafür werden zunächst die Häufigkeitsverteilungen auf die einzelnen Gemeindetypen betrachtet:

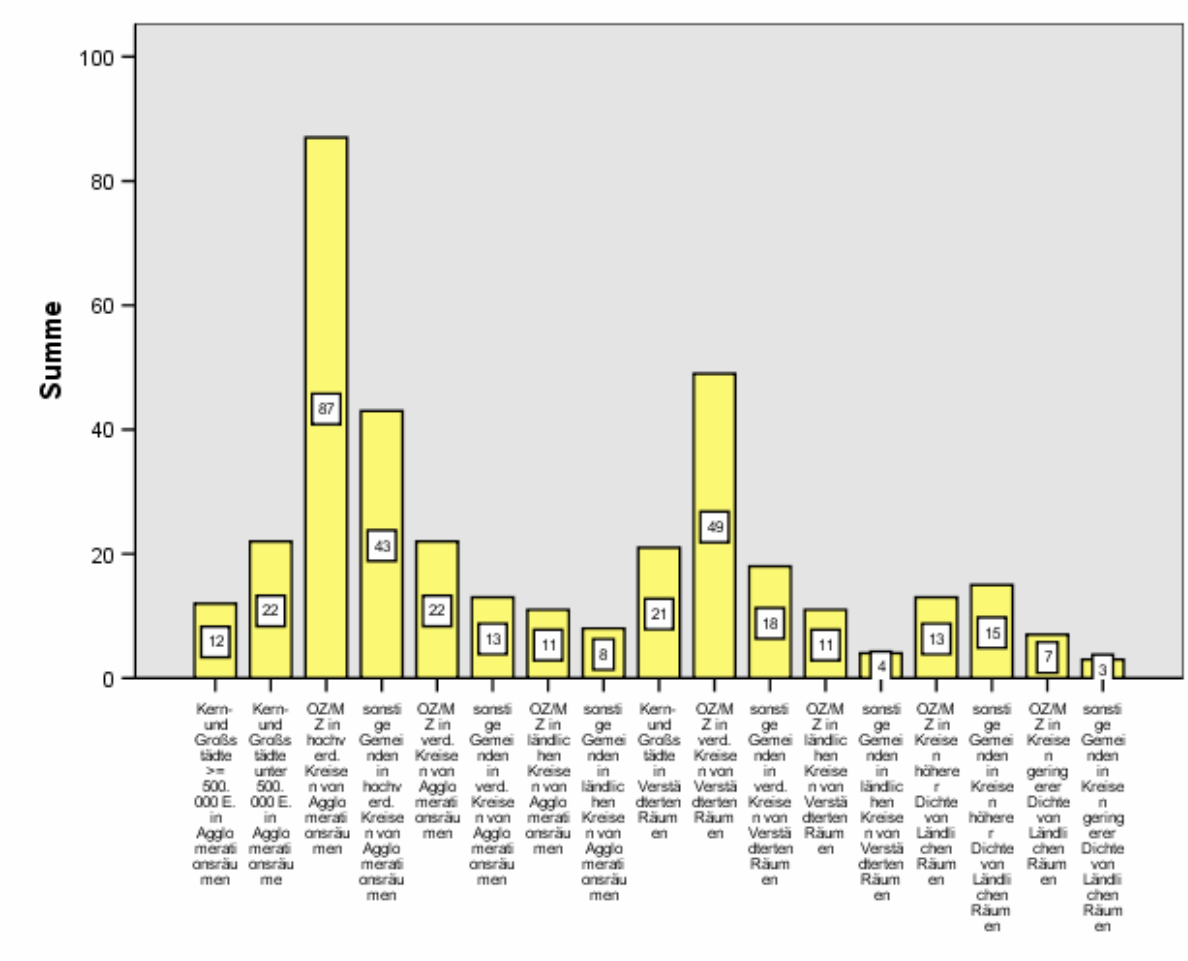


Abbildung 15: Häufigkeitsverteilungen der siedlungsstrukturellen Gemeindetypen

Als Referenzkategorie wird „Gemeindetyp 3“ (OZ/MZ in hochverdichteten Kreisen von Agglomerationsräumen) ausgewählt, da dieser am häufigsten besetzt ist und relativ am Rand liegt.

Als nächstes werden die Variablen „Anteil Wohnungen in…“ betrachtet. Addiert man die drei Werte für eine Stadt oder Gemeinde auf, so ergeben diese immer Eins. Daher muss auch hier wieder eine Variable als Referenzkategorie festgelegt werden, in diesem Fall die Variable „Anteil Wohnungen in Mehrfamilienhäusern“.

5.3 Erstes Überprüfen auf Multikollinearität

Lineare Messwertverteilungen unter den Prädiktoren geben einen ersten Hinweis auf Multikollinearität. Bei einer ersten Analyse der Korrelationsmatrix fallen die hohen Korrelationskoeffizienten zwischen den Variablen „Anzahl Einwohner“ und „Sozialversicherungspflichtig Beschäftigte am Wohnort“ sowie zwischen den Variablen „Sozialversicherungspflichtig Beschäftigte in 30km Umkreis“ und „Anzahl Einwohner in 30km Umkreis“ auf (s. Tabelle 26).

Korrelationen

		SVB_ins_Wohn Sozialverspflichtig Besch. am Wohnort zum 30.06.2006 lt. BA	EW Anzahl Einwohner zum 31.12.2006
SVB_ins_Wohn Sozialverspflichtig Besch. am Wohnort zum 30.06.2006 lt. BA	Korrelation nach Pearson	1	,995(**)
	Signifikanz (2-seitig)		,000
	N	359	359
EW Anzahl Einwohner zum 31.12.2006	Korrelation nach Pearson	,995(**)	1
	Signifikanz (2-seitig)	,000	
	N	359	359

** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

		km30_SVBinsg_antlg Soz.vers.pflichtig Besch. in 30km Umkreis	km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis
km30_SVBinsg_antlg Soz.vers.pflichtig Besch. in 30km Umkreis	Korrelation nach Pearson	1	,994(**)
	Signifikanz (2-seitig)		,000
	N	358	358
km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	Korrelation nach Pearson	,994(**)	1
	Signifikanz (2-seitig)	,000	
	N	358	358

** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Tabelle 26: Korrelationsanalyse

Auch die bivariaten Streudiagramme zeigen einen eindeutigen linearen Zusammenhang:

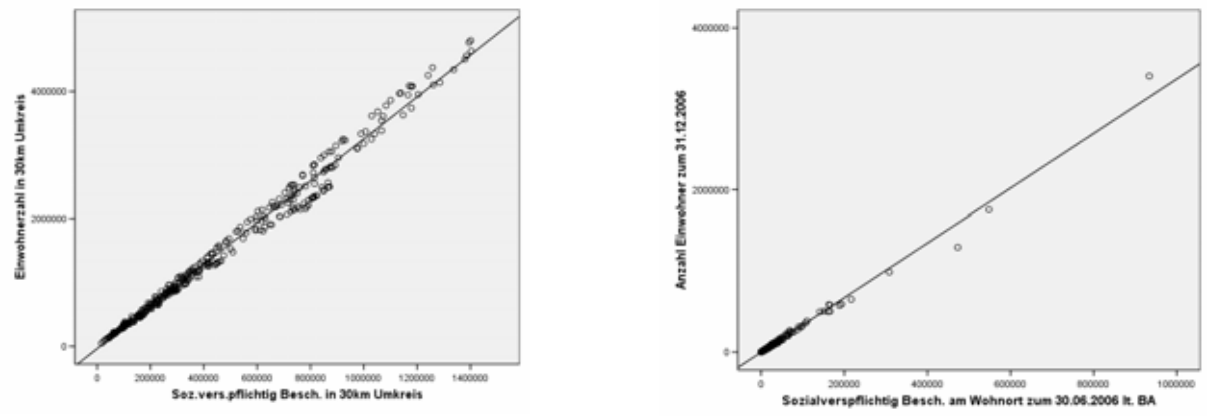


Abbildung 16: Bivariate Streudiagramme

Daher werden zwei neue Variablen gebildet, indem die Anzahl der sozialversicherungspflichtig Beschäftigten am Wohnort durch die Einwohnerzahl geteilt und die Anzahl sozialversicherungspflichtig Beschäftigter in 30km Umkreis durch die Einwohnerzahl in 30km Umkreis geteilt wird. Diese zwei neuen Variablen („SVB_pro_EW“ und „SVB_EW_30km“) fließen nun statt der bisherigen Variablen („Sozialversicherungspflichtig Beschäftigte am Wohnort“ und „Sozialversicherungspflichtig Beschäftigte in 30km Umkreis“) in die weiteren Analysen mit ein.

5.4 Analyse verschiedener Modellvarianten

Bevor die nachfolgenden Abschnitten beschriebenen Modelle 1 und 2 des Mietspiegeldatensatzes ausgewählt wurden, sind verschiedene Modellvarianten analysiert worden.

- Zum einen wurde für die unter 5.6 beschriebene Faktorenanalyse untersucht, ob mit Hilfe der Dummy-Variablen der siedlungsstrukturellen Gemeindetypen sinnvolle Faktoren extrahiert werden können. Das Einbeziehen dieser 17 Variablen brachte aber keine sinnvolle Lösung, da sich ein Großteil der Gemeindetypen als einzelne Faktoren ausbildete. Nach dem Kaiser-Kriterium wären insgesamt acht Faktoren zu extrahieren gewesen, die in einer anschließenden Regressionsanalyse aber eine schlechtere Modellgüte zur Folge gehabt hätten. Deshalb sind für Modell 2 die siedlungsstrukturellen Gemeindetypen erst in die anschließende Regressionsanalyse als unabhängige Variable mit eingeflossen.
- Für die Regressionsanalysen wurden ebenfalls verschiedene Variationen analysiert. Zum einen wurde versucht über die Auswahlmethode der Variablen (BACKWARD, STEPWISE) eine gute Modellanpassung zu finden. Zum anderen wurde versucht, siedlungsstrukturelle Gemeindetypen, die sich räumlich ähnelten und auch das gleiche Vorzeichen aufwiesen, in Klassen zusammenzufassen. Diese Variationen haben aber auch zu keiner Modellverbesserung geführt, was wahrscheinlich auf die geringen Fallzahlen zurückzuführen ist.
- Für die Regressionsanalysen wird der Einfluss der unabhängigen Variablen auf die abhängige Variable als linear angenommen. Die Überprüfung der Scatterplots hat kei-

nen kurvilinearen oder exponentiellen Zusammenhang erkennen lassen. Die Überprüfung der Residualplots brachte dasselbe Ergebnis. Aus Platzgründen werden die Scatterplots nicht extra aufgeführt.

Modell 1 und 2 haben sich als die am besten geeigneten Varianten herausgestellt und sollen deshalb nachfolgend näher erläutert werden.

5.5 Regressionsanalyse für Modell 1

In einem ersten Modell soll mittels einer Regressionsanalyse untersucht werden, ob und welchen Einfluss die verschiedenen Variablen auf den Wohnungsmietpreis haben. Es werden alle Fälle, für die keine Nettomiete ausgewiesen ist (Missing-Values), von den weiteren Analysen ausgeschlossen[73]. Als Auswahlverfahren für die Regressionsanalyse wird die Methode „BACKWARD“ gewählt. Dabei werden zunächst alle Variablen in das Modell aufgenommen und danach sukzessive anhand ihres Signifikanzniveaus (beginnend mit dem höchsten α-Wert) wieder aus dem Modell entfernt, bis das Modell nur noch signifikante Variablen enthält.

5.5.1 Festlegung des Signifikanzniveau α

Bevor die Regressionsanalyse durchgeführt werden kann, muss zunächst das Signifikanzniveau α festgelegt werden.

Für den globalen F-Test, der die Nullhypothese testet, dass alle Steigungsparameter gleich Null sind, wird das Signifikanzniveau auf 1% festgelegt. Für die Bewertung der einzelnen Koeffizienten (t-Test) wird ein α von 10% festgelegt.

5.5.2 Überprüfung der Voraussetzungen[74]

Bevor ein Regressionsmodell aufgrund seiner Modellgüte oder der statistischen Signifikanz seiner Parameter angenommen werden kann, muss es sorgfältig daraufhin überprüft werden ob alle Modellprämissen eingehalten wurden. Ist das gefundene Modell valide, kann es inhaltlich interpretiert werden.

5.5.2.1 Multikollinearität

Die Regressoren im linearen Regressionsmodell sollten nach Möglichkeit untereinander nicht linear abhängig sein. Multikollinearität kann mit Hilfe des Variance Inflation Factor (VIF) und der Toleranz aufgedeckt werden.

Toleranzwerte von unter 0,1 lassen nahezu sicher auf Kollinearität schließen. In der hier betrachteten Regression liegt der kleinste Toleranzwert bei 0,362, so dass offenbar keine Multikollinearität vorliegt. Der VIF sollte nach Möglichkeit unter 3 liegen, ein Wert größer 5 ist

[73] Siehe Kapitel 5.1; Ausschluss erfolgt Listenweise

[74] Vgl. Kapitel 3.6

als kritisch zu betrachten. In diesem Modellansatz liegt der höchste VIF bei 2,763. Somit kann Multikollinearität ausgeschlossen werden.

	Kollinearitätsstatistik	
	Toleranz	VIF
(Konstante)		
EW_HH Einwohner pro Haushalt	,362	2,763
KKI Kaufkraftindex je EW	,790	1,266
Dichte Bevölkerungsdichte	,463	2,159
Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	,385	2,600
Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	,408	2,451
gtyp14 OZ/MZ in Kreisen höherer Dichte von Ländlichen Räumen	,961	1,040
gtyp15 sonstige Gemeinden in Kreisen höherer Dichte von Ländlichen Räumen	,842	1,188

Tabelle 27: Kollinearitätsstatistik

5.5.2.2 Normalverteilung der Residuen

Eine weitere Modellannahme ist die Normalverteilung der Residuen (Abweichungen der Schätzwerte vom tatsächlichen Wert). Dieses Kriterium lässt sich zum einen grafisch mit Hilfe eines Histogramms der standardisierten Residuen überprüfen:

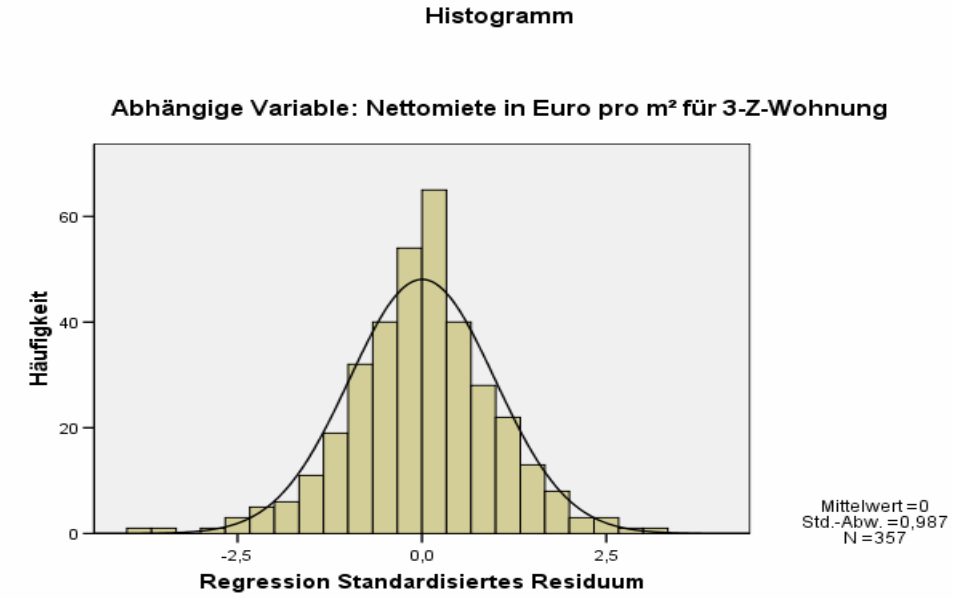


Abbildung 17: Histogramm der standardisierten Residuen

Das Histogramm bildet die Häufigkeitsverteilung der standardisierten Residuen ab und zeigt, dass sich die standardisierten Residuen recht gut an eine Standardnormalverteilung anpassen. In der Spitze und am linken Rand sind Abweichungen erkennbar. Geringfügige Abweichungen sind in der Praxis aber nicht ungewöhnlich und durchaus tolerierbar. Der Erwartungswert liegt bei 0, die Standardabweichung bei 0,987.

Der PP-Plot, der als zweites grafisches Kriterium herangezogen wird, ergibt ein ähnlich gutes Bild: Die standardisierten Residuen der abhängigen Variablen folgen mit ganz leichten Abweichungen der Referenzverteilung (Normalverteilung).

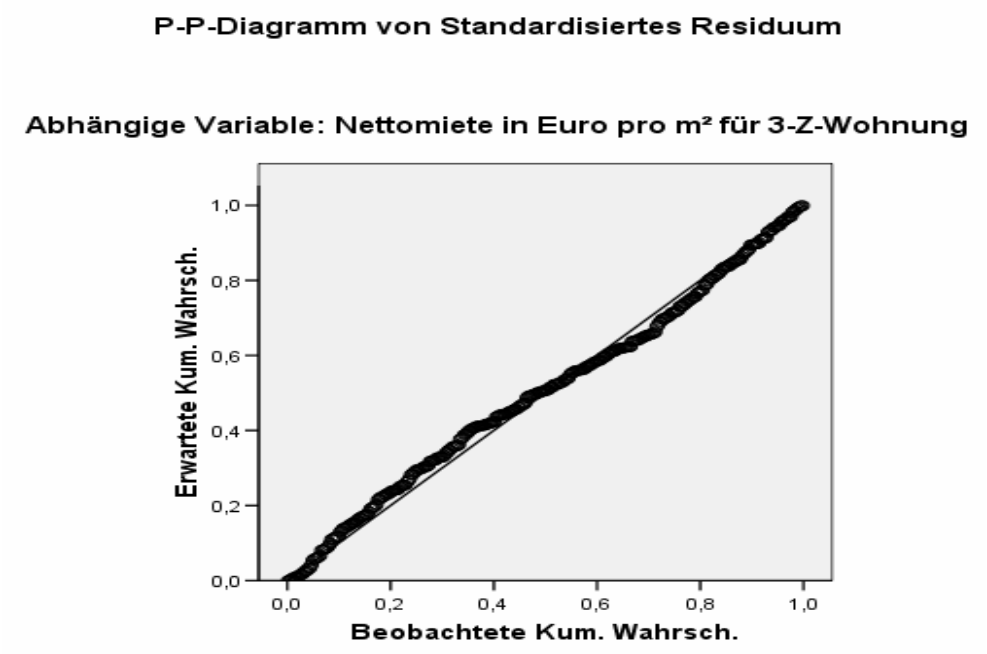


Abbildung 18: PP-Plot der standardisierten Residuen

Das Vorliegen einer Normalverteilung der standardisierten Residuen als Voraussetzung einer linearen Regressionsanalyse kann somit als erfüllt betrachtet werden.

5.5.2.3 Heteroskedastizität

Heteroskedastizität liegt dann vor, wenn die Streuung der Residuen nicht konstant ist. Die Streuung darf weder zeitlich noch betragsmäßig von den Beobachtungen abhängig sein. Grafisch kann Heteroskedastizität mit Hilfe eines Streudiagramms identifiziert werden, auf dessen x-Achse die standardisierten geschätzten Werte und auf der y-Achse die standardisierten Residuen abgetragen werden.

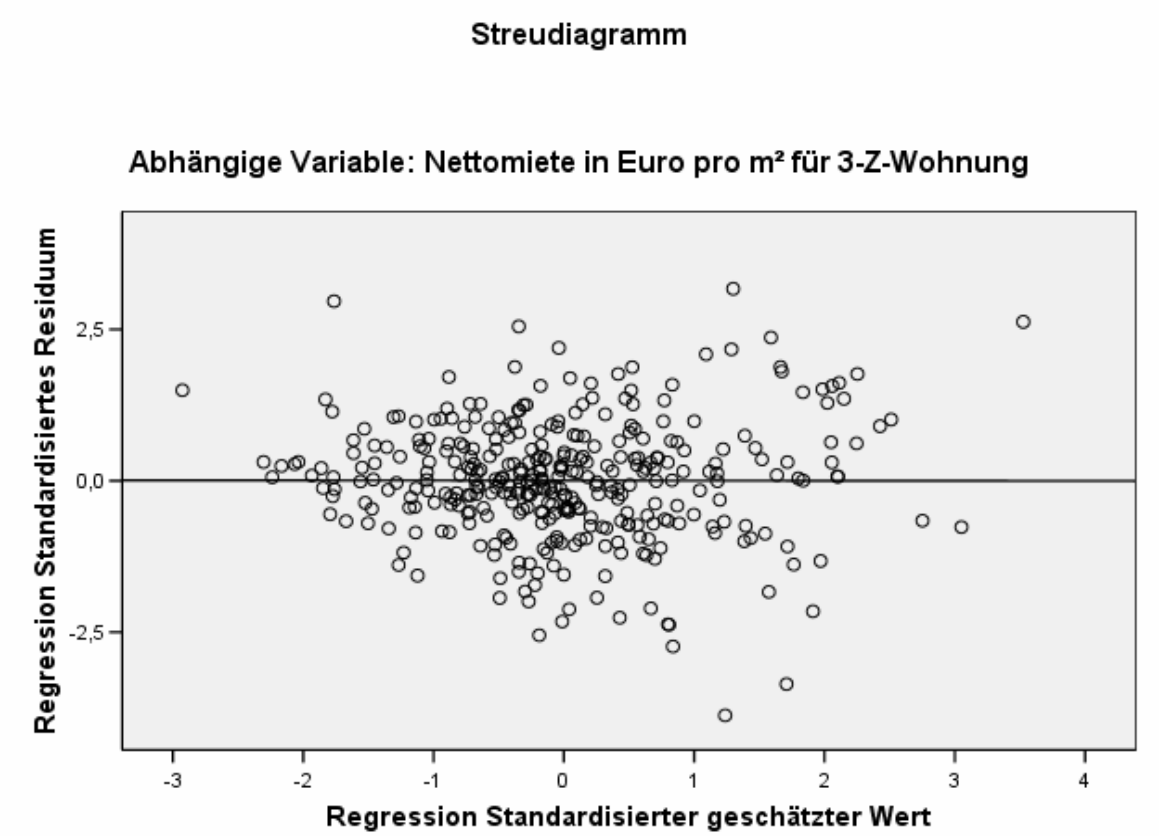


Abbildung 19: Streudiagramm für den Test auf Varianzgleichheit

Das Streudiagramm lässt mit wachsenden geschätzten Werten eine leichte Zunahme der Streuung der Residuen erkennen, d.h. die Extrembereiche streuen etwas stärker. Diese Streuung liegt aber noch im akzeptablen Bereich und ist als normal zu interpretieren, da für die Extrembereiche nur wenige Daten vorliegen (Einzelfälle). Damit liegt keine Verletzung der Modellannahme vor. Auch bei der Auswertung der Streudiagramme der einzelnen Variablen konnten keine Auffälligkeiten festgestellt werden.

5.5.2.4 Autokorrelation

Die serielle Unkorreliertheit der Störgrößen kann mit Hilfe der Durbin-Watson-Statistik überprüft werden. Liegt die Teststatistik nahe bei 2 ist davon auszugehen, dass keine Autokorrelation vorliegt. Der Index-Wert zur Prüfung auf Autokorrelation liegt in diesem Modellansatz bei 2,032 (s. Tabelle 28), somit kann Autokorrelation ausgeschlossen werden.

5.5.3 Überprüfung der Modellgüte[75]

Nachdem die Voraussetzungen der Regressionsanalyse überprüft wurden, kann nun die Güte des berechneten Modells beurteilt werden.

[75] Vgl. Kapitel 3.4

Das Bestimmtheitsmaß R^2 gibt den Anteil der erklärten Streuung an, der durch das Modell beschrieben wird und kann Werte zwischen 0 und 1 annehmen. Ein R^2 von 0,65 bedeutet dabei z.B. dass 65% der Streuung in den y-Werten durch die Regressionsgerade erklärt werden.

Durch das eben ermittelte Regressionsmodell kann demnach 47,7% der (senkrechten) Streuung in den y-Werten erklärt werden.

Modellzusammenfassung(s)

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standard-fehler des Schätzers	Durbin-Watson-Statistik
18	,691(r)	,477	,467	,78968	2,032

r Einflussvariablen : (Konstante), gtyp15 sonstige Gemeinden in Kreisen höherer Dichte von Ländlichen Räumen, gtyp14 OZ/MZ in Kreisen höherer Dichte von Ländlichen Räumen, KKI Kaufkraftindex je EW, EW_HH Einwohner pro Haushalt, Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern, Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern, Dichte Bevölkerungsdichte

s Abhängige Variable: nmqm_3Z Nettomiete in Euro pro m² für 3-Z-Wohnung

Tabelle 28: SPSS-Output der Modellbeurteilung

Das korrigierte R-Quadrat $\overline{R}^2$ dient zur Beurteilung des Modells unter Berücksichtigung der Anzahl der Regressoren. Der Wert dieser Größe kann in der vierten Spalte von Tabelle 25 entnommen werden und beträgt für dieses Regressionsmodell 46,7%. Der Standardfehler des Schätzers nimmt in diesem Regressionsmodell einen Wert von 0,78968 an und beträgt, bezogen auf den Mittelwert der abhängigen Variablen, 14,7%.

ANOVA

	Quadrat-summe	df	Mittel der Quadrate	F	Signifikanz
Regression	196,855	7	28,122	45,097	,000(s)
Residuen	215,764	346	,624		
Gesamt	412,618	353			

Tabelle 29: SPSS-Ausgabe der Varianzanalyse

In den Spalten der Tabelle 29 ist neben den Freiheitsgraden (df) auch die Quadratsummenzerlegung angegeben. Die Quadratsumme der Residuen (SSE) beträgt in dem Modellansatz 215,764, die Quadratsumme der erklärten Abweichungen (SSR) nimmt einen Wert von 196,855 an. Des Weiteren kann die Schätzung der Restvarianz abgelesen werden, die in diesem Modell 0,624 beträgt.

Zur Überprüfung der Signifikanz des Regressionsmodells wird der F-Test herangezogen. Dieser formuliert die Nullhypothese, dass in der Grundgesamtheit kein linearer Zusammenhang zwischen Regressand und Regressoren besteht, also die Regressionskoeffizienten in der Grundgesamtheit Null sind.

Diese Hypothese kann zum unter 5.3.1 bestimmten Signifikanzniveau von 1% verworfen werden.

5.5.4 Überprüfung der Koeffizienten[76]

Als nächstes sind nun die Regressionskoeffizienten der einzelnen Variablen zu überprüfen. Dies geschieht mittels der T-Statistik und den angegebenen Konfidenzintervallen.

Koeffizienten(a)

	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B	
	B	Standardfehler	Beta			Untergrenze	Obergrenze
(Konstante)	-,514	,720		-,714	,475	-1,929	,901
EW_HH Einwohner pro Haushalt	,732	,341	,139	2,148	,032	,062	1,403
KKI Kaufkraftindex je EW	,049	,004	,584	13,343	,000	,041	,056
Dichte Bevölkerungsdichte	,000	,000	,106	1,859	,064	,000	,000
Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	-1,078	,466	-,145	-2,315	,021	-1,993	-,162
Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	-1,722	,678	-,155	-2,541	,011	-3,055	-,389
gtyp14 OZ/MZ in Kreisen höherer Dichte von Ländlichen Räumen	-,703	,237	-,118	-2,974	,003	-1,169	-,238
gtyp15 sonstige Gemeinden in Kreisen höherer Dichte von Ländlichen Räumen	-,595	,227	-,111	-2,622	,009	-1,042	-,149

a Abhängige Variable: nmqm_3Z Nettomiete in Euro pro m² für 3-Z-Wohnung

Tabelle 30: SPSS-Output der Koeffizientenstatistik

Das Signifikanzniveau α wurde bereits im Vorfeld unter 5.3.1 mit 10% festgelegt. Da die Methode „BACKWARD“ verwendet wurde, befinden sich in dem Regressionsmodell nur noch zu dem Niveau signifikante Variablen, die anderen wurden sukzessive aus dem Modell entfernt.

Neben den nicht standardisierten Regressionskoeffizienten gibt SPSS auch die standardisierten Regressionskoeffizienten aus, die einen ersten Hinweis zur Interpretation liefern: Je betragsmäßig größer ein standardisierter Regressionskoeffizient ist (Maximum: 1), umso größer ist der Einfluss des betreffenden Prädiktors. Sehr gut lässt sich der hohe Einfluss der Variablen „Kaufkraftindex“ ablesen, der in positiver Richtung verläuft. Das bedeutet, dass auf hohe bzw. kleine Werte des Kaufkraftindex hohe bzw. kleine Nettomieten folgen. Fließt die

[76] Vgl. Kapitel 3.5

Variable „Kaufkraftindex“ als einzige unabhängige Variable in ein Regressionsmodell mit ein, so erklärt sie bereits 37,6% der Streuung in den y-Werten.

Als nächstes sollen die 95%-Konfidenzintervalle der signifikanten Koeffizienten näher betrachtet werden. Diese werden von SPSS nur für die nichtstandardisierten Regressionskoeffizienten (B) ausgegeben. Liegt sowohl die untere als auch die obere Intervallgrenze im positiven oder negativen Bereich, kann die Richtung des Koeffizienten (Zu- oder Abschlag) bestimmt werden. An der Breite der Konfidenzintervalle kann auch die Schätzgenauigkeit abgelesen werden. So liegt z.B. für die Variable „Kaufkraftindex“ mit 95%iger Wahrscheinlichkeit der wahre Wert zwischen 0,041 (untere Intervallgrenze) und 0,056 (obere Intervallgrenze). Bei anderen Variablen fallen die Konfidenzintervalle zum Teil erheblich größer aus.

Die Schätzfunktion für die Nettomiete in Euro pro m² (nmqm) unter diesem Modell lautet:

$$\begin{aligned} nmqm = &-0{,}514 + 0{,}732 \cdot EW\_HH + 0{,}049 \cdot KKI + 0{,}00017 \cdot Dichte \\ &- 1{,}078 \cdot Anteil\_Whg\_efh - 1{,}722 \cdot Anteil\_Whg\_zfh \\ &- 0{,}703 \cdot gtyp14 - 0{,}595 \cdot gtyp15 \end{aligned}$$

5.5.5 Überprüfung auf räumliche Korrelation

Zum Schluss sollen die Residuen grafisch daraufhin untersucht werden, ob sich räumliche Muster identifizieren lassen. Dazu werden die Residualwerte mit Hilfe des Geoinformationssystems Regiograph® in einer Deutschlandkarte als farbige Punkte dargestellt (Farbverlauf von rot (negativ) nach grün (positiv)).

Die Auswertung der Grafik (s. Abbildung 20) lässt keine räumlichen Muster erkennen, die einen Hinweis auf nicht betrachtete Einflussfaktoren geben würde. Es lässt sich weder eine Clusterung der Residualwerte nach Bundesländern erkennen, noch lässt sich ein Unterschied in der Verteilung zwischen West- und Ostdeutschland bzw. Nord- und Süddeutschland identifizieren. Eine systematische Über- bzw. Unterschätzung der Miethöhen in Großstädten oder Ballungsräumen ist ebenfalls nicht erkennbar.

Vielmehr scheinen die Residualwerte unsystematisch und gleichmäßig verteilt zu liegen, so dass eine räumliche Korrelation ausgeschlossen werden kann.

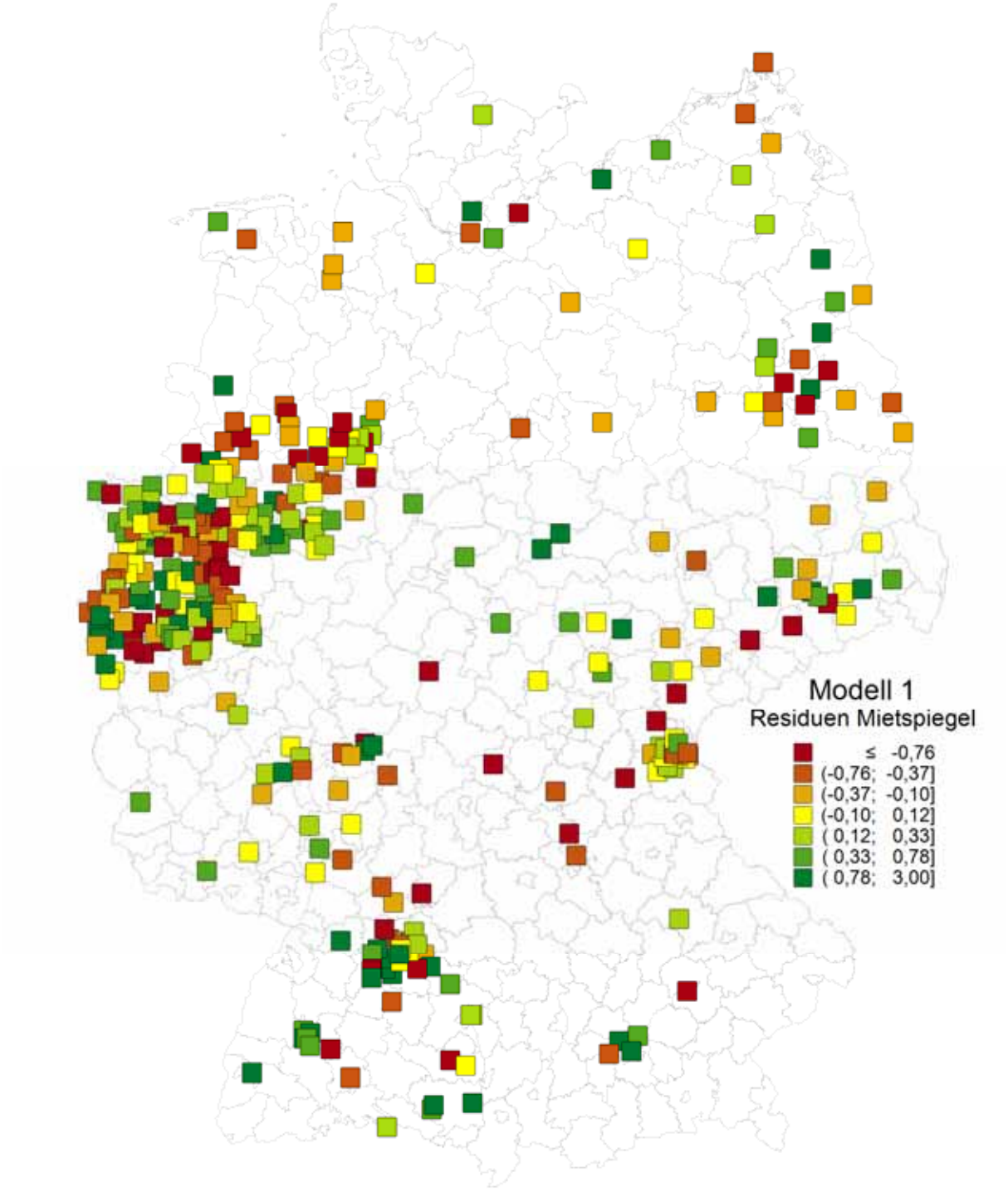


Abbildung 20: Darstellung der Residuen aus dem Mietspiegel-Modell 1[77]

[77] Quelle: Eigene Darstellung

5.6 Faktorenanalyse für Modell 2

Im nachfolgenden Abschnitt wird nun versucht, die Variablen die zueinander in enger Beziehung stehen zu identifizieren und nach Möglichkeit zu einem Faktor zusammenzufassen. Dieser gefundene Faktor muss sinnvoll zu interpretieren sein und geht anschließend als unabhängige Variable in eine Regressionsanalyse mit ein. Als Extraktionsmethode wird die Hauptkomponentenanalyse gewählt, da nach einem Sammelbegriff gesucht wird, der die zusammengefassten Variablen beschreibt[78]. Es ist darauf hinzuweisen, dass in dieser Ausarbeitung der Begriff „Komponente“ gleichbedeutend mit dem Begriff „Faktor“ ist.

5.6.1 Korrelationsanalyse

Um überhaupt eine Faktorenanalyse durchführen zu können, muss zunächst die Korrelationsmatrix auf ihre Eignung überprüft werden[79].

[78] Vgl. Kapitel 2

[79] Vgl. Kapitel 2.4

Korrelationsmatrix

	EW Anzahl Einwohner zum 31.12.2006	EW_HH Einwohner pro Haushalt	SVB_pro_Einw Sozialversp flichtig Besch. am Wohnort pro Einwohner	ALQ Arbeitslosenquote	KKI Kaufkraftindex je EW	km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	SVB_EW_30km	Dichte Bevölkerungsdichte	Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern
EW	1,000	-,335	-,100	,150	,024	,256	-,020	,603	-,312	-,326
EW_HH	-,335	1,000	,072	-,523	,110	-,073	-,194	-,575	,734	,683
SVB_pro_Einw	-,100	,072	1,000	-,414	,464	,037	,573	-,080	-,006	,197
ALQ	,150	-,523	-,414	1,000	-,697	-,160	-,176	,146	-,452	-,594
KKI	,024	,110	,464	-,697	1,000	,472	,269	,201	,109	,156
km30_EwGfK_antlg	,256	-,073	,037	-,160	,472	1,000	-,251	,524	-,184	-,176
SVB_EW_30km	-,020	-,194	,573	-,176	,269	-,251	1,000	,078	-,369	-,063
Dichte	,603	-,575	-,080	,146	,201	,524	,078	1,000	-,613	-,568
Anteil_Whg_efh	-,312	,734	-,006	-,452	,109	-,184	-,369	-,613	1,000	,632
Anteil_Whg_zfh	-,326	,683	,197	-,594	,156	-,176	-,063	-,568	,632	1,000

Tabelle 31: SPSS- Output der Korrelationsmatrix

An der Korrelationsmatrix ist zu erkennen, dass bei einigen Paaren der Korrelationskoeffizient 0,3 übersteigt, aber unter 0,5 liegt. Hier ist also nur ein mittlerer Effekt zu erkennen. Des Weiteren treten auch einige Korrelationskoeffizienten mit Werten < 0,3 oder > 0,5 auf, so dass die Korrelationsmatrix selbst kein eindeutiges Urteil über die Eignung der Daten zur Faktorenanalyse zulässt.

5.6.2 Variablenauswahl

Wie im oberen Abschnitt bereits beschrieben, weist die Korrelationsmatrix neben sehr kleinen Werten auch hohe Korrelationskoeffizienten auf und es lässt sich soweit noch keine sinnvolle Aussage zur Eignung für eine Faktorenanalyse treffen. Darum werden nun weitere Kriterien zur Überprüfung herangezogen.

5.6.2.1 Signifikanzniveau der Korrelation[80]

Zu den jeweiligen Korrelationen gibt SPSS auch die Signifikanzniveaus aus. Kleine Werte geben dabei die Wahrscheinlichkeit an, dass sich die Korrelation signifikant von Null unterscheidet. Die Analyse der Signifikanzniveaus ergibt, dass sich die meisten Korrelationen signifikant von Null unterscheiden. Des Weiteren lassen sich keine Variablen feststellen, die mit keiner anderen Variablen signifikant korrelieren. Es muss allerdings darauf hingewiesen werden, dass die Korrelation zwischen zwei Variablen keine Aussage über die Zuverlässigkeit einer Faktorenanalyse machen kann, da dort auch Abhängigkeiten zwischen mehreren Merkmalen erfasst werden können[81].

Signifikanz (1-seitig)

	EW Anzahl Einwohner zum 31.12.2006	EW_HH Einwohner pro Haushalt	SVB_pro_Einw Sozialversp flichtig Besch. am Wohnort pro Einwohner	ALQ Arbeitslosenquote	KKI Kaufkraftindex je EW	km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	SVB_EW_30km	Dichte Bevölkerungsdichte	Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern
EW		,000	,030	,002	,329	,000	,352	,000	,000	,000
EW_HH	,000		,087	,000	,019	,084	,000	,000	,000	,000
SVB_pro_Einw	,030	,087		,000	,000	,241	,000	,065	,452	,000
ALQ	,002	,000	,000		,000	,001	,000	,003	,000	,000
KKI	,329	,019	,000	,000		,000	,000	,000	,020	,002
km30_EwGfK_antlg	,000	,084	,241	,001	,000		,000	,000	,000	,000
SVB_EW_30km	,352	,000	,000	,000	,000	,000		,072	,000	,120
Dichte	,000	,000	,065	,003	,000	,000	,072		,000	,000
Anteil_Whg_efh	,000	,000	,452	,000	,020	,000	,000	,000		,000
Anteil_Whg_zfh	,000	,000	,000	,000	,002	,000	,120	,000	,000	

Tabelle 32: Signifikanzniveau der Korrelation

[80] Vgl. Kapitel 2.5.1

[81] Vgl. PFEI S. 210

5.6.2.2 KMO-Maß und Bartlett-Test[82]

Mit Hilfe des KMO-Maßes und des Bartlett-Test kann die Gesamtheit der einbezogenen Variablen auf ihre Tauglichkeit für eine Faktorenanalyse überprüft werden.

KMO- und Bartlett-Test

Maß der Stichprobeneignung nach Kaiser-Meyer-Olkin.		,706
Bartlett-Test auf Sphärizität	Ungefähres Chi-Quadrat	1836,582
	df	36
	Signifikanz nach Bartlett	,000

Tabelle 33: KMO-Maß und Bartlett-Test

Das KMO-Maß dient als Maß für die Angemessenheit einer Stichprobe und setzt die einfachen Korrelationskoeffizienten mit den partiellen Korrelationskoeffizienten in Beziehung. Für diesen Datensatz nimmt das KMO-Maß einen Wert von 0,706 an, womit die Stichprobe als „mittelprächtig“ geeignet angesehen werden kann.

Der Bartlett-Test überprüft, ob die Korrelationsmatrix der Grundgesamtheit einer Einheitsmatrix entspricht und die Variablen somit in ihrer Erhebungsgesamtheit unkorreliert sind. Für den Datensatz „Mietspiegel.sav“ erbrachte der Bartlett-Test eine Prüfgröße von 1836,582 bei einem Signifikanzniveau von 0,000. Dies bedeutet, dass mit einer Wahrscheinlichkeit von nahezu 100% davon auszugehen ist, dass die Variablen untereinander korrelieren.

5.6.2.3 Image-Analyse nach Guttman[83]

Die Image-Analyse nach Guttman wird ebenfalls als Kriterium zur Variablenauswahl herangezogen. Guttman unterteilt die Varianz eines Merkmals in das Image und das Anti-Image. Der Varianzanteil, der durch die übrigen Variablen anhand einer Regressionsanalyse erklärt werden kann, wird als Image bezeichnet. Das Anti-Image stellt den von den übrigen Variablen unabhängigen Teil dar.

Auf der Hauptdiagonalen der Anti-Image-Korrelationsmatrix ist das variablenspezifische KMO-Maß (MSA-Wert) aufgetragen. Die Variablen „SVB_EW_30km“ (sozialversicherungspflichtig Beschäftigte je Einwohner in 30km Umkreis) und „km30_EwGfK_antlg“ (Einwohner in 30km Umkreis) weisen einen Wert < 0,5 auf und müssen sukzessive aus dem Modell entfernt werden. Zuerst wird die Variable mit dem kleinsten MSA-Wert entfernt („SVB_EW_30km“).

[82] Vgl. Kapitel 2.5.2 und 2.5.3

[83] Vgl. Kapitel 2.5.4

Anti-Image- Korrelation

	EW Anzahl Einwohner zum 31.12.2006	EW_HH Einwohner pro Haushalt	SVB_pro_Einw Sozialverspflichtig Besch. am Wohnort pro Einwohner	ALQ Arbeitslosenquote	KKI Kaufkraftindex je EW	km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	SVB_EW_30km	Dichte Bevölkerungsdichte	Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern
EW	,742(a)	-,002	-,037	-,059	,044	,054	,020	-,499	-,094	-,049
EW_HH	-,002	,763(a)	,101	,303	,274	-,319	-,144	,254	-,399	-,169
SVB_pro_Einw	-,037	,101	,679(a)	,064	-,106	-,216	-,497	,191	-,068	-,073
ALQ	-,059	,303	,064	,674(a)	,604	-,125	,086	,253	,126	,448
KKI	,044	,274	-,106	,604	,546(a)	-,508	-,293	,007	-,264	,188
km30_EwGfK_antlg	,054	-,319	-,216	-,125	-,508	,415(a)	,584	-,364	,344	-,001
SVB_EW_30km	,020	-,144	-,497	,086	-,293	,584	,407(a)	-,051	,533	,048
Dichte	-,499	,254	,191	,253	,007	-,364	-,051	,719(a)	,249	,260
Anteil_Whg_efh	-,094	-,399	-,068	,126	-,264	,344	,533	,249	,712(a)	-,063
Anteil_Whg_zfh	-,049	-,169	-,073	,448	,188	-,001	,048	,260	-,063	,835(a)

Tabelle 34: Anti-Image-Korrelationsmatrix

Nachdem die Variable „SVB_EW_30km“ aus dem Modell entfernt wurde ergibt sich für die Anti-Image-Korrelationsmatrix folgendes Bild: Alle Variablen weisen nun einen MSA-Wert größer 0,5 auf und können somit für die Faktorenanalyse verwendet werden. Die Variablen „Anteil_Whg_efh“ und „Anteil_Whg_zfh“ können mit Werten größer 0,8 sogar als „verdienstvoll“ geeignet angesehen werden, wogegen die Variablen „KKI“ und „km30_EwGfK_antlg“ nur als kläglich anzusehen sind.

Anti-Image-Korrelation

	EW Anzahl Einwohner zum 31.12.2006	EW_HH Einwohner pro Haushalt	SVB_pro_Einw Sozialverspflichtig Besch. am Wohnort pro Einwohner	ALQ Arbeitslosenquote	KKI Kaufkraftindex je EW	km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	Dichte Bevölkerungsdichte	Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern
EW	,737(a)	,001	-,031	-,061	,052	,052	-,499	-,124	-,050
EW_HH	,001	,782(a)	,034	,320	,245	-,292	,250	-,385	-,164
SVB_pro_Einw	-,031	,034	,659(a)	,124	-,304	,106	,191	,268	-,057
ALQ	-,061	,320	,124	,639(a)	,661	-,216	,259	,095	,446
KKI	,052	,245	-,304	,661	,546(a)	-,434	-,008	-,133	,212
km30_EwGfK_antlg	,052	-,292	,106	-,216	-,434	,565(a)	-,412	,048	-,036
Dichte	-,499	,250	,191	,259	-,008	-,412	,695(a)	,327	,263
Anteil_Whg_efh	-,124	-,385	,268	,095	-,133	,048	,327	,813(a)	-,105
Anteil_Whg_zfh	-,050	-,164	-,057	,446	,212	-,036	,263	-,105	,830(a)

a Maß der Stichprobeneignung

Tabelle 35: Anti-Image-Korrelationsmatrix

Dziuban und Shirkey[84] schlagen vor, eine Korrelationsmatrix dann als ungeeignet zu betrachten, wenn der Anteil der Nicht-Diagonal-Elemente der Anti-Image-Kovarianzmatrix, die einen kritischen Wert von 0,09 überschreiten, über 25% liegt.

[84] Vgl. BACK S. 275f.

Anti-Image- Kovarianz

	EW Anzahl Einwohner zum 31.12.2006	EW_HH Einwohner pro Haushalt	SVB_pro_Einw Sozialverspflichtig Besch. am Wohnort pro Einwohner	ALQ Arbeitslosenquote	KKI Kaufkraftindex je EW	km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	Dichte Bevölkerungsdichte	Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern
EW	,611	,001	-,020	-,022	,022	,029	-,197	-,058	-,023
EW_HH	,001	,315	,016	,085	,073	-,117	,071	-,128	-,055
SVB_pro_Einw	-,020	,016	,664	,048	-,131	,062	,079	,129	-,028
ALQ	-,022	,085	,048	,225	,166	-,073	,062	,027	,126
KKI	,022	,073	-,131	,166	,281	-,164	-,002	-,042	,067
km30_EwGfK_antlg	,029	-,117	,062	-,073	-,164	,510	-,149	,020	-,015
Dichte	-,197	,071	,079	,062	-,002	-,149	,255	,098	,079
Anteil_Whg_efh	-,058	-,128	,129	,027	-,042	,020	,098	,351	-,037
Anteil_Whg_zfh	-,023	-,055	-,028	,126	,067	-,015	,079	-,037	,353

Tabelle 36: Anti-Image-Kovarianzmatrix

Dieses Kriterium wird von den in die Untersuchung einfließenden Variablen nicht erfüllt. Hier liegt der Anteil der Nicht-Diagonal-Element mit einem Wert > 0,09 bei fast 28%. Trotzdem soll die Korrelationsmatrix für eine Faktorenanalyse verwendet werden, da bei den anderen Prüfgrößen die Kriterien erfüllt werden.

5.6.3 Bestimmung der Anzahl der zu extrahierenden Faktoren

Die Kommunalitäten beschreiben den Teil der Gesamtvarianz einer Variablen, der durch die extrahierten Faktoren erklärt werden kann. Diese werden bei der Hauptkomponentenanalyse anfänglich immer auf Eins gesetzt[85]. Das bedeutet, dass die Varianz jeder Variablen bei einer 9-faktoriellen Lösung vollständig erklärt wird. Da nach der Extraktion der Faktoren die

[85] Vgl. Kapitel 2.7.1

Variablenanzahl größer sein sollte als die Anzahl der gefundenen Faktoren, kann nicht die gesamte Varianz erklärt werden und die Kommunalitäten nehmen Werte kleiner Eins an.

Kommunalitäten

	Anfänglich	Extraktion
EW Anzahl Einwohner zum 31.12.2006	1,000	,538
EW_HH Einwohner pro Haushalt	1,000	,814
SVB_pro_Einw Sozialverspflichtig Besch. am Wohnort pro Einwohner	1,000	,825
ALQ Arbeitslosenquote	1,000	,842
KKI Kaufkraftindex je EW	1,000	,838
km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	1,000	,691
Dichte Bevölkerungsdichte	1,000	,856
Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	1,000	,798
Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	1,000	,735

Extraktionsmethode: Hauptkomponentenanalyse.

Tabelle 37: SPSS-Output der anfänglichen und endgültigen Kommunalitäten

Tabelle 37 kann entnommen werden, dass durch alle extrahierten Faktoren 85,6% der Varianz der Variablen „Dichte“ erklärt werden kann. Am schlechtesten kann die Variable „EW“, mit nur 53,8% erklärter Varianz, durch die extrahierten Faktoren beschrieben werden.

Die Tabelle „Erklärte Gesamtvarianz“ enthält die Eigenwerte, die Varianzanteile die jeder einzelne Faktor erklärt sowie die kumulierten Gesamtvarianzaufklärungen der Faktoren. Die Eigenwerte der Faktoren berechnen sich aus der Summe der quadrierten Faktorladungen eines Faktors über alle Variablen und beschreiben den Varianzbeitrag eines Faktors in Hinblick auf die Varianz aller Faktoren.

Erklärte Gesamtvarianz

Komponente	Anfängliche Eigenwerte		
	Gesamt	% der Varianz	Kumulierte %
1	3,634	40,383	40,383
2	2,259	25,102	65,485
3	1,044	11,603	77,088
4	,709	7,877	84,965
5	,429	4,768	89,733
6	,365	4,060	93,793
7	,251	2,793	96,585
8	,189	2,102	98,688
9	,118	1,312	100,000

Extraktionsmethode: Hauptkomponentenanalyse.

Tabelle 38: SPSS-Ausgabe der erklärten Gesamtvarianz

Tabelle 38 ist zu entnehmen, dass z.B. die erste Komponente mit einem Eigenwert von 3,634 einen Varianzanteil von 40,383% erklärt. Die zweite Komponente erklärt 25,102% der Restvarianz und zusammen mit der ersten Komponente 65,485% der Varianz. Da nur drei

der anfänglich neun Faktoren einen Eigenwert größer Eins aufweisen, resultiert nach dem Kaiser-Kriterium eine 3-faktorielle Lösung.

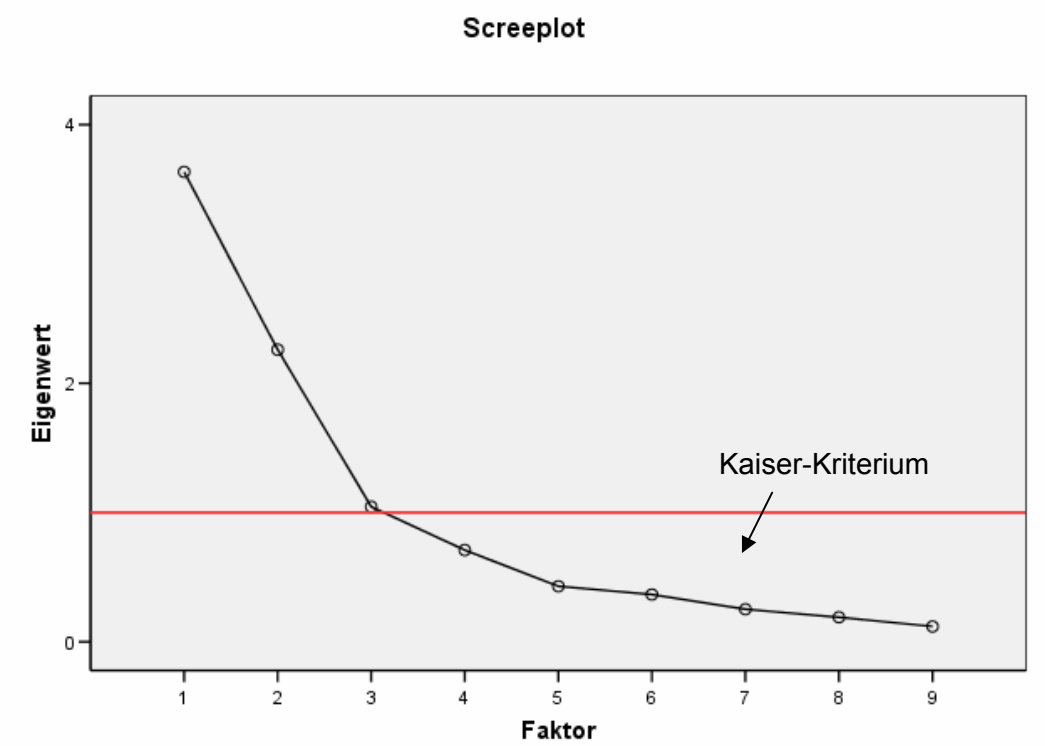


Abbildung 21: Scree-Plot und eingezeichnetes Kaiser-Kriterium

Als weitere Entscheidungshilfe für die Bestimmung der geeigneten Faktorenanzahl wird der Scree-Plot herangezogen. Auf der y-Achse sind die der Größe nach sortierten Eigenwerte abgetragen und auf der x-Achse die jeweils zugehörigen Faktornummern. Der Scree-Plot zeigt beim dritten Eigenwert einen Knick und legt somit eher eine Extraktion von zwei Faktoren nahe. In dieser Ausarbeitung soll das Kaiser-Kriterium als das relevantere Kriterium verwendet werden. Somit ergibt sich die Extraktion von drei Faktoren, die 77,088% der Gesamtvarianz erklären.

5.6.4 Auswertung der Ergebnisse

5.6.4.1 Reproduzierte Korrelationsmatrix und Residualmatrix[86]

Nachdem die Faktorladungsmatrix ermittelt ist, kann nun die reproduzierte Korrelationsmatrix bestimmt werden, um zu beurteilen wie gut die Korrelationsmatrix *R* durch diese abgebildet wird. Da die Korrelationsmatrix *R* nur dann vollständig reproduziert werden kann, wenn alle Faktoren extrahiert werden, treten mehr oder weniger große Abweichungen auf. Diese gibt SPSS als Residuen in der unteren Matrix aus.

Reproduzierte Korrelationen

		EW Anzahl Einwohner zum 31.12.2006	EW_HH Einwohner pro Haushalt	SVB_pro_Einw Sozialversp flichtig Besch. am Wohnort pro Einwohner	ALQ Arbeitslosenquote	KKI Kaufkraftindex je EW	km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	Dichte Bevölkerungsdichte	Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern
Reproduzierte Korrelation	EW	,538(b)	-,347	-,270	,174	,091	,487	,633	-,357	-,420
	EW_HH	-,347	,814(b)	-,001	-,563	,162	-,101	-,591	,802	,750
	SVB_pro_Einw	-,270	-,001	,825(b)	-,485	,576	,022	-,079	-,061	,178
	ALQ	,174	-,563	-,485	,842(b)	-,711	-,281	,189	-,493	-,592
	KKI	,091	,162	,576	-,711	,838(b)	,519	,239	,078	,226
	km30_EwGfK_antlg	,487	-,101	,022	-,281	,519	,691(b)	,605	-,153	-,150
	Dichte	,633	-,591	-,079	,189	,239	,605	,856(b)	-,617	-,610
	Anteil_Whg_efh	-,357	,802	-,061	-,493	,078	-,153	-,617	,798(b)	,730
	Anteil_Whg_zfh	-,420	,750	,178	-,592	,226	-,150	-,610	,730	,735(b)

[86] Vgl. Kapitel 2.7.3

		EW Anzahl Einwohner zum 31.12.2006	EW_HH Einwohner pro Haushalt	SVB_pro_Einw Sozialversp flichtig Besch. am Wohnort pro Einwohner	ALQ Arbeitslosenquote	KKI Kaufkraftindex je EW	km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	Dichte Bevölkerungsdichte	Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern
Residuum(a)	EW		,012	,171	-,024	-,067	-,232	-,031	,045	,094
	EW_HH	,012		,074	,040	-,051	,028	,016	-,068	-,067
	SVB_pro_Einw	,171	,074		,071	-,112	,015	-,002	,054	,019
	ALQ	-,024	,040	,071		,014	,121	-,044	,041	-,001
	KKI	-,067	-,051	-,112	,014		-,047	-,039	,031	-,070
	km30_EwGfK_antlg	-,232	,028	,015	,121	-,047		-,080	-,030	-,026
	Dichte	-,031	,016	-,002	-,044	-,039	-,080		,004	,042
	Anteil_Whg_efh	,045	-,068	,054	,041	,031	-,030	,004		-,098
	Anteil_Whg_zfh	,094	-,067	,019	-,001	-,070	-,026	,042	-,098	

Extraktionsmethode: Hauptkomponentenanalyse.

a Residuen werden zwischen beobachteten und reproduzierten Korrelationen berechnet. Es liegen 15 (41,0%) nicht redundante Residuen mit absoluten Werten größer 0,05 vor.

b Reproduzierte Kommunalitäten

Tabelle 39: SPSS-Ausgabe der reproduzierten Korrelationsmatrix und Residualmatrix

Auf der Hauptdiagonalen der reproduzierten Korrelationsmatrix können die endgültigen Kommunalitäten abgelesen werden. Die Nichtdiagonalelemente geben die durch die Faktorenstruktur reproduzierten Korrelationen wieder. Die Residuen geben die Differenz zwischen den ursprünglichen und den reproduzierten Korrelationen wieder.

Der reproduzierte Korrelationskoeffizient zwischen der Variablen „Dichte“ und der Variablen „Sozialversicherungspflichtig Beschäftigte am Wohnort je Einwohner“ weicht um -0,002, also minimal, vom ursprünglichen Korrelationskoeffizienten ab. Allerdings weisen 41% der Residuen einen Wert > 0,05 auf, was darauf schließen lässt, dass das Modell die ursprünglichen Daten nicht optimal darstellt.

5.6.4.2 Interpretation der gefundenen Faktoren[87]

Nachdem die Anzahl der Faktoren bestimmt wurde, soll nun geklärt werden, welche inhaltliche Bedeutung diese drei Faktoren besitzen. Um die Interpretation der Faktoren zu erleichtern soll die Komponentenmatrix mittels Rotation möglichst in die sog. „Einfachstruktur“ überführt werden. Die Rotation erfolgt orthogonal nach dem Varimax-Verfahren, da die ermittelten Faktoren für die anschließende Regressionsanalyse unabhängig voneinander sein müssen.

Rotierte Komponentenmatrix(a)

	Komponente		
	1	2	3
EW_HH Einwohner pro Haushalt	,891		
Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	,875		
Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	,792		
km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis		,801	
Dichte Bevölkerungsdichte	-,553	,741	
EW Anzahl Einwohner zum 31.12.2006		,655	
SVB_pro_Einw Sozialverspflichtig Besch. am Wohnort pro Einwohner			,878
KKI Kaufkraftindex je EW			,781
ALQ Arbeitslosenquote	-,620		-,655

Extraktionsmethode: Hauptkomponentenanalyse.
Rotationsmethode: Varimax mit Kaiser-Normalisierung.
a Die Rotation ist in 7 Iterationen konvergiert.

Tabelle 40: Rotierte Komponentenmatrix

Anhand der rotierten Komponentenmatrix kann nun die inhaltliche Interpretation der Faktoren vorgenommen werden. In der Praxis hat sich die Konvention durchgesetzt, dass eine Variable ab einer Ladungshöhe von |0,5| einem Faktor zugeordnet werden kann. Zur besseren Übersicht wurden deshalb die Faktorladungen der Größe nach sortiert und solche mit einem betragsmäßigen Wert von < 0,5 nicht angezeigt[88].

[87] Vgl. Kapitel 2.8

[88] SPSS-Befehl: /FORMAT SORT BLANK(.50)

Aus Tabelle 40 lässt sich ablesen, dass auf den ersten Faktor die Variablen „Einwohner pro Haushalt“, „Anteil Wohnungen in Einfamilienhäusern“ und „Anteil Wohnungen in Zweifamilienhäusern“ stark positiv laden, während die Variablen „Bevölkerungsdichte“ und „Arbeitslosenquote“ einen negativen Einfluss haben. Dieser Faktor nimmt also in Regionen mit

- vielen Personen pro Haushalt (Familien),
- einem hohen Anteil an Ein- und Zweifamilienhäusern,
- einer geringen Bevölkerungsdichte und
- einer geringen Arbeitslosigkeit

tendenziell hohe Werte an. Diese Konstellationen sind in Deutschland vor allem in den ländlichen Regionen zu finden.

Auf den zweiten Faktor laden die Variablen „Einwohnerzahl in 30km Umkreis“, „Bevölkerungsdichte“ und „Anzahl der Einwohner“ positiv. Das bedeutet, dass der zweite Faktor tendenziell hohe Werte in Regionen mit

- vielen Einwohnern in 30km Umkreis,
- einer hohen Bevölkerungsdichte und
- einer hohen Einwohnerzahl

annimmt. Dieser Faktor kann daher als „Ballungsindikator“ angesehen werden.

Der dritte der extrahierten Faktoren weist hohe positive Ladungen bei den Variablen „Sozialversicherungspflichtig Beschäftigte am Wohnort pro Einwohner“ und beim „Kaufkraftindex“ auf. Die „Arbeitslosenquote“ hat hier einen negativen Einfluss. Der dritte Faktor weist also hohe Werte in Regionen mit

- einem hohen Anteil sozialversicherungspflichtig Beschäftigter,
- hoher Kaufkraft und
- geringer Arbeitslosigkeit

auf. Dieser Faktor kann demnach als „Wirtschaftskraft“ interpretiert werden.

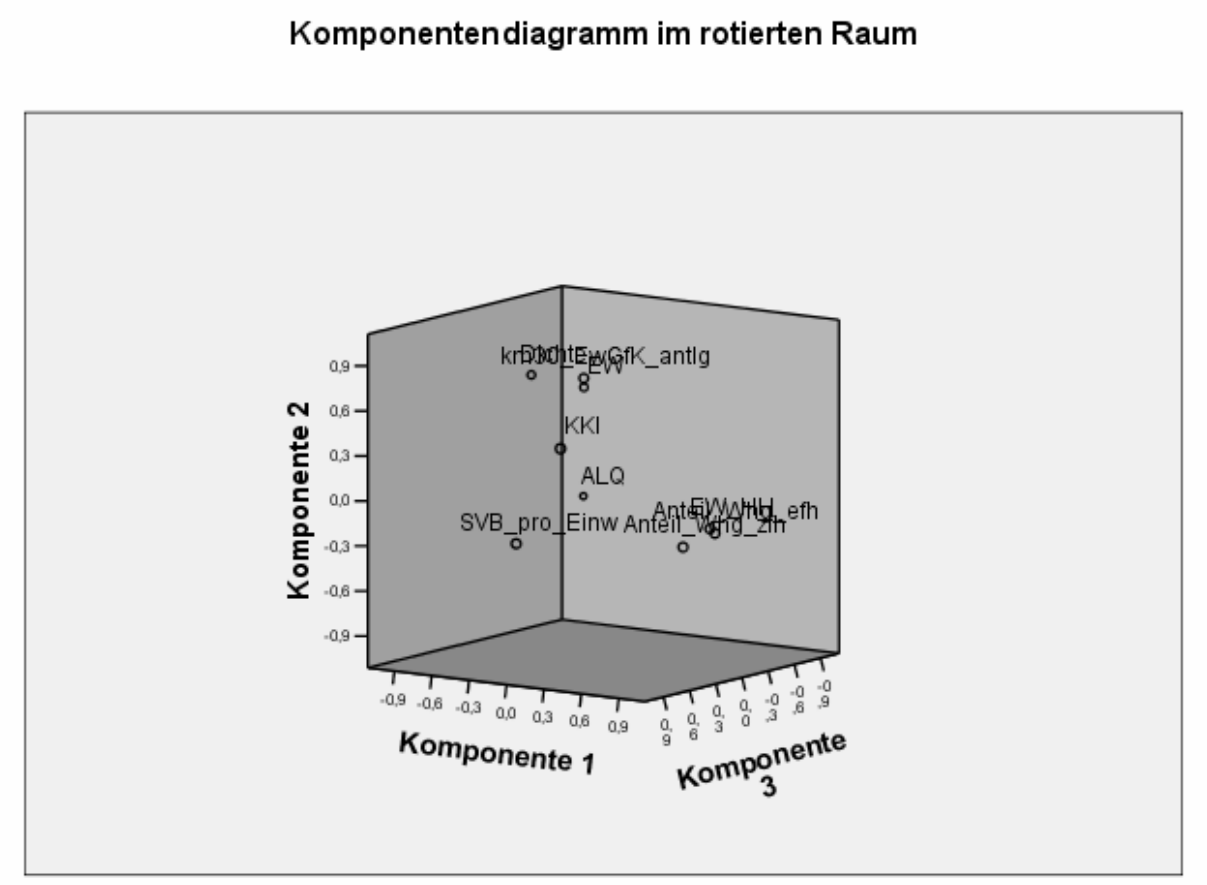


Abbildung 22: Komponentendiagramm im rotierten Raum

Anhand von Abbildung 22 kann entsprechend grafisch veranschaulicht werden, wie die Variablen im dreidimensionalen Faktorraum liegen.

5.6.4.3 Bestimmung der Faktorwerte[89]

Mit Hilfe der Koeffizientenmatrix der Komponentenwerte können die Faktorwerte für die einzelnen Städte und Gemeinden berechnet werden. SPSS gibt die berechneten Werte im Dateneditor aus.

[89] Vgl. Kapitel 2.9

Koeffizientenmatrix der Komponentenwerte

	Komponente		
	1	2	3
EW Anzahl Einwohner zum 31.12.2006	,054	,376	-,167
EW_HH Einwohner pro Haushalt	,353	,101	-,118
SVB_pro_Einw Sozialverspflichtig Besch. am Wohnort pro Einwohner	-,256	-,306	,593
ALQ Arbeitslosenquote	-,184	-,128	-,250
KKI Kaufkraftindex je EW	,031	,176	,364
km30_EwGfK_antlg Einwohner-zahl in 30km Umkreis	,130	,458	-,004
Dichte Bevölkerungsdichte	-,090	,333	,001
Anteil_Whg_efh Anteil Wohnun-gen in Einfamilienhäusern	,352	,084	-,156
Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	,254	-,017	,023

Extraktionsmethode: Hauptkomponentenanalyse.
Rotationsmethode: Varimax mit Kaiser-Normalisierung.

Tabelle 41: Koeffizientenmatrix der Komponentenwerte

Für den ersten Faktor ergibt sich die folgende Gleichung:

$$FACTOR\_1 = 0{,}054 \cdot EW + 0{,}353 \cdot EW\_HH - 0{,}256 \cdot SVB\_pro\_EW - 0{,}184 \cdot ALQ + 0{,}031 \cdot KKI + 0{,}130 \cdot km30\_EwGfK\_antlg - 0{,}090 \cdot Dichte + 0{,}352 \cdot Anteil\_Whg\_efh + 0{,}254 \cdot Anteil\_Whg\_zfh$$

Für die beiden anderen Faktoren werden die Gleichungen auf die gleiche Weise erstellt. Aus den ursprünglich neun Variablen sind nun drei Faktoren extrahiert worden, die 77,088% der Information enthalten.

5.6.5 Überprüfung der Merkmale anhand einer 50%-Zufallsstichprobe

Nachdem die Faktoren extrahiert und sinnvoll interpretiert wurden, soll nun anhand einer 50%-Zufallsstichprobe die Qualität der Untersuchung überprüft werden. Dazu wird die Stichprobe zufällig aufgeteilt und das Ergebnis anhand der beiden Teilstichproben kontrolliert. Wünschenswert hierbei ist es, dass sich ein ähnliches Bild wie bei der Gesamtstichprobe ergibt, da die Ergebnisse sonst als zufallsbedingt zu bewerten sind.

5.6.5.1 Variablenauswahl

Das KMO-Maß für die beiden Teilstichproben mit einem Umfang von 188 und 168 Fällen fällt erstaunlich gut aus mit Werten von 0,694 bzw. 0,691, was als mäßig betrachtet werden kann.

KMO- und Bartlett-Test

Maß der Stichprobeneignung nach Kaiser-Meyer-Olkin.		,694
Bartlett-Test auf Sphärizität	Ungefähres Chi-Quadrat	956,613
	df	36
	Signifikanz nach Bartlett	,000

Tabelle 42: KMO-Maß und Bartlett-Test der ersten 50%-Zufallsstichprobe

KMO- und Bartlett-Test

Maß der Stichprobeneignung nach Kaiser-Meyer-Olkin.		,691
Bartlett-Test auf Sphärizität	Ungefähres Chi-Quadrat	925,441
	df	36
	Signifikanz nach Bartlett	,000

Tabelle 43: KMO-Maß der zweiten 50%-Zufallsstichprobe

Auch der Bartlett-Test fällt bei beiden Stichproben signifikant aus.

Bei den Anti-Image-Korrelationsmatrizen ergibt sich das folgende Bild:

- Bei der 50%-Zufallsstichprobe mit den 188 Fällen fällt das MSA-Maß der Variablen „Kaufkraftindex“ mit 0,497 knapp unter die 0,5 Richtlinie. Die Variable wird trotzdem nicht aus der Analyse entfernt.
- Bei der zweiten 50%-Zufallsstichprobe mit den 168 Fällen ändern sich die MSA-Werte auf der Hauptdiagonalen leicht im Vergleich zur Gesamtstichprobe, es bleiben aber alle Werte > 0,5.

5.6.5.2 Bestimmung der Anzahl der zu extrahierenden Faktoren

Aus beiden untersuchten Teilstichproben resultiert nach dem Kaiser-Kriterium eine Extraktion von drei Faktoren, die jeweils 77,167% bzw. 78,192% der Gesamtvarianz erklären. Es ergibt sich somit ein nahezu identisches Bild zu der Gesamtstichprobe.

5.6.5.3 Vergleich der extrahierten Faktoren

Die drei extrahierten Faktoren aus der Gesamtstichprobe werden in den beiden 50%-Stichproben fast identisch abgebildet. Die auf die Faktoren hochladenden Variablen variieren auch in ihrer Höhe kaum. Dies deutet darauf hin, relativ gute Faktoren gefunden zu haben.

5.7 Regressionsanalyse für Modell 2

Die eben erhobenen Faktoren sollen nun mit den restlichen Variablen in einer Regressionsanalyse auf Art und Größe ihres Einflusses untersucht werden.

Die im Vorfeld getroffenen Annahmen für das erste Regressionsmodell (fehlende Werte, Signifikanzniveau etc.) sollen zur besseren Vergleichbarkeit auch für das zweite Modell gelten, so dass auf eine ausführliche Beschreibung verzichtet werden kann und auf Kapitel 5.4 verwiesen wird. In den folgenden Abschnitten werden die Ergebnisse der Regressionsanalyse kurz erläutert.

5.7.1 Überprüfung der Voraussetzungen[90]

Auch das zweite Regressionsmodell muss nun sorgfältig daraufhin überprüft werden, ob alle Modellprämissen eingehalten wurden. Ist das gefundene Modell valide, kann es inhaltlich interpretiert werden.

5.7.1.1 Multikollinearität

	Kollinearitätsstatistik	
	Toleranz	VIF
(Konstante)		
FAC1_1 REGR factor score 1 for analysis 1	,909	1,100
FAC2_1 REGR factor score 2 for analysis 1	,853	1,172
FAC3_1 REGR factor score 3 for analysis 1	,976	1,024
gtyp2 Kern- und Großstädte unter 500.000 E. in Agglomerationsräume	,862	1,160
gtyp8 sonstige Gemeinden in ländlichen Kreisen von Agglomerationsräumen	,980	1,021
gtyp14 OZ/MZ in Kreisen höherer Dichte von Ländlichen Räumen	,949	1,053
gtyp15 sonstige Gemeinden in Kreisen höherer Dichte von Ländlichen Räumen	,930	1,075

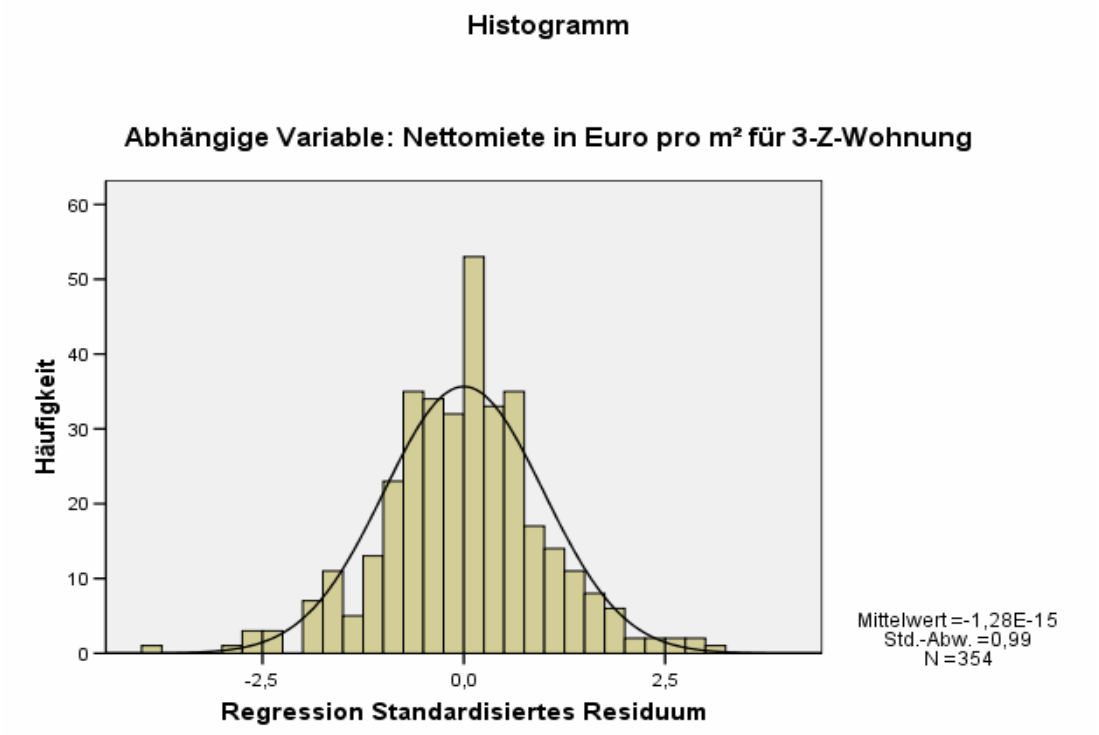
Tabelle 44: Kollinearitätsstatistik

In der hier betrachteten Regression liegen alle Toleranzwerte nahe bei Eins, der geringste Wert liegt bei 0,853. Auch der *VIF* nimmt bei allen signifikanten Variablen einen Wert um Eins an. Somit kann Multikollinearität eindeutig ausgeschlossen werden.

5.7.1.2 Normalverteilung der Residuen

Eine weitere Modellannahme ist die Normalverteilung der Residuen. Die grafische Überprüfung ergibt das folgende Ergebnis:

[90] Vgl. Kapitel 3.6



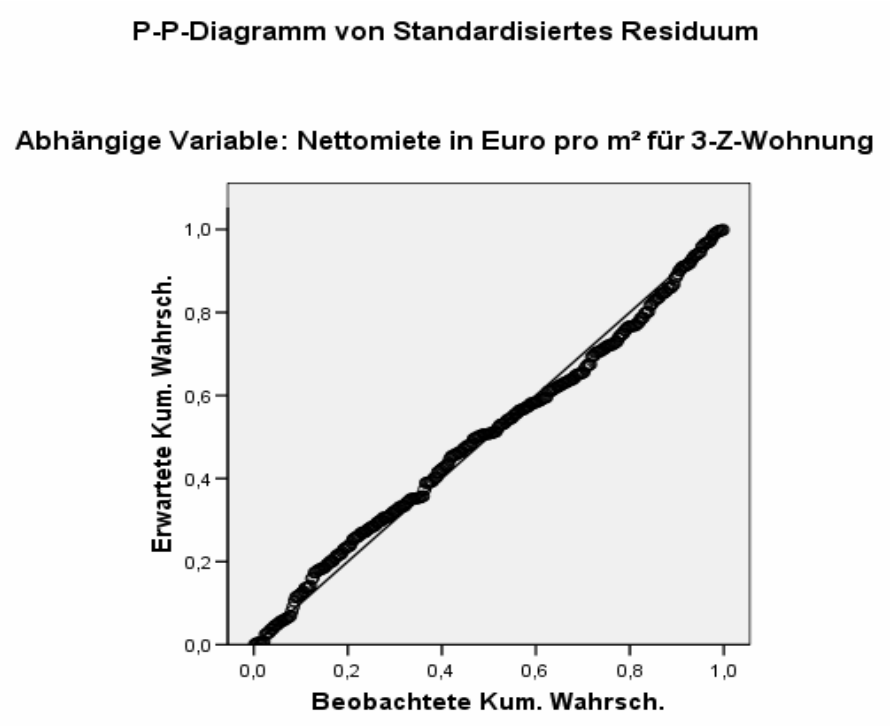


Abbildung 23: Histogramm und PP-Plot der standardisierten Residuen

Das Histogramm zeigt, dass sich die standardisierten Residuen recht gut an eine Standardnormalverteilung anpassen. In der Spitze und an den Rändern sind kleinere Abweichungen

erkennbar. Geringfügige Abweichungen sind in der Praxis aber nicht ungewöhnlich und durchaus tolerierbar. Der Erwartungswert liegt nahe 0, die Standardabweichung bei 0,99.

Das PP-Diagramm zeigt, dass die standardisierten Residuen der abhängigen Variablen mit ganz leichten Abweichungen der Referenzverteilung (Normalverteilung) folgen.

Das Vorliegen einer Normalverteilung der standardisierten Residuen als Voraussetzung einer linearen Regressionsanalyse kann somit als erfüllt betrachtet werden.

5.7.1.3 Heteroskedastizität

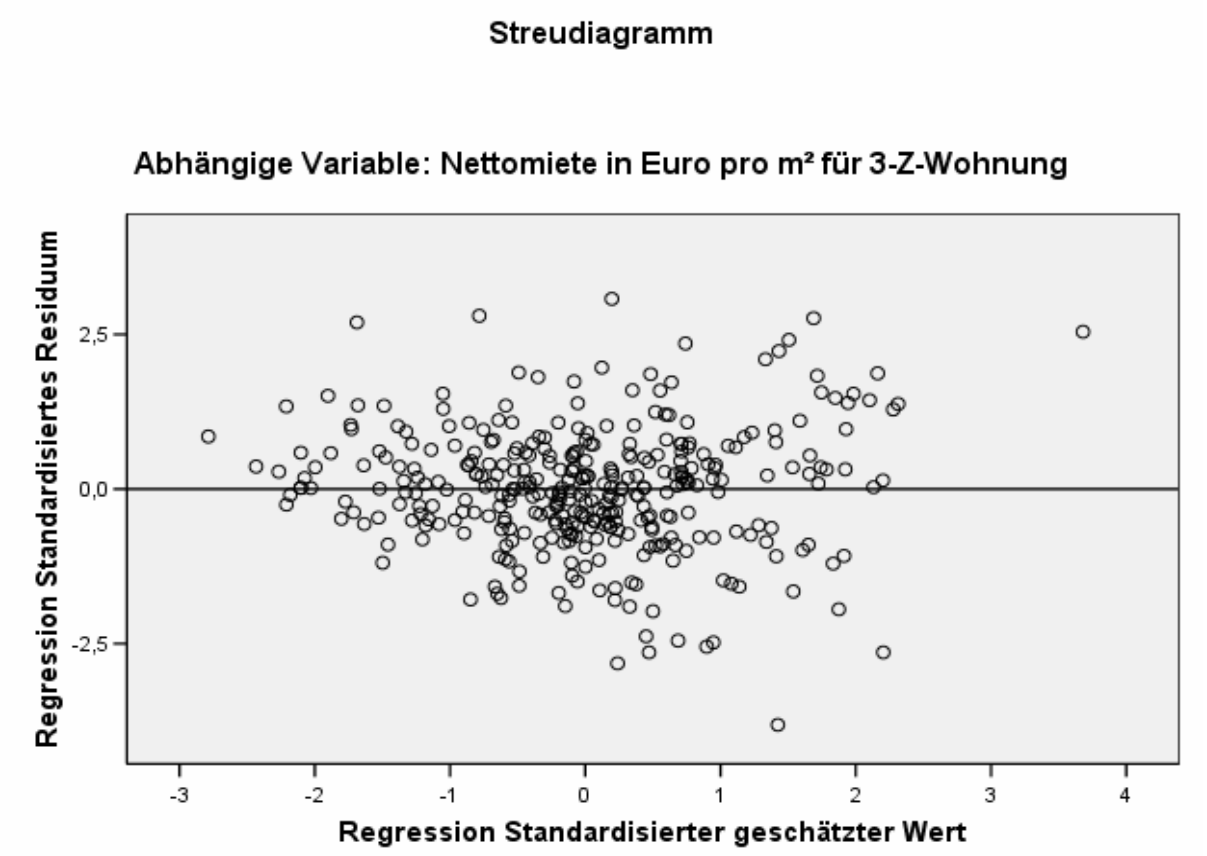


Abbildung 24: Streudiagramm für den Test auf Varianzgleichheit

Die grafische Überprüfung der Residuen zeigt auch hier, dass kaum Streuung (Heteroskedastizität) in den Daten vorliegt. Die Extrembereiche streuen etwas breiter, aber noch im akzeptablen Bereich, so dass hier keine Verletzung der Modellannahme vorliegt. Die Auswertung der Streudiagramme der einzelnen Variablen ergab ebenfalls keine Auffälligkeiten.

5.7.1.4 Autokorrelation

Der Index-Wert zur Prüfung auf Autokorrelation liegt in diesem Modellansatz bei 1,999 (s. Tabelle 45), so dass Autokorrelation ausgeschlossen werden kann.

5.7.2 Überprüfung der Modellgüte[91]

Das Bestimmtheitsmaß R^2 des eben ermittelten Regressionsmodells liegt bei 0,43, d.h. durch das Modell kann 43% der (senkrechten) Streuung in den y-Werten erklärt werden.

Modellzusammenfassung(o)

R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Durbin-Watson-Statistik
,656(n)	,430	,418	,82453	1,999

n Einflußvariablen : (Konstante), gtyp15 sonstige Gemeinden in Kreisen höherer Dichte von Ländlichen Räumen, gtyp2 Kern- und Großstädte unter 500.000 E. in Agglomerationsräume, gtyp8 sonstige Gemeinden in ländlichen Kreisen von Agglomerationsräumen, gtyp14 OZ/MZ in Kreisen höherer Dichte von Ländlichen Räumen, FAC3_1 REGR factor score 3 for analysis 1, FAC1_1 REGR factor score 1 for analysis 1, FAC2_1 REGR factor score 2 for analysis 1
o Abhängige Variable: nmqm_3Z Nettomiete in Euro pro m² für 3-Z-Wohnung

Tabelle 45: SPSS-Output der Modellbeurteilung

Das korrigierte R-Quadrat ($\overline{R}^2$) beträgt für dieses Regressionsmodell 41,8%. Der Standardfehler des Schätzers liegt im Vergleich zum ersten Regressionsmodell (0,78968) höher und nimmt in diesem Regressionsmodell einen Wert von 0,82453 an. Bezogen auf den Mittelwert der abhängigen Variablen (5,3710) beträgt er 15,4%.

ANOVA(o)

	Quadrat-summe	df	Mittel der Quadrate	F	Signifikanz
Regression	177,388	7	25,341	37,274	,000(n)
Residuen	235,230	346	,680		
Gesamt	412,618	353			

n Einflußvariablen : (Konstante), gtyp15 sonstige Gemeinden in Kreisen höherer Dichte von Ländlichen Räumen, gtyp2 Kern- und Großstädte unter 500.000 E. in Agglomerationsräume, gtyp8 sonstige Gemeinden in ländlichen Kreisen von Agglomerationsräumen, gtyp14 OZ/MZ in Kreisen höherer Dichte von Ländlichen Räumen, FAC3_1 REGR factor score 3 for analysis 1, FAC1_1 REGR factor score 1 for analysis 1, FAC2_1 REGR factor score 2 for analysis 1
o Abhängige Variable: nmqm_3Z Nettomiete in Euro pro m² für 3-Z-Wohnung

Tabelle 46: SPSS-Ausgabe der Varianzanalyse

Die Quadratsumme der Residuen (SSE) beträgt in dem Modellansatz 235,230, die Quadratsumme der erklärten Abweichungen (SSR) nimmt einen Wert von 177,388 an. Die geschätzte Restvarianz beträgt 0,680.

Die Nullhypothese des F-Tests (in der Grundgesamtheit besteht kein linearer Zusammenhang zwischen Regressand und Regressoren) kann zu dem im Voraus festgelegten Signifikanzniveau von 1% verworfen werden.

[91] Vgl. Kapitel 3.4

5.7.3 Überprüfung der Koeffizienten[92]

Als nächstes sind nun die Regressionskoeffizienten der einzelnen Variablen zu überprüfen. Dies geschieht mittels der T-Statistik und den angegebenen Konfidenzintervallen.

Koeffizienten(a)

	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B	
	B	Standardfehler	Beta			Untergrenze	Obergrenze
(Konstante)	5,477	,048		114,169	,000	5,382	5,571
FAC1_1 REGR factor score 1 for analysis 1	-,087	,046	-,081	-1,894	,059	-,178	,003
FAC2_1 REGR factor score 2 for analysis 1	,401	,048	,371	8,443	,000	,308	,494
FAC3_1 REGR factor score 3 for analysis 1	,489	,044	,451	10,989	,000	,401	,576
gtyp2 Kern- und Großstädte unter 500.000 E. in Agglomerationsräume	-,342	,196	-,077	-1,750	,081	-,727	,042
gtyp8 sonstige Gemeinden in ländlichen Kreisen von Agglomerationsräumen	-,790	,298	-,109	-2,653	,008	-1,376	-,204
gtyp14 OZ/MZ in Kreisen höherer Dichte von Ländlichen Räumen	-,702	,249	-,118	-2,826	,005	-1,191	-,214
gtyp15 sonstige Gemeinden in Kreisen höherer Dichte von Ländlichen Räumen	-,937	,226	-,175	-4,152	,000	-1,380	-,493

a Abhängige Variable: nmqm_3Z Nettomiete in Euro pro m² für 3-Z-Wohnung

Tabelle 47: SPSS-Output der Koeffizientenstatistik

Im zweiten Regressionsmodell üben der zweite und dritte Faktor einen hohen Einfluss auf die abhängige Variable aus. Ihre standardisierten Koeffizienten liegen bei 0,371 und 0,451. Damit üben beide einen positiven Einfluss aus. Für beide Variablen fallen außerdem die Konfidenzintervalle schmal aus, was auf eine gute Schätzung hindeutet. Die Qualität des

[92] Vgl. Kapitel 3.5

Koeffizienten „gtyp2“ ist dagegen anzuzweifeln, da die untere Intervallgrenze im negativen Bereich liegt und die Intervallobergrenze im positiven Bereich.

Die Schätzfunktion für die Nettomiete in Euro pro m² (nmqm) unter diesem Modell lautet:

$$nmqm = 5{,}477 - 0{,}087 \cdot FACTOR\_1 + 0{,}401 \cdot FACTOR\_2 + 0{,}489 \cdot FACTOR\_3 - 0{,}342 \cdot gtyp2 - 0{,}790 \cdot gtyp8 - 0{,}702 \cdot gtyp14 - 0{,}937 \cdot gtyp15$$

5.7.4 Überprüfung auf räumliche Korrelation[93]

Zum Schluss sollen auch für das zweite Regressionsmodell die Residuen grafisch daraufhin untersucht werden, ob sich räumliche Muster erkennen lassen.

Bei den Residuen des zweiten Regressionsmodells lassen sich eindeutig regionale Muster identifizieren. Besonders in Ostdeutschland werden die Nettomieten verstärkt überschätzt, die prosperierenden Ballungsräume werden dagegen eher unterschätzt. Im Ruhrgebiet lässt sich eine Unterschätzung des „Ballungszentrum“ erkennen, die Randgebiete werden dagegen überschätzt. Eine unsystematische Verteilung der Residualwerte ist hier nicht zu erkennen, vielmehr entsteht eine starke Clusterung nach großen Regionen, was darauf hindeutet, dass das Modell die Ursprungsdaten nicht optimal darstellt.

[93] Vgl. Kapitel 5.4.5

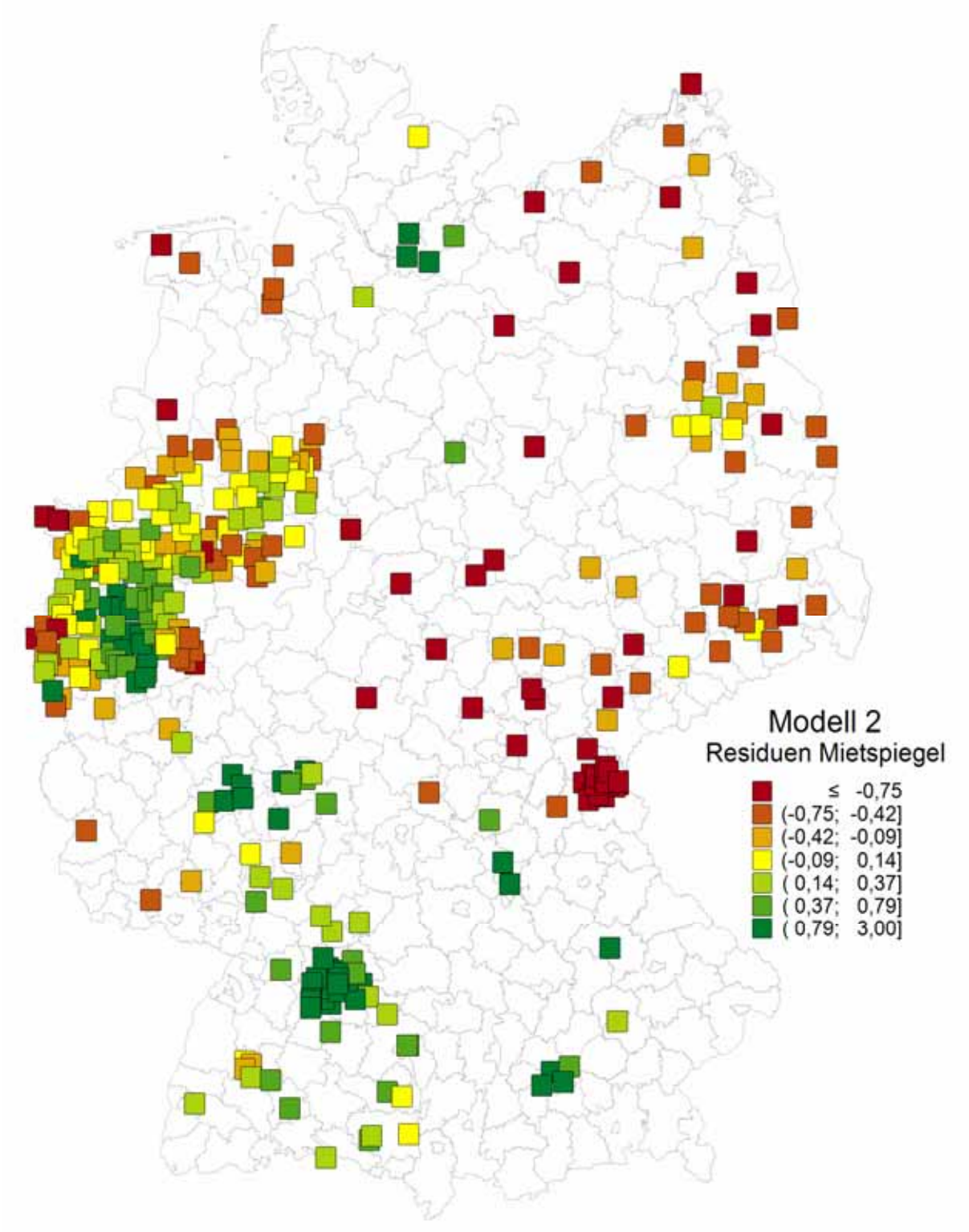


Abbildung 25: Darstellung der Residuen aus dem Mietspiegel-Modell 2[94]

[94] Quelle: Eigene Darstellung

5.8 Überprüfung der Modellstrukturen anhand einer 50%-Zufallsstichprobe

Um zu überprüfen, wie zufällig die gefundene Modellstruktur ist, soll Modell 1 anhand einer 50%-Zufallsstichprobe berechnet werden. Die Teilstichprobenbildung erfolgt analog der Überprüfung der Faktorenanalyse, weshalb auf Kapitel 5.5.5 verwiesen wird.

5.8.1 Vergleich der Voraussetzungen

Das Vorhandensein von Autokorrelation kann für die beiden Teilstichproben, mit einem Umfang von 189 und 168 Fällen, ausgeschlossen werden. Die jeweiligen Durbin-Watson-Statistiken liegen bei 1,708 und 2,058.

Bei Überprüfung der Normalverteilung der Residuen fällt auf, dass sich die Erwartungswerte und Standardabweichungen der 50%-Zufallsstichproben nur unwesentlich von denen der Gesamtstichprobe unterscheiden. Die grafische Analyse zeigt, dass sich die standardisierten Residuen bei beiden Stichproben relativ gut an eine Standardnormalverteilung anpassen, wie nachfolgend zu sehen ist:

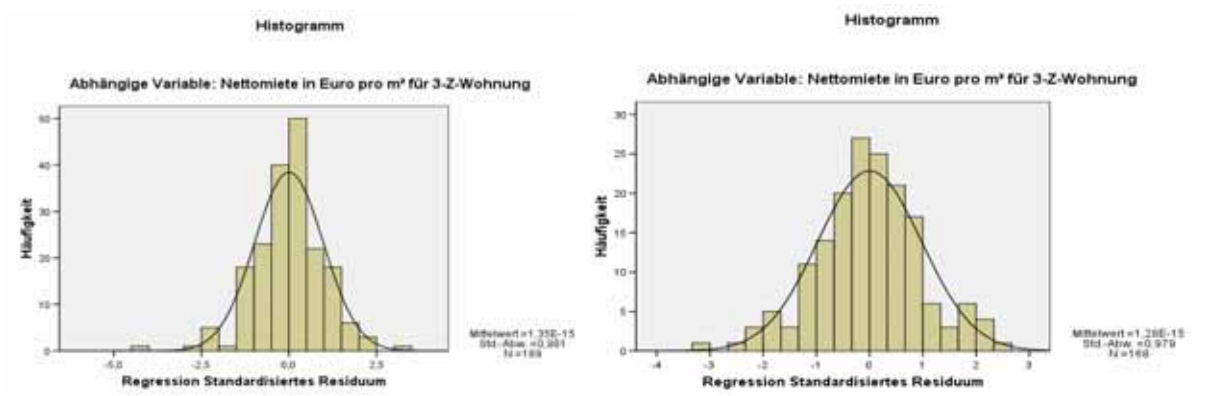


Abbildung 26: Histogramm der standardisierten Residuen (1. und 2. Teilstichprobe)

5.8.2 Vergleich der Modellgüte

Die Zufallsstichprobe mit 189 Fällen verbessert sich bei der Regressionsanalyse um knapp 6% beim R^2 (53,1%) und um knapp 5% beim korrigierten R-Quadrat $\overline{R}^2$ (51,3%). Die zweite Teilstichprobe mit 168 Fällen verschlechtert sich dagegen beim R^2 um 3% auf 44,6%, das $\overline{R}^2$ verschlechtert sich um 4,5% auf 42,2%. Die Fehlerquadratsummen können nicht mit der Gesamtstichprobe verglichen werden, da sich die Anzahl der Fälle halbiert hat.

Die Nullhypothese des F-Tests (in der Grundgesamtheit besteht kein linearer Zusammenhang zwischen Regressand und Regressoren) kann in beiden Teilstichproben zum Signifikanzniveau von 1% verworfen werden.

5.8.3 Vergleich der Koeffizienten

Die Modellstruktur in den Stichproben hat sich folgendermaßen verändert:

- Die ermittelten Koeffizienten haben sich betragsmäßig erhöht oder verringert. Einzig der Koeffizient des Kaufkraftindexes ist bei beiden Teilstichproben relativ stabil geblieben mit Werten von 0,053 bzw. 0,043 (Gesamtstichprobe: 0,049).
- Bei der 50%-Zufallsstichprobe mit den 189 Fällen werden die Variablen „Einwohner pro Haushalt“, „Anteil Wohnungen in Zweifamilienhäusern“ und „Gemeindetyp 14“ nicht mehr signifikant.
- Bei der zweiten 50%-Zufallsstichprobe mit 168 Fällen werden die „Einwohner pro Haushalt“, „Anteil Wohnungen in Einfamilienhäusern“ und „Bevölkerungsdichte“ nicht mehr signifikant.

Die Ergebnisse des Regressionsmodells haben sich als nicht sehr stabil erwiesen, was zum einen mit der geringen Fallzahl in den Teilstichproben zusammenhängen kann. Die Werte der einzelnen Koeffizienten sind daher mit einer gewissen Vorsicht zu interpretieren. Ein anderer Grund könnten die korrelativen Zusammenhänge zwischen einigen Variablen sein, wodurch sich eine Faktorenanalyse wiederum als sinnvoll rausstellen könnte.

6 Untersuchung des IVD-Datensatzes

Als nächstes sollen die gleichen Untersuchungen mit dem Datensatz des Immobilienverbands Deutschland durchgeführt werden und die Ergebnisse mit denen aus der Untersuchung der Mietspiegel verglichen werden[95].

Dieser Datensatz enthält die Mietwerte für 351 deutsche Städte und Gemeinden. Als Referenzwohnungstyp wird hier eine 3-Zimmer-Wohnung in Gebäuden mit Fertigstellung ab 1949 und gutem Wohnwert herangezogen, um die Vergleichbarkeit mit der sanierten 3-Zimmer-Wohnung aus dem Mietspiegel-Datensatz zu gewährleisten.

Als potenzielle erklärende Variable fließen die gleichen Merkmale ein, wie in die Untersuchung zuvor. Die beiden Variablen „Sozialversicherungspflichtig Beschäftigte am Wohnort“ und „Sozialversicherungspflichtig Beschäftigte in 30km Umkreis“ sind hier bereits durch die neu gebildeten Variablen „Sozialversicherungspflichtig Beschäftigte je Einwohner“ und „Sozialversicherungspflichtig Beschäftigte je Einwohner in 30km Umkreis“ ersetzt worden[96].

Es soll zunächst ebenfalls eine Regressionsanalyse durchgeführt werden. Im Anschluss daran wird versucht, mögliche Faktoren mittels einer Hauptkomponentenanalyse zu extrahieren, die dann wiederum in einer daran anschließenden Regressionsanalyse untersucht werden sollen.

6.1 Vergleich IVD-Daten mit Mietspiegel-Daten

Bevor mit der Analyse des IVD-Datensatzes begonnen wird, soll zunächst der Unterschied zu den Mietspiegel-Daten erläutert werden.

Vergleicht man beide Datensätze miteinander, so liegen für 150 Städte und Gemeinden sowohl vom IVD erhobene Mietwerte als auch durch die Auswertung von Mietspiegeln berechnete Nettomieten vor. Theoretisch sollten die vom IVD aufgestellten Nettomieten über denen der Mietspiegel liegen, da der IVD Neuvertragsmieten erhebt. Bei den Mietspiegeln sind dagegen die Bestandsmieten der letzten vier Jahre zu Grunde gelegt.

Eine Analyse der beiden Datensätze ergab, dass bei knapp einem Drittel der Daten die Nettomieten aus den Mietspiegel-Auswertungen über denen des IVD liegen. Ein Grund dafür kann die Wohnungseinschränkung sein, die bei der Auswertung der Mietspiegel vorgenommen wurde. Durch die Annahme, dass die 3-Z-Whg gut saniert wurde, ergaben sich bei vielen Mietspiegel-Berechnungen mehr oder weniger hohe Modernisierungszuschläge oder die Wohnung konnte einer neueren Baualtersklasse zugeordnet werden. Dies kann zu einem verfälschten Bild führen, da dieser Wohnungstyp dann bei der Berechnung der Nettomiete höhere Zuschläge erhält und somit der ermittelte Mietwert steigt. Des Weiteren existieren keine genauen Angaben darüber, wie der IVD Modernisierungen bei der Erhebung berücksichtigt. Die Grafik lässt zudem keine regionale Systematik bei den Differenzen zwischen

[95] Beschreibung der IVD-Daten unter 4.3

[96] Vgl. Kapitel 5.3

den Datensätzen erkennen. Daher werden alle Daten in die nachfolgenden Analysen mit einbezogen.

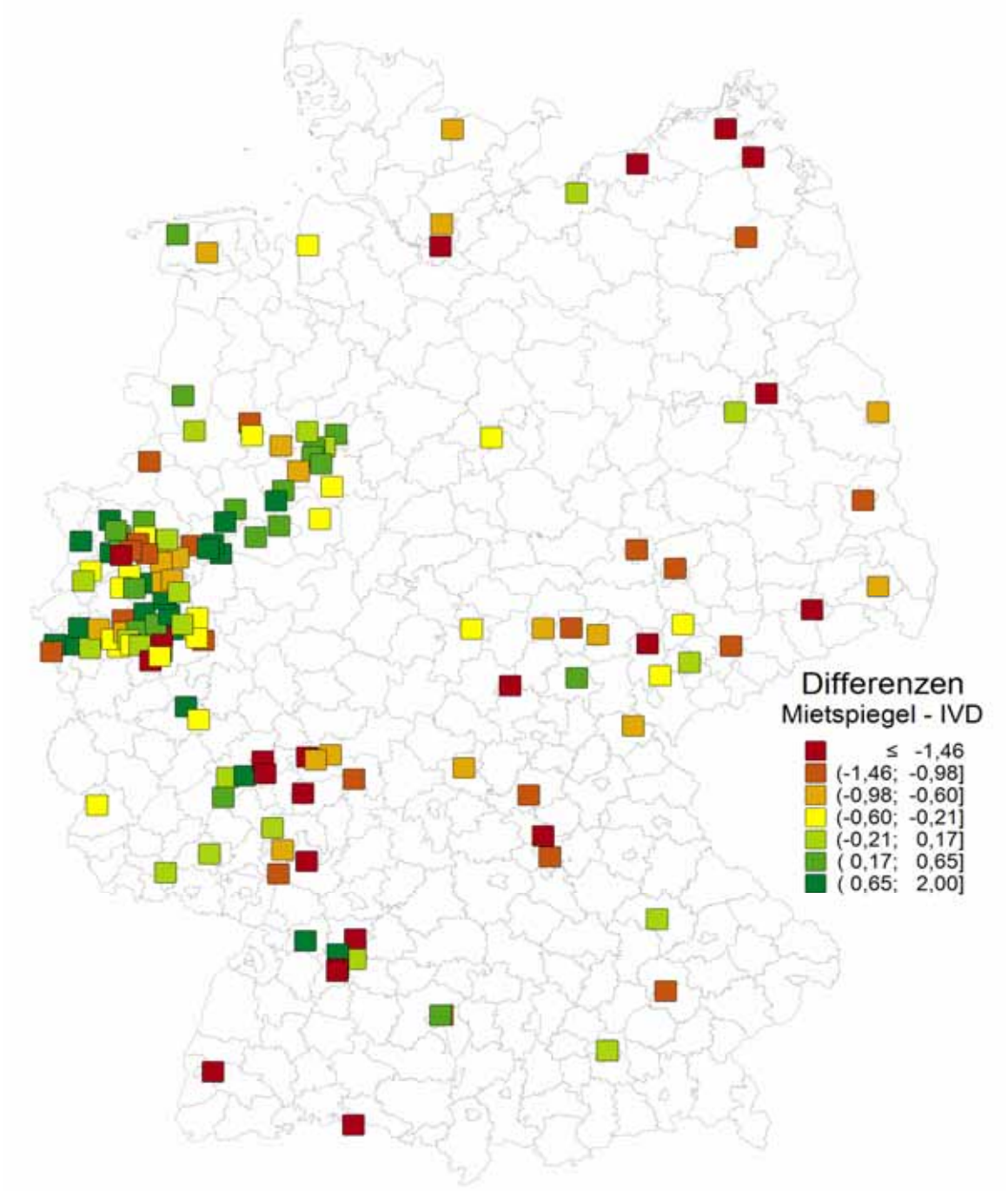


Abbildung 27: Differenz Mietspiegel-Mieten und IVD-Mieten[97]

[97] Quelle: Eigene Berechnungen

6.2 Übersicht über den Datensatz

Zunächst kann mit Hilfe der Tabelle „Statistiken“ ein kleiner Überblick über die Variablen gewonnen werden.

Statistiken

	N		Mittelwert	Standardabweichung	Minimum	Maximum
	Gültig	Fehlend				
nmqm49_gWW Nettomiete in €/m².Fertigstellung ab 1949: guter Wohnwert	345	6	5,8398	1,24279	3,90	12,00
siedl_gemtyp siedlungsstruktureller Gemeindetyp	349	2	7,30	4,343	1	16
EW Anzahl Einwohner zum 31.12.2006	349	2	101310,46	240207,565	2630	3404037
EW_HH Einwohner pro Haushalt	349	2	2,087324	,1750931	1,7242	2,6703
SVB_pro_Einw Sozialverspflichtig Besch. am Wohnort pro Einwohner	349	2	,306579	,0290548	,0871	,4212
ALQ Arbeitslosenquote	346	5	,130187	,0464186	,0522	,3226
KKI Kaufkraftindex je EW	349	2	100,239925	12,4259696	73,6749	151,3171
km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	348	3	1340437,74	1113404,614	11741	4914471
Dichte Bevölkerungsdichte	349	2	859,16	655,454	104	4171
SVB_EW_30km	348	3	,3148	,01775	,27	,36
Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	349	2	,266151	,1272591	,0710	,6901
Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	349	2	,171183	,0769233	,0228	,3599

Tabelle 48: Deskriptive Statistik der Variablen

Die erhobenen Nettomieten des IVD schwanken pro m² zwischen 3,90€ und 12,00€; die Durchschnittsmiete beträgt 5,84€.

Auch hier ist es wichtig zu beachten, dass alle Fälle, für die keine Nettomiete ausgewiesen ist (Missing-Values), von den weiteren Analysen ausgeschlossen werden. Der Ausschluss erfolgt ebenfalls listenweise (der Fall wird ausgeschlossen, falls eine der Variablen einen fehlenden Wert aufweist).

6.3 Abgeleitete Variablen

6.3.1 Bildung von Dummy-Variablen

Die Variable „Siedlungsstruktureller Gemeindetyp“ (siedl_gemtyp) nimmt in diesem Datensatz Werte von 1 bis 16 an (s. Tabelle 48). Im Gegensatz zum vorherigen Datensatz kommen die Gemeindetypisierungen 8, 15 und 17 hier nicht vor. Daher werden für sie auch keine Dummy-Variablen gebildet. Die anderen Gemeindetypisierungen werden genauso umprogrammiert wie in Kapitel 5.2. beschrieben.

6.3.2 Bildung von Referenzkategorien

Nachdem die Variable „siedl_gemtyp“ in 1/0-codierte Dummy-Variablen aufgelöst wurde, muss nun entschieden werden, welcher Gemeindetyp als Referenzkategorie dienen soll. Dafür werden wiederum die Häufigkeitsverteilungen der einzelnen Gemeindetypen betrachtet:

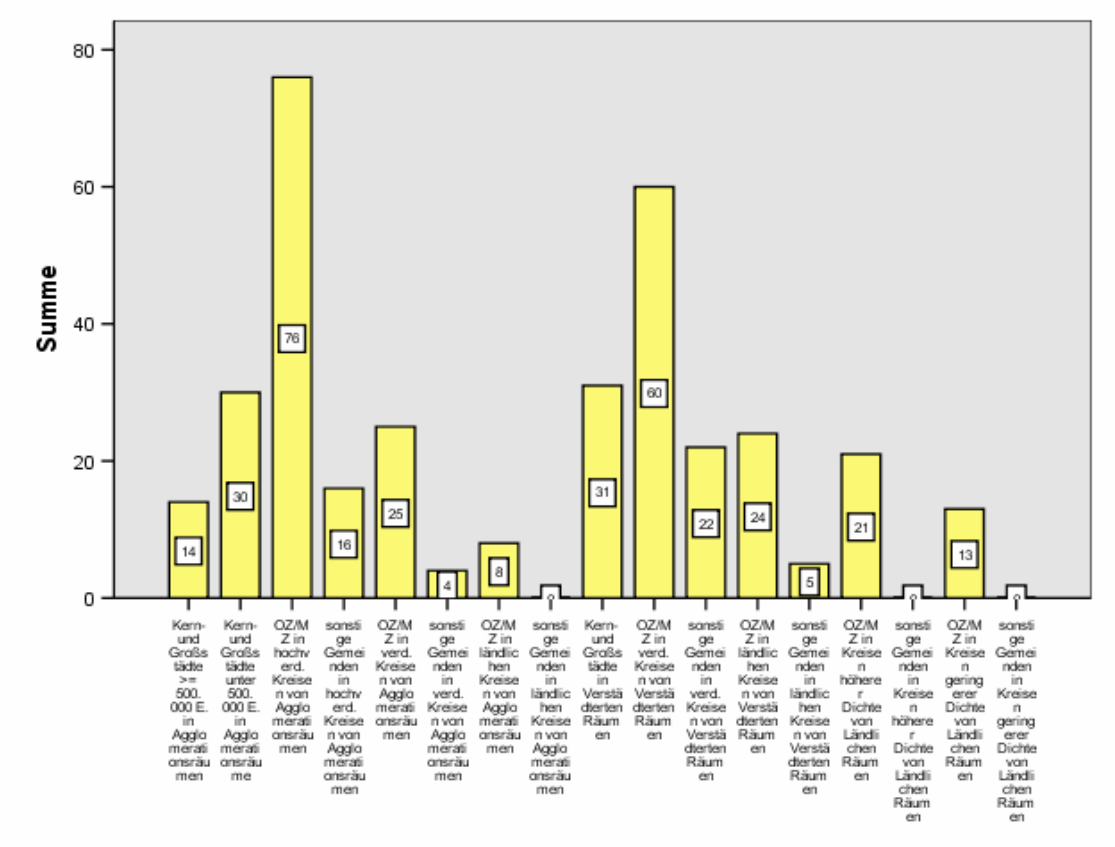


Abbildung 28: Häufigkeitsverteilung der siedlungsstrukturellen Gemeindetypen

Als Referenzkategorie wird ebenfalls der Gemeindetyp 3 (OZ/MZ in hochverdichteten Kreisen von Agglomerationsräumen) ausgewählt, da dieser am häufigsten besetzt ist.

6.4 Analyse verschiedener Modellvarianten

Auch für den IVD-Datensatz wurden verschiedene Modellvarianten analysiert, bevor die nachstehend beschriebenen Modelle 1 und 2 ausgewählt wurden. Die verschiedenen Untersuchungen entsprechen in der Durchführung und dem Ergebnis denen des Mietspiegel-Datensatzes, weshalb diese hier nicht näher erläutert werden sollen, sondern auf Kapitel 5.3. verwiesen wird. Für die Regressionsanalysen wird der Einfluss der unabhängigen Variablen auf die abhängige Variable als linear angenommen. Die Überprüfung der Scatterplots der unabhängigen Variablen mit der abhängigen Variablen hat keinen kurvilinearen oder exponentiellen Zusammenhang erkennen lassen. Die Überprüfung der Residualplots brachte dasselbe Ergebnis. Aus Platzgründen werden die Scatterplots nicht extra aufgeführt.

6.5 Regressionsanalyse für Modell 1

Um ein erstes Modell für den IVD-Datensatz zu bekommen, soll zunächst mittels einer Regressionsanalyse untersucht werden, ob und welchen Einfluss die verschiedenen Variablen auf den Wohnungsmietpreis haben. Die im Vorfeld getroffenen Annahmen für die Regressionsmodelle aus Kapitel 5 (fehlende Werte, Signifikanzniveau etc.) sollen zur besseren Vergleichbarkeit auch für den IVD-Datensatz gelten. Auf eine ausführliche Beschreibung wird deshalb verzichtet und auf das Kapitel 5.4 verwiesen. In den folgenden Abschnitten werden die Ergebnisse der Regressionsanalyse kurz erläutert.

6.5.1 Überprüfung der Voraussetzungen[98]

Das ermittelte Regressionsmodell muss zuerst sorgfältig daraufhin überprüft werden, ob alle Modellprämissen eingehalten wurden. Ist das gefundene Modell valide, kann es inhaltlich interpretiert werden.

6.5.1.1 Multikollinearität

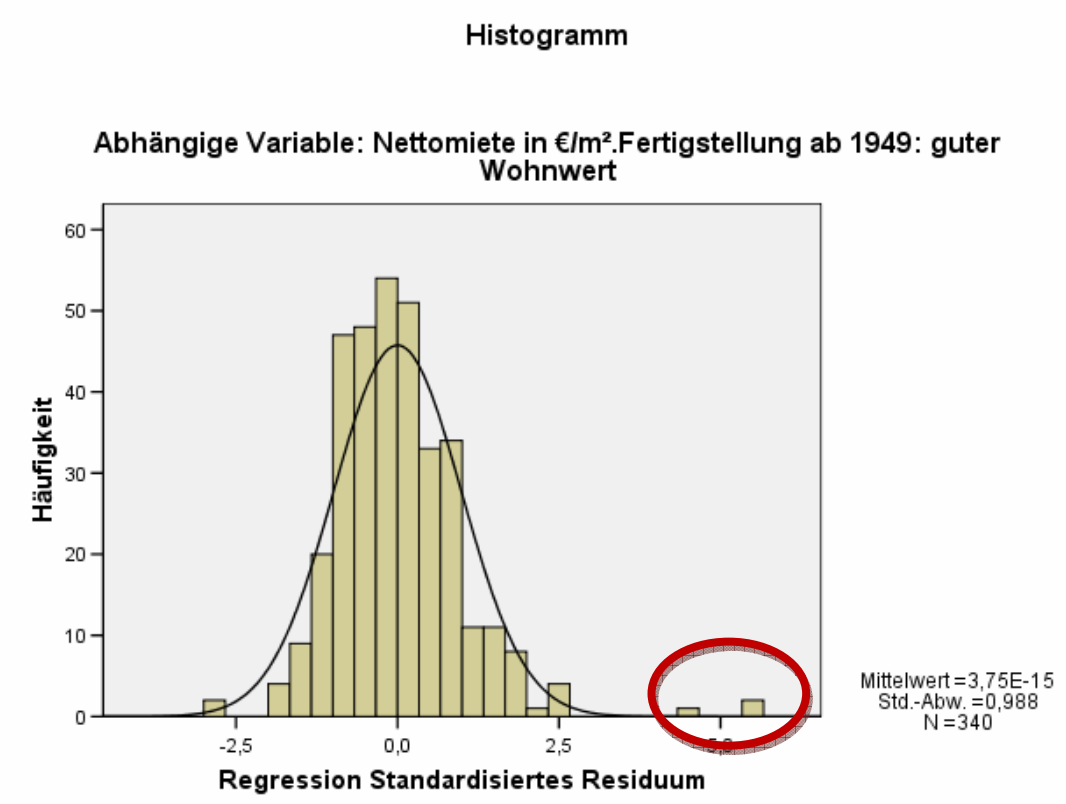
In der hier betrachteten Regression liegt der kleinste Toleranzwert bei 0,325, so dass hier keine Multikollinearität erkennbar ist. Der höchste VIF nimmt einen Wert von 3,073 an. Somit kann Multikollinearität ausgeschlossen werden.

[98] Vgl. Kapitel 3.6

	Kollinearitätsstatistik	
	Toleranz	VIF
(Konstante)		
EW Anzahl Einwohner zum 31.12.2006	,820	1,219
ALQ Arbeitslosenquote	,325	3,073
KKI Kaufkraftindex je EW	,431	2,318
SVB_EW_30km Soz.vers.pflichtig Beschäftigte je Einwohner in 30km Umkreis	,809	1,236
Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	,507	1,972
Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	,512	1,955
gtyp6 sonstige Gemeinden in verd. Kreisen von Agglomerationsräumen	,962	1,039
gtyp11 sonstige Gemeinden in verd. Kreisen von Verstädterten Räumen	,915	1,093

Tabelle 49: Kollinearitätsstatistik

6.5.1.2 Normalverteilung der Residuen



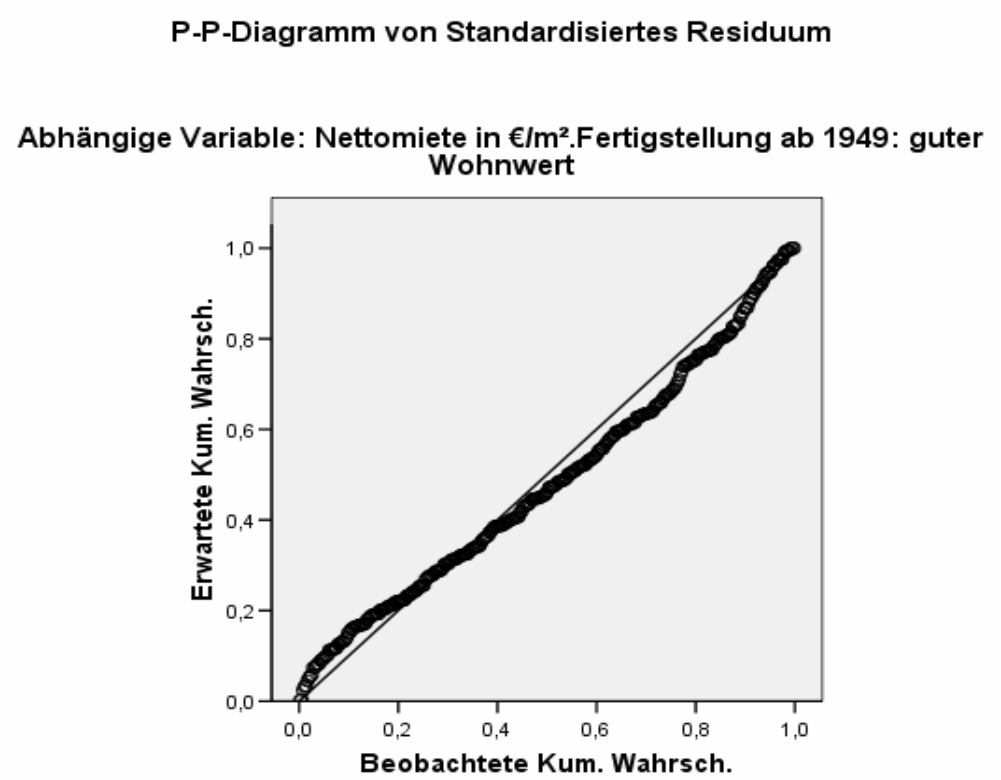


Abbildung 29: Histogramm und PP-Plot der standardisierten Residuen

Bei der Betrachtung des Histogramms der standardisierten Residuen des Regressionsmodells fallen, neben der linksschiefe der Verteilung, die Ausreißer im rechten Teil des Histogramms auf. Auch der PP-Plot zeigt eine leichte Abweichung der Residuen von der Referenzverteilung (Normalverteilung).

Die Analyse der Ausreißer ergibt, dass es sich um die folgenden Städte handelt:

- Borkum (Nettomiete: 11,00€)
- Norderney (Nettomiete: 12,00€)
- Oestrich-Winkel (Nettomiete: 12,00€)

Da es sich bei den Orten um bekannte Touristengebiete handelt, kommt die Idee auf, diesen Faktor in die Untersuchungen mit einzubeziehen. Dies erscheint auch in sofern sinnvoll, da bekanntermaßen in den Orten mit einem hohen Touristenaufkommen auch hohe Mieten verlangt werden. Um diesen Umstand zu berücksichtigen wurde die Variable „Anzahl der Gästeübernachtungen je Einwohner“ gebildet[99] und als unabhängige Variable mit in die Regressionsanalyse aufgenommen. Diese Variante soll nun im weiteren Verlauf untersucht werden und der erste Ansatz wird verworfen.

Auf eine nachträgliche Analyse des Mietspiegel-Datensatzes mit der Variablen „Anzahl der Gästeübernachtungen je Einwohner“ wird im Rahmen dieser Diplomarbeit verzichtet, da für

[99] Vgl. 4.4.1.12

die relevanten kleinen Gemeinden mit vielen Gästeübernachtungen keine Mietspiegel vorliegen.

6.5.1.3 Multikollinearität

Nach Hinzunahme der Variablen „Anzahl der Übernachtungsgäste pro Einwohner“ kann Multikollinearität weiterhin ausgeschlossen werden.

	Kollinearitätsstatistik	
	Toleranz	VIF
(Konstante)		
ALQ Arbeitslosenquote	,319	3,135
KKI Kaufkraftindex je EW	,400	2,498
SVB_EW_30km	,914	1,095
Dichte Bevölkerungsdichte	,577	1,733
Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	,438	2,284
Gast_EW Übernachtungsgäste pro Einwohner im Jahr 2006	,915	1,093
gtyp6 sonstige Gemeinden in verd. Kreisen von Agglomerationsräumen	,971	1,030

Tabelle 50: Kollinearitätsstatistik

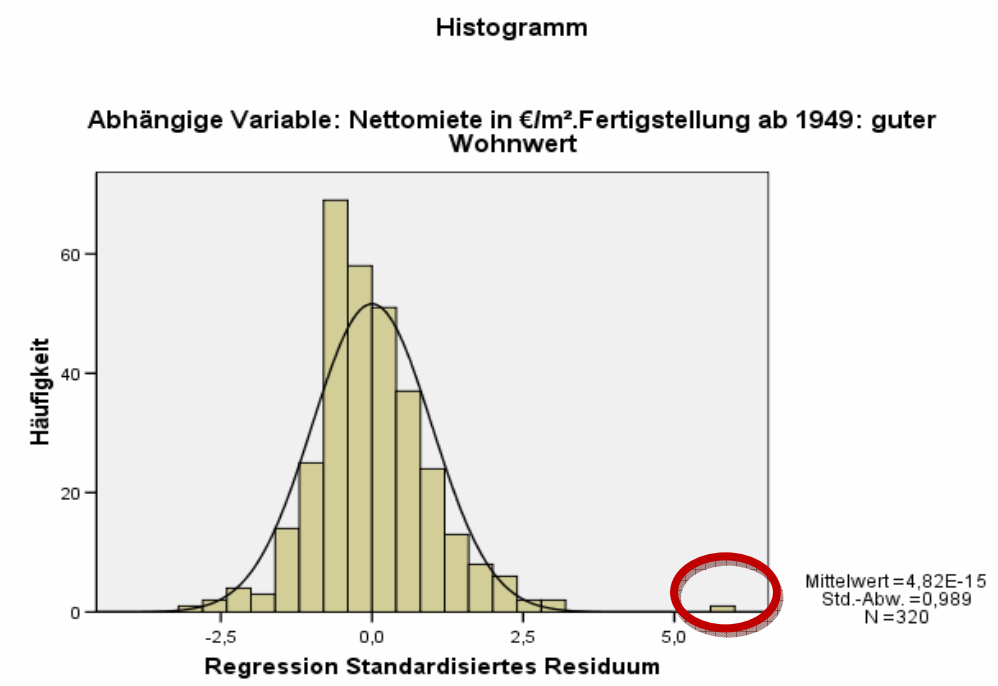
In der hier betrachteten Regression liegt der kleinste Toleranzwert bei 0,319, der höchste VIF nimmt einen Wert von 3,135 an. Dies ist jedoch als unproblematisch zu betrachten.

6.5.1.4 Normalverteilung der Residuen

Das Histogramm zeigt, dass sich die standardisierten Residuen jetzt besser an eine Standardnormalverteilung anpassen. In der Spitze und an den Rändern sind kleinere Abweichungen erkennbar. Geringfügige Abweichungen sind in der Praxis aber nicht ungewöhnlich und durchaus zu tolerieren. Der Erwartungswert liegt nahe 0, die Standardabweichung bei 0,989. Im rechten Randbereich kann aber auch bei diesem Modellansatz ein Ausreißer identifiziert werden, welcher noch genauer zu untersuchen wäre. Im Rahmen dieser Diplomarbeit wird allerdings auf eine Diagnostik der Ausreißer in den Modellen verzichtet und auf weiterführende Untersuchungen verwiesen.

Das PP-Diagramm zeigt, dass die standardisierten Residuen der abhängigen Variablen mit ganz leichten Abweichungen der Referenzverteilung (Normalverteilung) folgen.

Das Vorliegen einer Normalverteilung der standardisierten Residuen als Voraussetzung einer linearen Regressionsanalyse kann somit als erfüllt betrachtet werden.



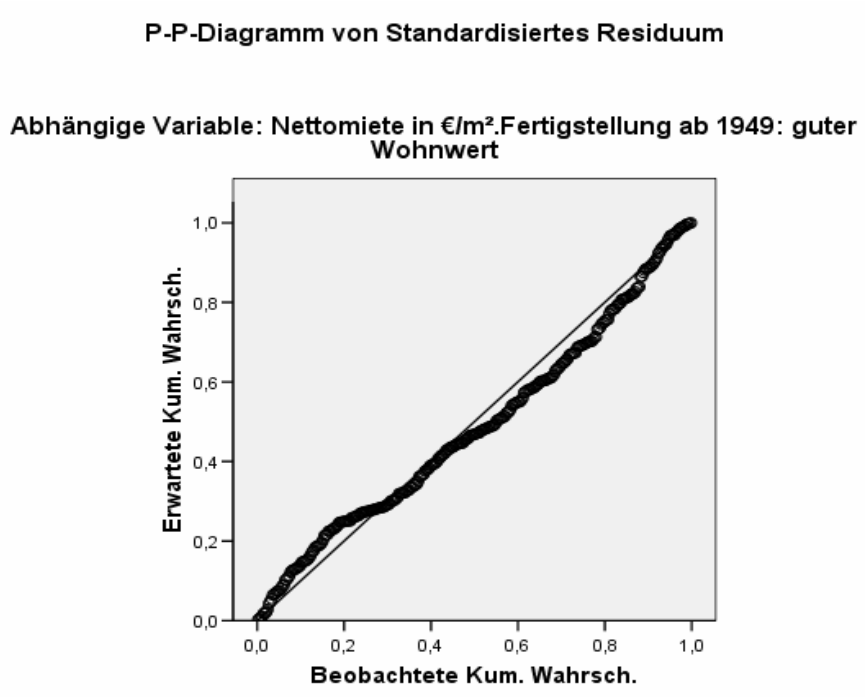


Abbildung 30: Histogramm und PP-Plot der standardisierten Residuen

6.5.1.5 Heteroskedastizität

Die Werte streuen ausgewogen und zufällig um die Nulllinie, d.h. es ist keine Systematik in der Streuung erkennbar. Es liegen somit keine Hinweise auf Heteroskedastizität vor. Die Auswertung der Streudiagramme der einzelnen Variablen ergab ebenfalls keine Auffälligkeiten.

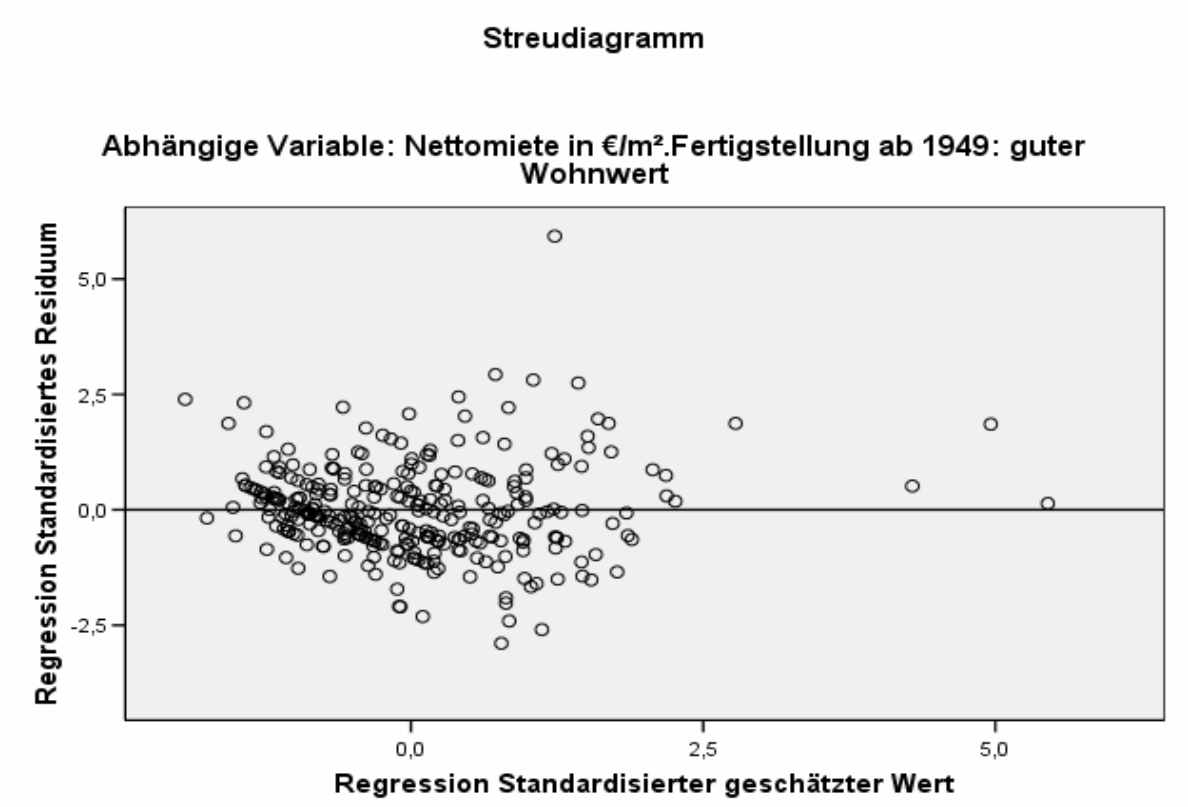


Abbildung 31: Streudiagramm für den Test auf Varianzgleichheit

6.5.1.6 Autokorrelation

Der Index-Wert zur Prüfung auf Autokorrelation liegt in diesem Modellansatz bei 1,982 (s. Tabelle 51). Da dieser Wert nahe bei 2 liegt, kann Autokorrelation ausgeschlossen werden.

6.5.2 Überprüfung der Modellgüte[100]

Das Bestimmtheitsmaß R^2 des eben ermittelten Regressionsmodells liegt bei 0,549, d.h. durch das Modell kann 54,9% der (senkrechten) Streuung in den y-Werten erklärt werden. Das korrigierte R-Quadrat $\overline{R}^2$ dient zur Beurteilung des Modells unter Berücksichtigung der Anzahl der Regressoren. Der Wert dieser Größe beträgt für dieses Regressionsmodell 53,9%. Der Standardfehler des Schätzers nimmt in diesem Regressionsmodell einen Wert von 0,84466 an und beträgt, bezogen auf den Mittelwert der abhängigen Variablen, 14,4%.

[100] Vgl. Kapitel 3.4

Modellzusammenfassung(s)

R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Durbin-Watson-Statistik
,741(r)	,549	,539	,84466	1,982

r Einflussvariablen : (Konstante), SVB_EW_30km, gtyp6 sonstige Gemeinden in verd. Kreisen von Agglomerationsräumen, Gast_EW Übernachtungsgäste pro Einwohner im Jahr 2006, KKI Kaufkraftindex je EW, Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern, ALQ Arbeitslosenquote, Dichte Bevölkerungsdichte
s Abhängige Variable: nmqm49_gWW Nettomiete in €/m².Fertigstellung ab 1949: guter Wohnwert

Tabelle 51: SPSS-Output der Modellbeurteilung

ANOVA(s)

	Quadrat-summe	df	Mittel der Quadrate	F	Signifikanz
Regression	271,263	7	38,752	54,316	,000(r)
Residuen	222,596	312	,713		
Gesamt	493,860	319			

r Einflußvariablen : (Konstante), SVB_EW_30km, gtyp6 sonstige Gemeinden in verd. Kreisen von Agglomerationsräumen, Gast_EW Übernachtungsgäste pro Einwohner im Jahr 2006, KKI Kaufkraftindex je EW, Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern, ALQ Arbeitslosenquote, Dichte Bevölkerungsdichte
s Abhängige Variable: nmqm49_gWW Nettomiete in €/m².Fertigstellung ab 1949: guter Wohnwert

Tabelle 52: SPSS-Ausgabe der Varianzanalyse

Die Quadratsumme der Residuen (SSE) beträgt in dem Modellansatz 222,596, die Quadratsumme der erklärten Abweichungen (SSR) nimmt einen Wert von 271,263 an. Außerdem kann die geschätzte Restvarianz $\hat{s}^2 = 0{,}713$ abgelesen werden. Zur Überprüfung der Signifikanz des Regressionsmodells wird der F-Test herangezogen. Dieser formuliert die Nullhypothese, dass in der Grundgesamtheit kein linearer Zusammenhang zwischen Regressand und Regressoren besteht, also die Regressionskoeffizienten in der Grundgesamtheit Null sind. Diese Hypothese kann zu dem im Voraus festgelegten Signifikanzniveau von 1% verworfen werden.

6.5.3 Überprüfung der Koeffizienten[101]

Als nächstes sind nun die Regressionskoeffizienten der einzelnen Variablen zu überprüfen.

[101] Vgl. Kapitel 3.5

Koeffizienten(a)

	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B	
	B	Standardfehler	Beta			Untergrenze	Obergrenze
(Konstante)	,077	1,212		,063	,950	-2,309	2,462
ALQ Arbeitslosenquote	-3,843	1,814	-,143	-2,118	,035	-7,413	-,273
KKI Kaufkraftindex je EW	,037	,006	,364	6,062	,000	,025	,049
SVB_EW_30km	9,571	2,725	,140	3,513	,001	4,210	14,932
Dichte Bevölkerungsdichte	,000	,000	,211	4,223	,000	,000	,001
Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	-5,399	,929	-,334	-5,809	,000	-7,228	-3,570
Gast_EW Übernachtungsgäste pro Einwohner im Jahr 2006	,022	,003	,316	7,962	,000	,016	,027
gtyp6 sonstige Gemeinden in verd. Kreisen von Agglomerationsräumen	1,390	,431	,124	3,223	,001	,542	2,239

a Abhängige Variable: nmqm49_gWW Nettomiete in €/m².Fertigstellung ab 1949: guter Wohnwert

Tabelle 53: SPSS-Output der Koeffizientenstatistik

Bei diesem Regressionsmodell lässt sich ein starker Einfluss in negativer Richtung der Variablen „Anteil Wohnungen in Zweifamilienhäusern“ erkennen. Ihr standardisierter Koeffizient beträgt -0,334. Auch die Variable „Kaufkraftindex“ hat in diesem Modell wieder einen hohen positiven Einfluss, mit einem standardisierten Regressionskoeffizienten von 0,364. Interessant ist aber vor allem der hohe positive Einfluss der neu aufgenommenen Variablen “Übernachtungsgäste pro Einwohner“ deren standardisierter Koeffizient einen Wert von 0,316 aufweist.

Bei der Betrachtung der Konfidenzintervalle fällt außerdem auf, dass der Wert der Variablen „Übernachtungsgäste pro Einwohner“ relativ genau geschätzt werden kann. Mit 95%-iger Wahrscheinlichkeit liegt der wahre Wert zwischen 0,016 (Intervalluntergrenze) und 0,027 (Intervallobergrenze).

Die Schätzfunktion für die Nettomiete in Euro pro m² (nmqm) unter diesem Modell lautet:

$$nmqm = 0{,}077 - 3{,}843 \cdot ALQ + 0{,}037 \cdot KKI + 9{,}571 \cdot SVB\_EW\_30km + 0{,}0003934 \cdot Dichte - 5{,}399 \cdot Anteil\_Whg\_zfh + 0{,}022 \cdot Gast\_EW + 1{,}390 \cdot gtyp6$$

6.5.4 Überprüfung auf räumliche Korrelation

Die Residuen des Regressionsmodells sollen grafisch daraufhin untersucht werden, ob sich räumliche Muster identifizieren lassen.

Die Auswertung der Grafik (s. Abb. 33) lässt keine räumlichen Muster erkennen, die einen Hinweis auf nicht betrachtete Einflussfaktoren geben würde. Es lässt sich weder eine Clusterung der Residualwerte nach Bundesländern erkennen, noch lässt sich ein Unterschied in der Verteilung zwischen West- und Ostdeutschland bzw. Nord- und Süddeutschland identifizieren. Eine systematische Über- bzw. Unterschätzung der Miethöhen in Großstädten oder Ballungsräumen ist ebenfalls nicht erkennbar. Vielmehr scheinen die Residualwerte unsystematisch und gleichmäßig verteilt zu liegen, so dass eine räumliche Korrelation ausgeschlossen werden kann.

Im Vergleich zu den Mietspiegelmodellen fällt auf, dass die Schätzung der Nettomieten in diesem Modell wesentlich genauer ausfällt. Der Hauptanteil der Residuen liegt in einem Bereich von (-0,10; 0,12].

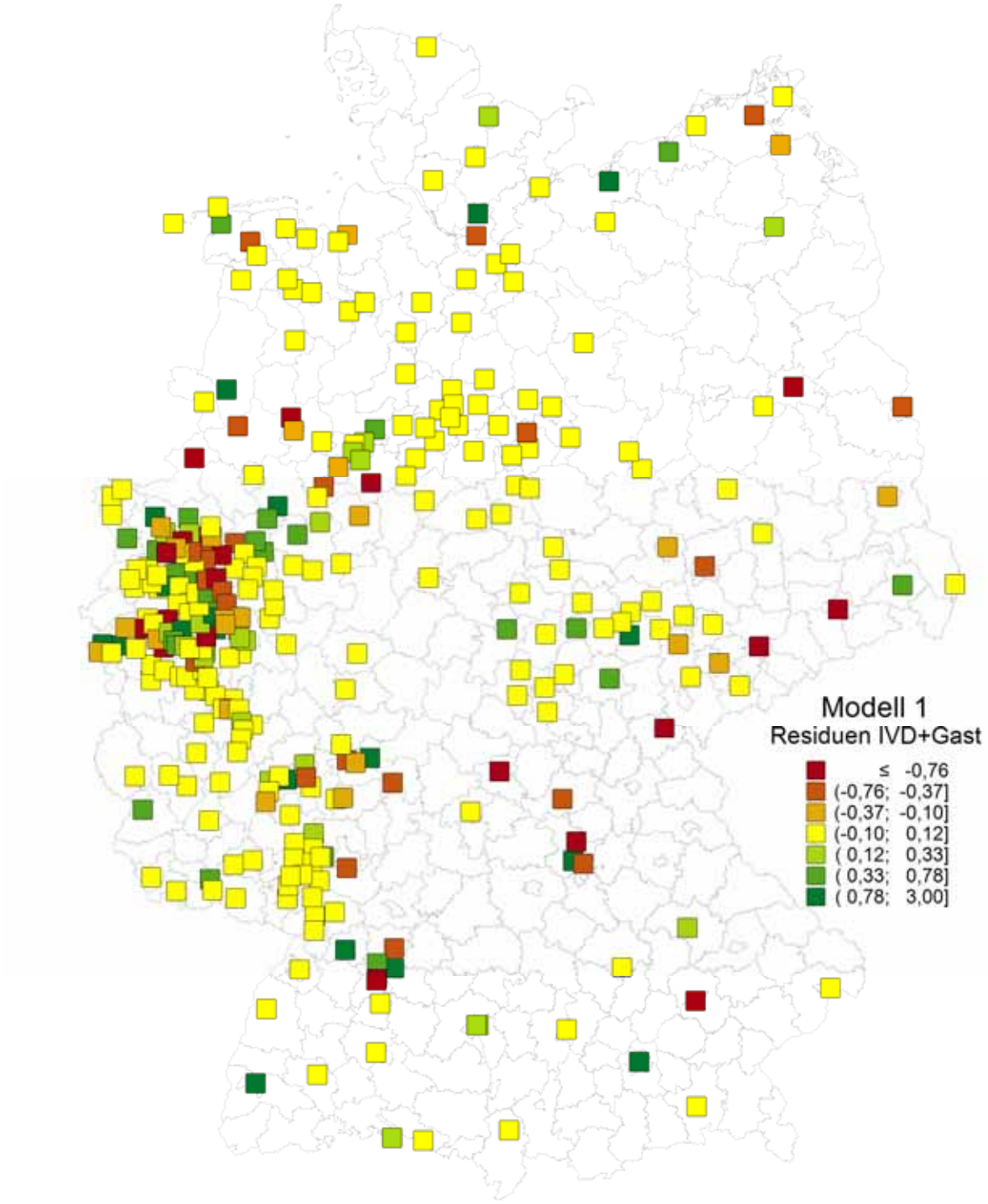


Abbildung 32: Darstellung der Residuen aus dem IVD-Modell 1[102]

[102] Quelle: Eigene Darstellung

6.6 Faktorenanalyse für Modell 2

Für den Aufbau eines zweiten Modells wird versucht, die Variablen die zueinander in enger Beziehung stehen zu identifizieren und nach Möglichkeit zu einem Faktor zusammenzufassen. Dieser Faktor wird dann im Anschluss als unabhängige Variable in eine Regressionsanalyse mit einfließen. Als Extraktionsmethode wird ebenfalls die Hauptkomponentenanalyse eingesetzt[103].

6.6.1 Korrelationsanalyse

Bevor eine Faktorenanalyse durchgeführt werden kann, muss zunächst die Korrelationsmatrix auf ihre Eignung überprüft werden.

[103] Vgl. Kapitel 2

Korrelationsmatrix

	EW Anzahl Einwohner	EW_HH Einwohner pro Haushalt	SVB_pro_Einw Sozialversp flichtig Besch. am Wohnort pro Einwohner	ALQ Arbeitslosenquote	KKI Kaufkraftindex je EW	km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	SVB_EW_30km	Dichte Bevölkerungsdichte	Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	Gast_EW Übernachtungsgäste pro Einwohner im Jahr 2006
EW	1,000	-,323	-,038	,112	,027	,281	-,023	,627	-,319	-,348	-,040
EW_HH	-,323	1,000	,146	-,437	,201	,136	-,185	-,451	,720	,633	-,041
SVB_pro_Einw	-,038	,146	1,000	-,427	,410	,083	,542	-,020	-,019	,057	,241
ALQ	,112	-,437	-,427	1,000	-,734	-,211	-,113	,094	-,420	-,484	-,117
KKI	,027	,201	,410	-,734	1,000	,466	,172	,139	,220	,200	-,007
km30_EwGfK_antlg	,281	,136	,083	-,211	,466	1,000	-,221	,533	-,119	-,071	-,153
SVB_EW_30km	-,023	-,185	,542	-,113	,172	-,221	1,000	,021	-,305	-,133	,032
Dichte	,627	-,451	-,020	,094	,139	,533	,021	1,000	-,571	-,567	-,126
Anteil_Whg_efh	-,319	,720	-,019	-,420	,220	-,119	-,305	-,571	1,000	,598	,049
Anteil_Whg_zfh	-,348	,633	,057	-,484	,200	-,071	-,133	-,567	,598	1,000	-,023
Gast_EW	-,040	-,041	,241	-,117	-,007	-,153	,032	-,126	,049	-,023	1,000

Tabelle 54: SPSS-Output der Korrelationsmatrix

Ähnlich wie auch beim Mietspiegel-Datensatz lässt die Korrelationsmatrix kein eindeutiges Urteil über die Eignung der Daten zur Faktorenanalyse zu. Einige Korrelationen liegen in einem Intervall von 0,3 bis 0,5. Deswegen muss eine Faktorenanalyse aber nicht sinnlos sein, da sie auch Zusammenhänge höheren Grades erfasst[104], die bei paarweisen Korrelationen nicht auffallen.

[104] Vgl. PFEI S. 210

6.6.2 Variablenauswahl

Da sich mittels der Korrelationsmatrix noch keine sinnvolle Aussage zur Eignung für eine Faktorenanalyse treffen lässt sollen nun weitere Kriterien zur Überprüfung herangezogen werden.

6.6.2.1 Signifikanzniveau der Korrelation[105]

Kleine Werte des Signifikanzniveaus gebe die Wahrscheinlichkeit an, dass sich die Korrelation signifikant von Null unterscheidet. Die Analyse der Signifikanzniveaus ergibt, dass sich die meisten Korrelationen signifikant von Null unterscheiden. Des Weiteren lassen sich keine Variablen feststellen, die mit keiner anderen Variablen signifikant korrelieren.

[105] Vgl. Kapitel 2.5.1

Signifikanz (1-seitig)

	EW Anzahl Einwohner	EW_HH Einwohner pro Haushalt	SVB_pro_Einw Sozialversp flichtig Besch. am Wohnort pro Einwohner	ALQ Arbeitslosenquote	KKI Kaufkraftindex je EW	km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	SVB_EW_30km	Dichte Bevölkerungsdichte	Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	Gast_EW Übernachtungsgäste pro Einwohner im Jahr 2006
EW		,000	,247	,023	,316	,000	,339	,000	,000	,000	,236
EW_HH	,000		,004	,000	,000	,007	,000	,000	,000	,000	,232
SVB_pro_Einw	,247	,004		,000	,000	,069	,000	,362	,367	,156	,000
ALQ	,023	,000	,000		,000	,000	,022	,047	,000	,000	,018
KKI	,316	,000	,000	,000		,000	,001	,006	,000	,000	,450
km30_EwGfK_antlg	,000	,007	,069	,000	,000		,000	,000	,016	,103	,003
SVB_EW_30km	,339	,000	,000	,022	,001	,000		,352	,000	,009	,282
Dichte	,000	,000	,362	,047	,006	,000	,352		,000	,000	,012
Anteil_Whg_efh	,000	,000	,367	,000	,000	,016	,000	,000		,000	,192
Anteil_Whg_zfh	,000	,000	,156	,000	,000	,103	,009	,000	,000		,338
Gast_EW	,236	,232	,000	,018	,450	,003	,282	,012	,192	,338	

Tabelle 55: Signifikanzniveau der Korrelation

6.6.2.2 KMO-Maß und Bartlett-Test[106]

Mit Hilfe des KMO-Maß und des Bartlett-Test kann die Gesamtheit der einbezogenen Variablen auf ihre Tauglichkeit geprüft werden.

KMO- und Bartlett-Test

Maß der Stichprobeneignung nach Kaiser-Meyer-Olkin.		,620
Bartlett-Test auf Sphärizität	Ungefähres Chi-Quadrat	1869,965
	df	55
	Signifikanz nach Bartlett	,000

Tabelle 56: KMO-Maß und Bartlett-Test

Für den IVD-Datensatz nimmt das KMO-Maß einen Wert von 0,620 an, womit die Stichprobe als „mäßig“ geeignet angesehen werden kann.

Der Bartlett-Test erbrachte eine Prüfgröße von 1869,965 bei einem Signifikanzniveau von 0,000. Dies bedeutet, dass mit einer Wahrscheinlichkeit von nahezu 100% davon auszugehen ist, dass die Variablen untereinander korrelieren.

[106] Vgl. Kapitel 2.5.2, 2.5.3 und 5.5.2.3

6.6.2.3 Image-Analyse nach Guttman[107]

Anti-Image- Korrelation

	EW Anzahl Einwohner	EW_HH Einwohner pro Haushalt	SVB_pro_Einw Sozialversp flichtig Besch. am Wohnort pro Einwohner	ALQ Arbeitslosenquote	KKI Kaufkraftindex je EW	km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	SVB_EW_30km	Dichte Bevölkerungsdichte	Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	Gast_EW Übernachtungsgäste pro Einwohner im Jahr 2006
EW	,751(a)	,093	-,058	-,053	,052	,022	,019	-,487	-,124	-,072	-,041
EW_HH	,093	,658(a)	-,167	,167	,363	-,457	-,105	,116	-,549	-,250	,138
SVB_pro_Einw	-,058	-,167	,641(a)	,244	-,068	-,094	-,478	,136	,110	,163	-,280
ALQ	-,053	,167	,244	,661(a)	,611	-,237	-,057	,250	,102	,369	,174
KKI	,052	,363	-,068	,611	,526(a)	-,536	-,263	,048	-,306	,058	,123
km30_EwGfK_antlg	,022	-,457	-,094	-,237	-,536	,409(a)	,409	-,446	,303	-,072	,040
SVB_EW_30km	,019	-,105	-,478	-,057	-,263	,409	,463(a)	-,026	,342	,018	,164
Dichte	-,487	,116	,136	,250	,048	-,446	-,026	,680(a)	,259	,388	,124
Anteil_Whg_efh	-,124	-,549	,110	,102	-,306	,303	,342	,259	,701(a)	,020	-,038
Anteil_Whg_zfh	-,072	-,250	,163	,369	,058	-,072	,018	,388	,020	,799(a)	,127
Gast_EW	-,041	,138	-,280	,174	,123	,040	,164	,124	-,038	,127	,366(a)

Tabelle 57: Anti-Image-Korrelationsmatrix

Das auf der Hauptdiagonalen der Anti-Image-Korrelationsmatrix aufgetragene variablenspezifische KMO-Maß (MSA-Wert) weist für die Variablen „Gast_EW“ (Übernachtungsgäste pro Einwohner), „km30_EwGfK_antlg“ (Einwohner in 30km Umkreis) und „SVB_EW_30km“ (so-

[107] Vgl. Kapitel 2.5.4 und 5.5.2.2

zialversicherungspflichtig Beschäftigte je Einwohner in 30km Umkreis) einen Wert < 0,5 auf. Diese müssen daher sukzessive aus dem Modell entfernt werden. Zuerst wird die Variable mit dem kleinsten MSA-Wert (0,366) entfernt. Nachdem die Variable „Gast_EW“ aus dem Modell entfernt wurde, weist die Variable „km30_EwGfK_antlg“ noch immer einen MSA-Wert < 0,5 und muss somit ebenfalls aus dem Modell entfernt werden. Danach ergibt sich für die Anti-Image-Korrelationsmatrix folgendes Bild: Alle Variablen weisen nun einen MSA-Wert größer 0,5 auf und können somit für die Faktorenanalyse verwendet werden.

Anti-Image-Korrelation

	EW Anzahl Einwohner zum 31.12.2006	EW_HH Einwohner pro Haushalt	SVB_pro_Einw Sozialverspflichtig Besch. am Wohnort pro Einwohner	ALQ Arbeitslosenquote	KKI Kaufkraftindex je EW	SVB_EW_30km Soz.vers.pflichtig Besch. je EW in 30km Umkreis	Dichte Bevölkerungsdichte	Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern
EW	,694(a)	,116	-,056	-,049	,076	,011	-,533	-,137	-,071
EW_HH	,116	,768(a)	-,237	,069	,157	,101	-,110	-,484	-,319
SVB_pro_Einw	-,056	-,237	,619(a)	,230	-,140	-,484	,106	,146	,157
ALQ	-,049	,069	,230	,696(a)	,590	,044	,166	,188	,363
KKI	,076	,157	-,140	,590	,634(a)	-,057	-,252	-,178	,023
SVB_EW_30km	,011	,101	-,484	,044	-,057	,571(a)	,191	,251	,052
Dichte	-,533	-,110	,106	,166	-,252	,191	,619(a)	,462	,399
Anteil_Whg_efh	-,137	-,484	,146	,188	-,178	,251	,462	,726(a)	,044
Anteil_Whg_zfh	-,071	-,319	,157	,363	,023	,052	,399	,044	,783(a)

a Maß der Stichprobeneignung

Tabelle 58: Anti-Image-Korrelationsmatrix

Nach Dziuban und Shirkey[108] ist eine Korrelationsmatrix dann als ungeeignet zu betrachten, wenn der Anteil der Nicht-Diagonal-Elemente der Anti-Image-Kovarianzmatrix, die einen kritischen Wert von 0,09 überschreiten, über 25% liegt.

[108] Vgl. BACK S. 275f.

Dieses Kriterium wird von den in die Untersuchung einfließenden Variablen, wie schon beim Datensatz zuvor, nicht erfüllt. Hier liegt der Anteil der Nicht-Diagonal-Element mit einem Wert > 0,09 bei ca. 30%. Trotzdem soll die Korrelationsmatrix für eine Faktorenanalyse verwendet werden, da bei den anderen Prüfgrößen die Kriterien erfüllt werden.

Anti-Image- Kovarianz

	EW Anzahl Einwohner zum 31.12.2006	EW_HH Einwohner pro Haushalt	SVB_pro_Einw Sozialverspflichtig Besch. am Wohnort pro Einwohner	ALQ Arbeitslosenquote	KKI Kaufkraftindex je EW	SVB_EW_30km Soz.vers.pflichtig Besch. je EW in 30km Umkreis	Dichte Bevölkerungsdichte	Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern
EW	,585	,055	-,031	-,021	,036	,006	-,229	-,059	-,034
EW_HH	,055	,384	-,106	,023	,061	,048	-,038	-,167	-,124
SVB_pro_Einw	-,031	-,106	,521	,091	-,064	-,267	,043	,059	,071
ALQ	-,021	,023	,091	,299	,203	,018	,051	,057	,124
KKI	,036	,061	-,064	,203	,397	-,027	-,089	-,063	,009
SVB_EW_30km	,006	,048	-,267	,018	-,027	,582	,082	,107	,025
Dichte	-,229	-,038	,043	,051	-,089	,082	,315	,145	,140
Anteil_Whg_efh	-,059	-,167	,059	,057	-,063	,107	,145	,311	,015
Anteil_Whg_zfh	-,034	-,124	,071	,124	,009	,025	,140	,015	,391

Tabelle 59: Anti-Image-Kovarianzmatrix

6.6.3 Bestimmung der Anzahl der zu extrahierenden Faktoren

Die Kommunalitäten beschreiben den Teil der Gesamtvarianz einer Variablen, der durch die extrahierten Faktoren erklärt werden kann.

Kommunalitäten

	Anfänglich	Extraktion
EW Anzahl Einwohner zum 31.12.2006	1,000	,660
EW_HH Einwohner pro Haushalt	1,000	,701
SVB_pro_Einw Sozialverspflichtig Besch. am Wohnort pro Einwohner	1,000	,748
ALQ Arbeitslosenquote	1,000	,849
KKI Kaufkraftindex je EW	1,000	,781
SVB_EW_30km Soz.vers.pflichtig Beschäftigte je Einwohner in 30km Umkreis	1,000	,816
Dichte Bevölkerungsdichte	1,000	,846
Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	1,000	,785
Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	1,000	,703

Extraktionsmethode: Hauptkomponentenanalyse.

Tabelle 60: SPSS-Output der anfänglichen und endgültigen Kommunalitäten

Der obigen Tabelle kann entnommen werden, dass durch alle extrahierten Faktoren z.B. 84,9% der Varianz der Variablen „Arbeitslosenquote“ erklärt werden können. Am geringsten liegt der Anteil an erklärter Varianz wiederum bei der Variablen „Einwohner“, bei der nur 66% durch alle extrahierten Faktoren erklärt werden können (im Datensatz „Mietspiegel.sav“ waren es sogar nur 53,8%).

Erklärte Gesamtvarianz

Komponente	Anfängliche Eigenwerte		
	Gesamt	% der Varianz	Kumulierte %
1	3,481	38,674	38,674
2	2,122	23,579	62,253
3	1,288	14,310	76,564
4	,626	6,954	83,517
5	,453	5,029	88,546
6	,361	4,014	92,561
7	,314	3,493	96,053
8	,202	2,242	98,295
9	,153	1,705	100,000

Extraktionsmethode: Hauptkomponentenanalyse.

Tabelle 61: SPSS-Ausgabe der erklärten Gesamtvarianz

Von den anfänglich neun Faktoren weisen ebenfalls nur die ersten drei einen Eigenwert größer Eins auf. Somit sind nach dem Kaiser-Kriterium auch hier drei Faktoren zu extrahieren.

Der Erklärungsgehalt dieser Faktoren ist etwas geringer als es beim vorherigen Datensatz der Fall war. Die ersten drei Komponenten erklären bei dieser Lösung einen Varianzanteil von 76,564%.

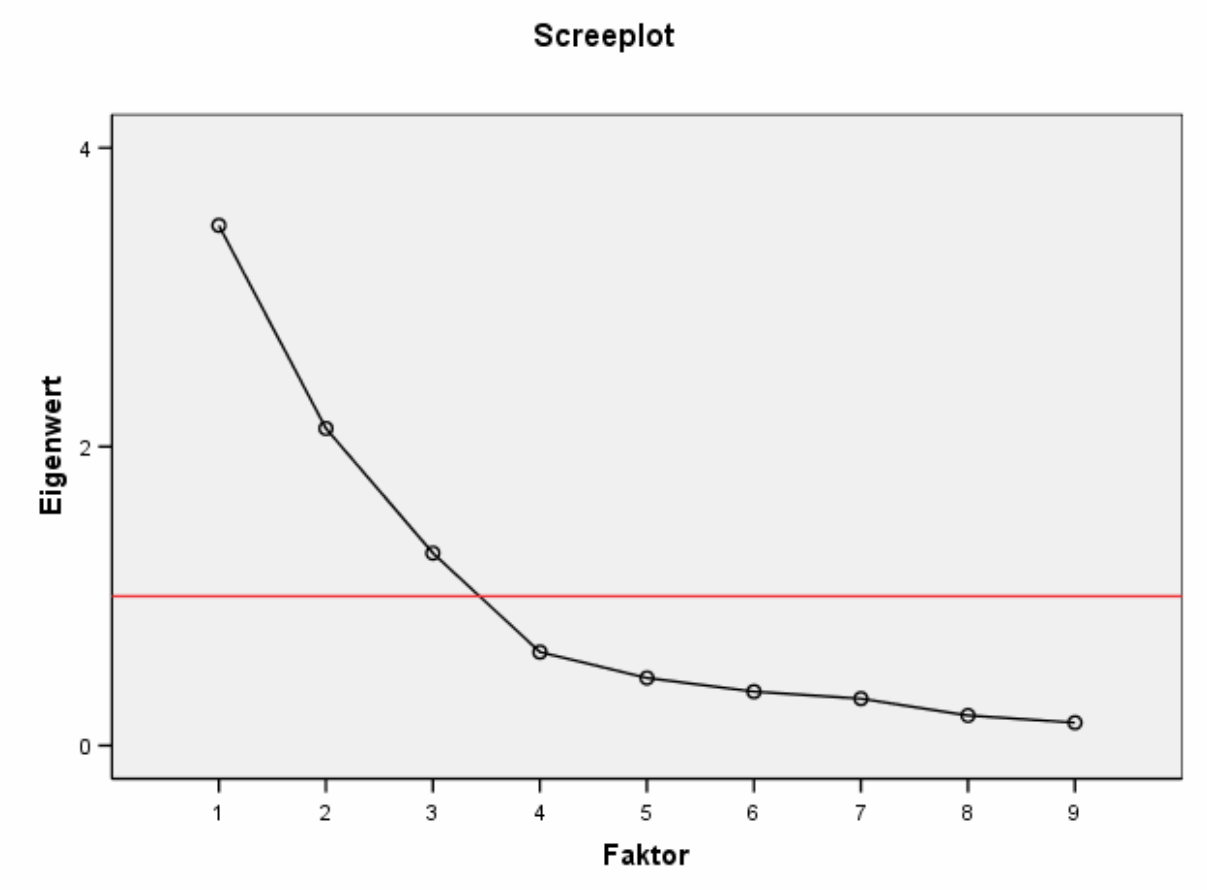


Abbildung 33: Scree-Plot mit eingezeichnetem Kaiser-Kriterium

Der Scree-Plot zeigt beim vierten Eigenwert einen Knick, d.h. auch hier wird eine 3-faktorielle Lösung gewählt, was mit dem Kaiser-Kriterium übereinstimmt.

6.6.4 Auswertung der Ergebnisse

6.6.4.1 Reproduzierte Korrelationsmatrix und Residualmatrix

Reproduzierte Korrelationen

Reproduzierte Korrelationen

		EW Anzahl Einwohner zum 31.12.2006	EW_HH Einwohner pro Haushalt	SVB_pro_Einw Sozialversicherungspflichtig Besch. am Wohnort pro Einwohner	ALQ Arbeitslosenquote	KKI Kaufkraftindex je EW	km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	Dichte Bevölkerungsdichte	Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern
Reproduzierte Korrelation	EW	,660(b)	-,408	-,122	,062	,166	-,140	,730	-,397	-,452
	EW_HH	-,408	,701(b)	,073	-,516	,272	-,216	-,552	,735	,700
	SVB_pro_Einw	-,122	,073	,748(b)	-,502	,524	,678	-,025	-,026	,095
	ALQ	,062	-,516	-,502	,849(b)	-,761	-,134	,102	-,493	-,499
	KKI	,166	,272	,524	-,761	,781(b)	,211	,183	,235	,248
	km30_EwGfK_antlg	-,140	-,216	,678	-,134	,211	,816(b)	,012	-,332	-,175
	Dichte	,730	-,552	-,025	,102	,183	,012	,846(b)	-,563	-,592
	Anteil_Whg_efh	-,397	,735	-,026	-,493	,235	-,332	-,563	,785(b)	,728
	Anteil_Whg_zfh	-,452	,700	,095	-,499	,248	-,175	-,592	,728	,703(b)

		EW Anzahl Einwohner zum 31.12.2006	EW_HH Einwohner pro Haushalt	SVB_pro_Einw Sozialversp flichtig Besch. am Wohnort pro Einwohner	ALQ Arbeitslosenquote	KKI Kaufkraftindex je EW	km30_EwGfK_antlg Einwohnerzahl in 30km Umkreis	Dichte Bevölkerungsdichte	Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern
Residuum(a)	EW		,082	,079	,064	-,149	,113	-,102	,068	,096
	EW_HH	,082		,079	,078	-,106	,046	,085	-,019	-,060
	SVB_pro_Einw	,079	,079		,070	-,110	-,130	,006	,028	-,030
	ALQ	,064	,078	,070		,038	,009	,008	,050	,004
	KKI	-,149	-,106	-,110	,038		-,030	-,043	-,017	-,052
	km30_EwGfK_antlg	,113	,046	-,130	,009	-,030		,008	,058	,054
	Dichte	-,102	,085	,006	,008	-,043	,008		-,014	,016
	Anteil_Whg_efh	,068	-,019	,028	,050	-,017	,058	-,014		-,119
	Anteil_Whg_zfh	,096	-,060	-,030	,004	-,052	,054	,016	-,119	

Extraktionsmethode: Hauptkomponentenanalyse.

a Residuen werden zwischen beobachteten und reproduzierten Korrelationen berechnet. Es liegen 21 (58,0%) nicht redundante Residuen mit absoluten Werten größer 0,05 vor.

b Reproduzierte Kommunalitäten

Tabelle 62: Reproduzierte Korrelationsmatrix und Residualmatrix

Der reproduzierte Korrelationskoeffizient zwischen der Variablen „Arbeitslosenquote“ und der Variablen „Anteil Wohnungen in Zweifamilienhäusern“ weicht um 0,004, also minimal, vom ursprünglichen Korrelationskoeffizienten ab. Allerdings weisen 58% der Residuen einen Wert > 0,05 auf, was darauf schließen lässt, dass das Modell die ursprünglichen Daten nicht optimal darstellt.

6.6.4.2 Interpretation der gefundenen Faktoren

Um die Interpretation der Faktoren zu erleichtern, wird die Komponentenmatrix mittels des Varimax-Verfahrens orthogonal rotiert. Damit ist auch gewährleistet, dass die Faktoren für die anschließende Regressionsanalyse voneinander unabhängig sind.

Rotierte Komponentenmatrix(a)

	Komponente		
	1	2	3
Dichte Bevölkerungsdichte	-,911		
EW Anzahl Einwohner zum 31.12.2006	-,770		
Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	,712		
Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	,694		
EW_HH Einwohner pro Haushalt	,676		
ALQ Arbeitslosenquote		-,884	
KKI Kaufkraftindex je EW		,852	
SVB_EW_30km Soz.vers.pflichtig Beschäftigte je Einwohner in 30km Umkreis			,901
SVB_pro_Einw Sozialverspflichtig Besch. am Wohnort pro Einwohner			,746

Extraktionsmethode: Hauptkomponentenanalyse.
Rotationsmethode: Varimax mit Kaiser-Normalisierung.
a Die Rotation ist in 12 Iterationen konvergiert.

Tabelle 63: Rotierte Komponentenmatrix

Anhand der rotierten Komponentenmatrix kann nun die inhaltliche Interpretation der Faktoren vorgenommen werden. In der Praxis hat sich die Konvention durchgesetzt, dass eine Variable ab einer Ladungshöhe von |0,5| einem Faktor zugeordnet werden kann. Zur besseren Übersicht wurden deshalb die Faktorladungen der Größe nach sortiert und solche mit einem betragsmäßigen Wert von < 0,5 nicht angezeigt[109].

Aus Tabelle 63 lässt sich ablesen, dass auf den ersten Faktor die Variablen „Einwohner pro Haushalt“, „Anteil Wohnungen in Einfamilienhäusern“ und „Anteil Wohnungen in Zweifamilienhäusern“ stark positiv laden, während die Variablen „Bevölkerungsdichte“ und „Anzahl der Einwohner“ einen starken negativen Einfluss haben. Dieser Faktor nimmt also in Regionen mit

[109] SPSS-Befehl: /FORMAT SORT BLANK(.50)

- vielen Personen pro Haushalt (Familien)
- einem hohen Anteil an Ein- und Zweifamilienhäusern
- einer geringen Bevölkerungsdichte und
- einer geringen Anzahl Einwohner

tendenziell hohe Werte an. Diese Konstellationen sind in Deutschland vor allem in den ländliche Regionen zu finden.

Auf den zweiten Faktor lädt die Variable „Kaufkraftindex“ stark positiv während die „Arbeitslosenquote“ einen starken negativen Einfluss hat. Das bedeutet, dass der zweite Faktor tendenziell hohe Werte in Regionen mit

- einer hohen Kaufkraft und
- einer geringen Arbeitslosenquote

annimmt. Dieser Faktor kann daher als ein Indikator für „Wirtschaftskraft“ angesehen werden.

Der dritte der extrahierten Faktoren weist hohe positive Ladungen bei den Variablen „Sozialversicherungspflichtig Beschäftigte am Wohnort pro Einwohner“ und „Sozialversicherungspflichtig Beschäftigte je Einwohner in 30km Umkreis“ auf. Der dritte Faktor weist also hohe Werte in Regionen mit

- einem hohen Anteil sozialversicherungspflichtig Beschäftigter am Wohnort und
- einem hohen Anteil sozialversicherungspflichtig Beschäftigter im Umkreis

auf. Dieser Faktor kann ebenfalls als „Wirtschaftskraft“ interpretiert werden kann.

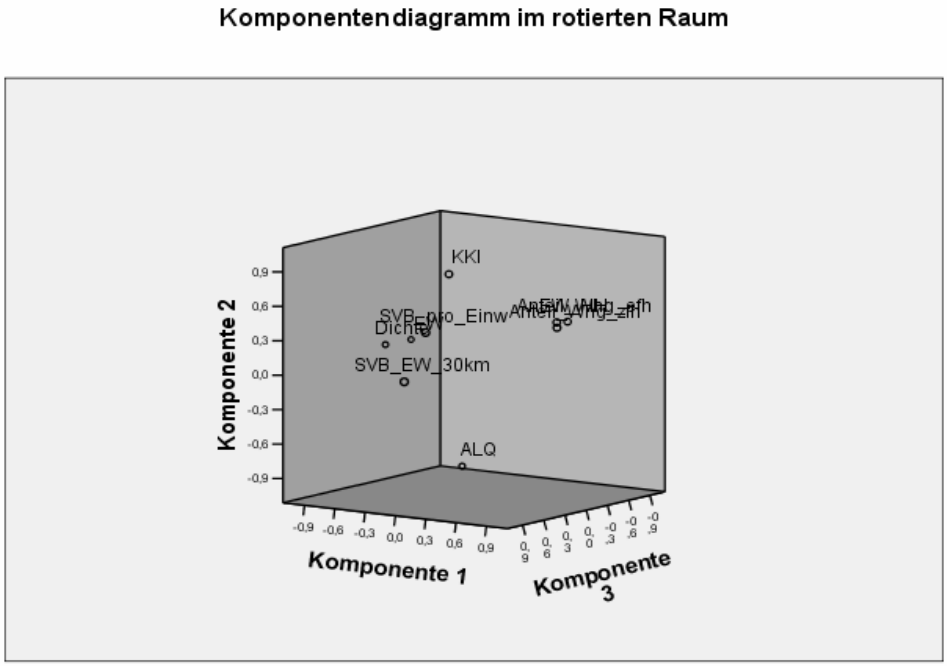


Abbildung 34: Komponentendiagramm im rotierten Raum

Anhand von Abbildung 34 kann entsprechend grafisch veranschaulicht werden, wie die Variablen im dreidimensionalen Faktorraum liegen.

6.6.4.3 Bestimmung der Faktorwerte[110]

Mit Hilfe der Koeffizientenmatrix der Komponentenwerte können die Faktorwerte für die einzelnen Städte und Gemeinden berechnet werden. SPSS gibt die berechneten Werte im Dateneditor aus.

Koeffizientenmatrix der Komponentenwerte

	Komponente		
	1	2	3
EW Anzahl Einwohner zum 31.12.2006	-,344	,217	-,225
EW_HH Einwohner pro Haushalt	,178	,150	-,121
SVB_pro_Einw Sozialverspflichtig Besch. am Wohnort pro Einwohner	,023	,124	,431
ALQ Arbeitslosenquote	,032	-,394	-,009
KKI Kaufkraftindex je EW	-,147	,418	,024
SVB_EW_30km Soz.vers.pflichtig Beschäftigte je Einwohner in 30km Umkreis	,059	-,089	,571
Dichte Bevölkerungsdichte	-,379	,201	-,132
Anteil_Whg_efh Anteil Wohnungen in Einfamilienhäusern	,174	,157	-,200
Anteil_Whg_zfh Anteil Wohnungen in Zweifamilienhäusern	,202	,120	-,082

Extraktionsmethode: Hauptkomponentenanalyse.
Rotationsmethode: Varimax mit Kaiser-Normalisierung.
Komponentenwerte.

Tabelle 64: Koeffizientenmatrix der Komponentenwerte

Für den ersten Faktor ergibt sich somit die folgende Gleichung:

$$FACTOR\_1 = -0{,}344 \cdot EW + 0{,}178 \cdot EW\_HH + 0{,}023 \cdot SVB\_pro\_EW + 0{,}032 \cdot ALQ - 0{,}147 \cdot KKI + 0{,}059 \cdot SVB\_EW\_30km - 0{,}379 \cdot Dichte + 0{,}174 \cdot Anteil\_Whg\_efh + 0{,}202 \cdot Anteil\_Whg\_zfh$$

Für die beiden anderen Faktoren werden die Gleichungen auf dieselbe Weise erstellt. Aus den ursprünglich neun Variablen sind nun drei Faktoren extrahiert worden, die 76,564% der Information enthalten.

[110] Vgl. Kapitel 2.9

6.6.5 Überprüfung der Merkmale anhand einer 50%-Zufallsstichprobe

Nachdem die Faktoren extrahiert und sinnvoll interpretiert wurden, soll nun anhand einer 50%-Zufallsstichprobe die Qualität der Untersuchung überprüft werden. Dazu wird die Stichprobe zufällig aufgeteilt und das Ergebnis anhand der beiden Teilstichproben kontrolliert. Wünschenswert hierbei ist es, dass sich ein ähnliches Bild wie bei der Gesamtstichprobe ergibt, da die Ergebnisse sonst als zufallsbedingt zu bewerten sind.

6.6.5.1 Variablenauswahl

Das KMO-Maß für die beiden Teilstichproben mit einem Umfang von 182 und 163 Fällen fällt auch hier erstaunlich gut aus mit Werten von 0,681 bzw. 0,705, was als mäßig bzw. als mittelprächtig eingestuft werden kann.

KMO- und Bartlett-Test

Maß der Stichprobeneignung nach Kaiser-Meyer-Olkin.		,681
Bartlett-Test auf Sphärizität	Ungefähres Chi-Quadrat	873,303
	df	36
	Signifikanz nach Bartlett	,000

Tabelle 65: KMO-Maß und Bartlett-Test der ersten 50%-Zufallsstichprobe

KMO- und Bartlett-Test

Maß der Stichprobeneignung nach Kaiser-Meyer-Olkin.		,705
Bartlett-Test auf Sphärizität	Ungefähres Chi-Quadrat	761,555
	df	36
	Signifikanz nach Bartlett	,000

Tabelle 66: KMO-Maß und Bartlett-Test der zweiten 50%-Zufallsstichprobe

Auch der Bartlett-Test fällt bei beiden Stichproben signifikant aus.

Bei den Anti-Image-Korrelationsmatrizen ändern sich bei beiden Stichproben die variablenspezifischen MSA-Maße geringfügig; alle bleiben aber größer 0,5, so dass keine weitere Variable ausgeschlossen werden musste. Die Anti-Image-Kovarianzmatrizen erfüllen das Kriterium dagegen bei beiden Zufallsstichproben ebenfalls nicht.

6.6.5.2 Bestimmung der Anzahl der zu extrahierenden Faktoren

Aus beiden untersuchten Teilstichproben resultiert nach dem Kaiser-Kriterium eine Extraktion von drei Faktoren, die jeweils 77,010% bzw. 77,563% der Gesamtvarianz erklären. Es ergibt sich somit ein nahezu identisches Bild zu der Gesamtstichprobe.

6.6.5.3 Vergleich der extrahierten Faktoren

Die drei aus der Gesamtstichprobe extrahierten Faktoren werden in den beiden 50%-Zufallsstichproben fast identisch abgebildet. Die auf die Faktoren hochladenden Variablen variieren auch in ihrer Höhe kaum. Dies deutet darauf hin, relativ gute Faktoren gefunden zu haben.

6.7 Regressionsanalyse für Modell 2

Die eben erhobenen Faktoren sollen nun mit den restlichen Variablen in einer Regressionsanalyse auf Art und Größe ihres Einflusses untersucht werden. Es gelten dieselben Annahmen wie für die vorherigen Regressionsanalysen. Auf eine ausführliche Beschreibung wird deshalb verzichtet und auf das Kapitel 5.4 verwiesen.

6.7.1 Überprüfung der Voraussetzungen[111]

6.7.1.1 Multikollinearität

	Kollinearitätsstatistik	
	Toleranz	VIF
(Konstante)		
FAC1_1 REGR factor score 1 for analysis 1	,609	1,643
FAC2_1 REGR factor score 2 for analysis 1	,962	1,039
FAC3_1 REGR factor score 3 for analysis 1	,968	1,033
Gast_EW Übernachtungsgäste pro Einwohner im Jahr 2006	,971	1,030
gtyp1 Kern- und Großstädte >= 500.000 E. in Agglomerationsräumen	,602	1,661
gtyp6 sonstige Gemeinden in verd. Kreisen von Agglomerationsräumen	,976	1,025

Tabelle 67: Kollinearitätsstatistik

In der hier betrachteten Regression liegt der kleinste Toleranzwert bei 0,602, der höchste VIF nimmt einen Wert von 1,661 an. Damit kann Multikollinearität ausgeschlossen werden.

[111] Vgl. Kapitel 3.6

6.7.1.2 Normalverteilung der Residuen

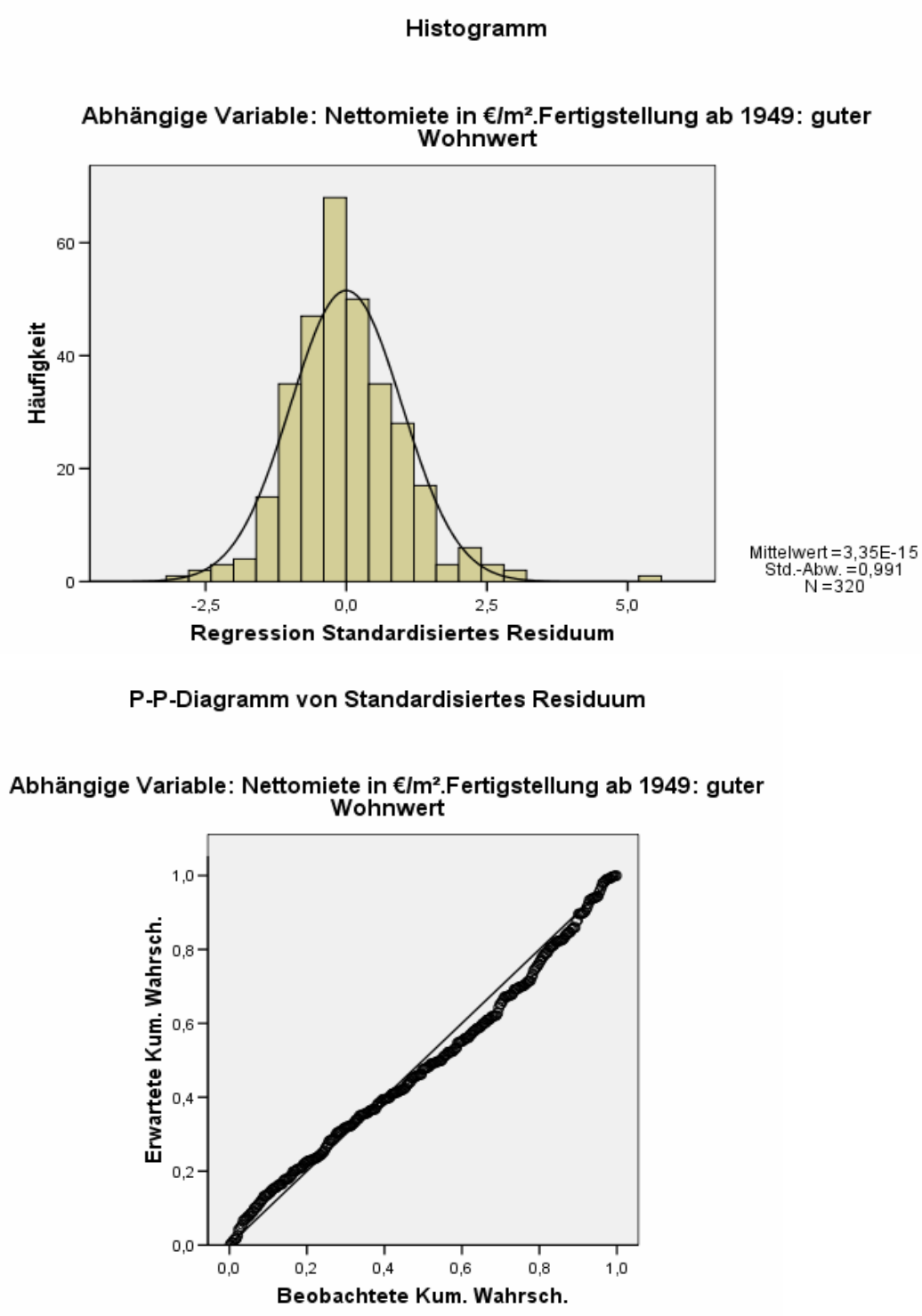


Abbildung 35: Histogramm der standardisierten Residuen

Das Histogramm zeigt, dass sich die standardisierten Residuen recht gut an eine Standardnormalverteilung anpassen. In der Spitze und an den Rändern sind kleinere Abweichungen erkennbar. Geringfügige Abweichungen sind in der Praxis aber nicht ungewöhnlich und durchaus zu tolerieren. Der Erwartungswert liegt nahe 0; die Standardabweichung bei 0,991.

Das PP-Diagramm zeigt, dass die standardisierten Residuen der abhängigen Variablen mit ganz leichten Abweichungen der Referenzverteilung (Normalverteilung) folgen.

Das Vorliegen einer Normalverteilung der standardisierten Residuen als Voraussetzung einer linearen Regressionsanalyse kann somit als erfüllt betrachtet werden.

6.7.1.3 Heteroskedastizität

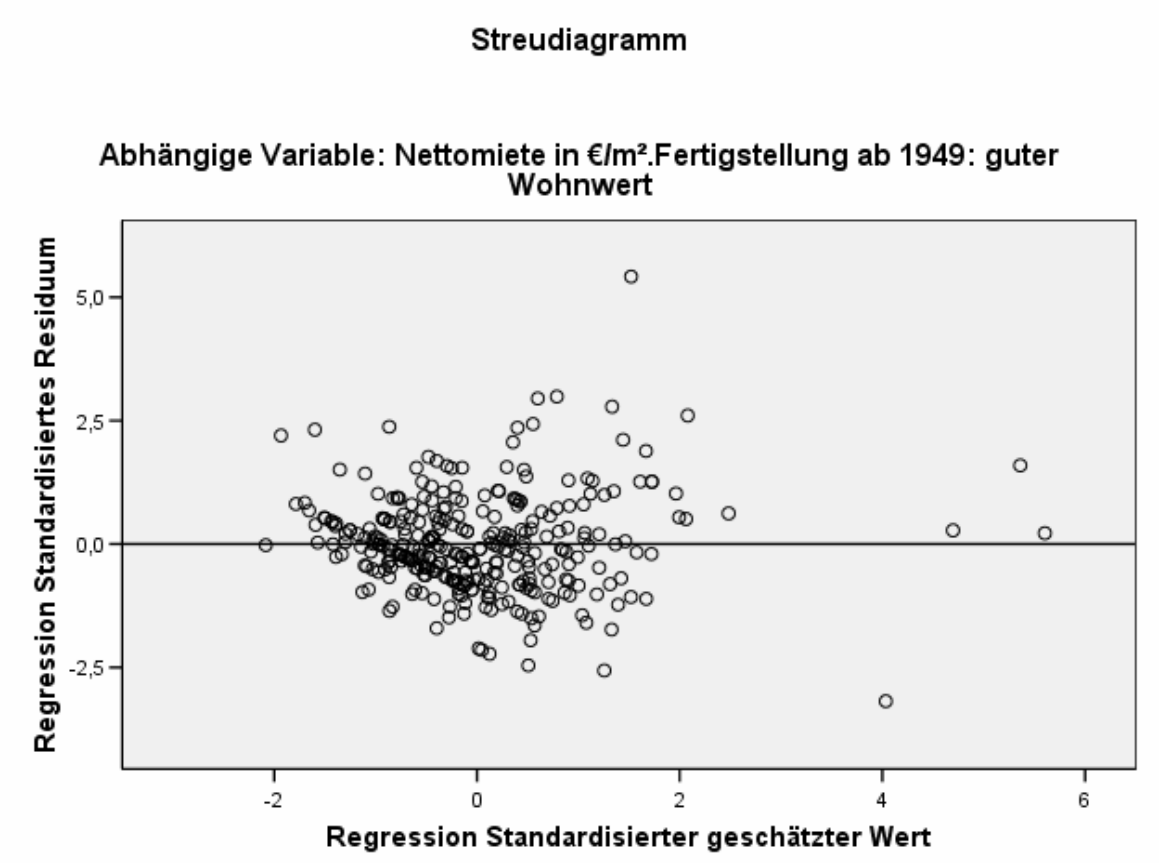


Abbildung 36: Streudiagramm für den Test auf Varianzgleichheit

Die Werte streuen ausgewogen und zufällig um die Nulllinie, d.h. es ist keine Systematik in der Streuung erkennbar. Es liegen somit keine Hinweise auf Heteroskedastizität vor. Auch bei der Auswertung der Streudiagramme der einzelnen Variablen konnten keine Auffälligkeiten festgestellt werden.

6.7.1.4 Autokorrelation

Der Index-Wert zur Prüfung auf Autokorrelation liegt in diesem Modellansatz bei 1,868 (s. Tabelle 68), was auf keine Autokorrelation hindeutet.

6.7.2 Überprüfung der Modellgüte[112]

.Das Bestimmtheitsmaß R^2 des eben ermittelten Regressionsmodells liegt bei 0,503, d.h. durch das Modell kann 50,3% der (senkrechten) Streuung in den y-Werten erklärt werden.

Modellzusammenfassung(n)

R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Durbin-Watson-Statistik
,709(m)	,503	,494	,88546	1,868

m Einflußvariablen : (Konstante), gtyp6 sonstige Gemeinden in verd. Kreisen von Agglomerationsräumen, gtyp1 Kern- und Großstädte >= 500.000 E. in Agglomerationsräumen, FAC3_1 REGR factor score 3 for analysis 1, Gast_EW Übernachtungsgäste pro Einwohner im Jahr 2006, FAC2_1 REGR factor score 2 for analysis 1, FAC1_1 REGR factor score 1 for analysis 1
n Abhängige Variable: nmqm49_gWW Nettomiete in €/m².Fertigstellung ab 1949: guter Wohnwert

Tabelle 68: SPSS-Output der Modellbeurteilung

Das korrigierte R-Quadrat $\overline{R}^2$ dient zur Beurteilung des Modells unter Berücksichtigung der Anzahl der Regressoren. Der Wert dieser Größe beträgt für dieses Regressionsmodell 49,9%. Der Standardfehler des Schätzers fällt im Vergleich zum ersten Regressionsmodell (0,84466) höher aus und nimmt in diesem Regressionsmodell einen Wert von 0,88546 an. Bezogen auf den Mittelwert der abhängigen Variablen beträgt er damit 15,1%.

ANOVA(n)

	Quadrat-summe	df	Mittel der Quadrate	F	Signifikanz
Regression	248,456	6	41,409	52,815	,000(m)
Residuen	245,404	313	,784		
Gesamt	493,860	319			

m Einflussvariablen : (Konstante), gtyp6 sonstige Gemeinden in verd. Kreisen von Agglomerationsräumen, gtyp1 Kern- und Großstädte >= 500.000 E. in Agglomerationsräumen, FAC3_1 REGR factor score 3 for analysis 1, Gast_EW Übernachtungsgäste pro Einwohner im Jahr 2006, FAC2_1 REGR factor score 2 for analysis 1, FAC1_1 REGR factor score 1 for analysis 1
n Abhängige Variable: nmqm49_gWW Nettomiete in €/m².Fertigstellung ab 1949: guter Wohnwert

Tabelle 69: SPSS-Ausgabe der Varianzanalyse

Die Quadratsumme der Residuen (SSE) beträgt in dem Modellansatz 245,404 die Quadratsumme der erklärten Abweichungen (SSR) nimmt einen Wert von 248,456 an. Die geschätzte Restvarianz beträgt im zweiten Modell 0,784. Die Nullhypothese des F-Tests (in der Grundgesamtheit besteht kein linearer Zusammenhang zwischen Regressand und Regressoren) kann zu dem im Voraus festgelegten Signifikanzniveau von 1% verworfen werden.

[112] Vgl. Kapitel 3.4

6.7.3 Überprüfung der Koeffizienten[113]

Koeffizienten(a)

	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B	
	B	Standardfehler	Beta			Untergrenze	Obergrenze
(Konstante)	5,756	,053		108,225	,000	5,651	5,861
FAC1_1 REGR factor score 1 for analysis 1	-,653	,063	-,527	-10,315	,000	-,777	-,528
FAC2_1 REGR factor score 2 for analysis 1	,428	,050	,347	8,534	,000	,330	,527
FAC3_1 REGR factor score 3 for analysis 1	,307	,050	,251	6,192	,000	,210	,405
Gast_EW Übernachtungsgäste pro Einwohner im Jahr 2006	,021	,003	,300	7,427	,000	,015	,026
gtyp1 Kern- und Großstädte >= 500.000 E. in Agglomerationsräumen	-,708	,312	-,117	-2,270	,024	-1,322	-,094
gtyp6 sonstige Gemeinden in verd. Kreisen von Agglomerationsräumen	1,361	,451	,122	3,018	,003	,474	2,249

a Abhängige Variable: nmqm49_gWW Nettomiete in €/m².Fertigstellung ab 1949: guter Wohnwert

Tabelle 70: SPSS-Output der Koeffizientenstatistik

Bei diesem Regressionsmodell lässt sich ein starker Einfluss in negativer Richtung durch den ersten Faktor erkennen. Der standardisierte Koeffizient beträgt -0,527. Auch die neu aufgenommene Variable “Übernachtungsgäste pro Einwohner“ hat in diesem Modell wieder einen hohen positiven Einfluss, mit einem standardisierten Regressionskoeffizienten von 0,300. Bei der Betrachtung der Konfidenzintervalle fällt außerdem auf, dass der Wert der Variablen „Übernachtungsgäste pro Einwohner“ relativ genau geschätzt werden kann. Mit 95%-iger Wahrscheinlichkeit liegt der wahre Wert zwischen 0,015 (Intervalluntergrenze) und 0,027 (Intervallobergrenze).

Die Schätzfunktion für die Nettomiete in Euro pro m² (nmqm) unter diesem Modell lautet:

$$nmqm = 5{,}756 - 0{,}653 \cdot FACTOR\_1 + 0{,}428 \cdot FACTOR\_2 + 0{,}307 \cdot FACTOR\_3 + 0{,}021 \cdot Gast\_EW - 0{,}708 \cdot gtyp1 + 1{,}361 \cdot gtyp6$$

[113] Vgl. Kapitel 3.5

6.7.4 Überprüfung auf räumliche Korrelation

Abschließend sollen auch für das zweite Regressionsmodell die Residuen grafisch daraufhin untersucht werden, ob sich räumliche Muster erkennen lassen.

Bei den Residuen lassen sich eindeutig regionale Muster identifizieren. Besonders in Ostdeutschland werden die Nettomieten verstärkt überschätzt, die prosperierenden Ballungsräume werden dagegen eher unterschätzt. Im Ruhrgebiet lässt sich eine Unterschätzung des „Ballungszentrum“ erkennen, die Randgebiete werden dagegen überschätzt. Eine unsystematische Verteilung der Residualwerte ist hier nicht zu erkennen, vielmehr entsteht eine starke Clusterung nach großen Regionen, was darauf hindeutet, dass das Modell die Ursprungsdaten nicht optimal darstellt.

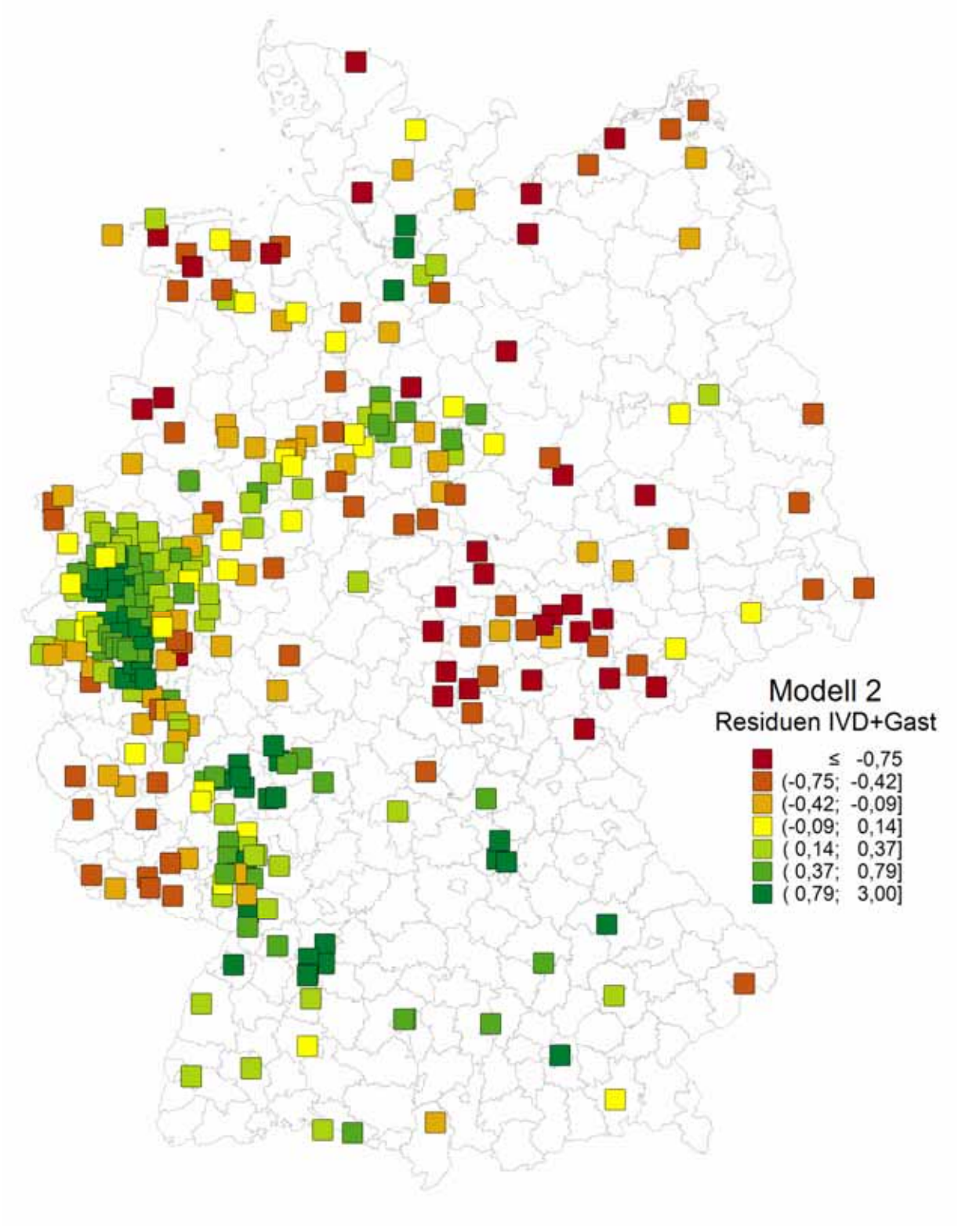


Abbildung 37: Darstellung der Residuen aus dem IVD-Modell 2[114]

[114] Quelle: Eigene Darstellung

6.8 Überprüfung der Modellstrukturen anhand einer 50%-Zufallsstichprobe

Um zu überprüfen, wie zufällig die gefundene Modellstruktur ist, soll Modell 1 (aufgrund der besseren Modellgüte) anhand einer 50%-Zufallsstichprobe berechnet werden. Die Teilstichprobenbildung erfolgt wie bei der Überprüfung der Faktorenanalyse im Mietspiegel-Datensatz, weshalb auf Kapitel 5.5.5 verwiesen wird.

6.8.1 Vergleich der Voraussetzungen

Das Vorhandensein von Autokorrelation kann für die beiden Teilstichproben, mit einem Umfang von 172 und 148 gültigen Fällen, ausgeschlossen werden. Die jeweiligen Durbin-Watson-Statistiken liegen bei 1,803 und 2,082.

Bei Überprüfung der Normalverteilung der Residuen fällt auf, dass sich die Erwartungswerte und Standardabweichungen der 50%-Zufallsstichproben nur unwesentlich von denen der Gesamtstichprobe unterscheiden. Die grafische Analyse zeigt zudem, dass sich beide Stichproben recht gut an eine Normalverteilung anpassen, wie nachfolgend zu sehen ist:

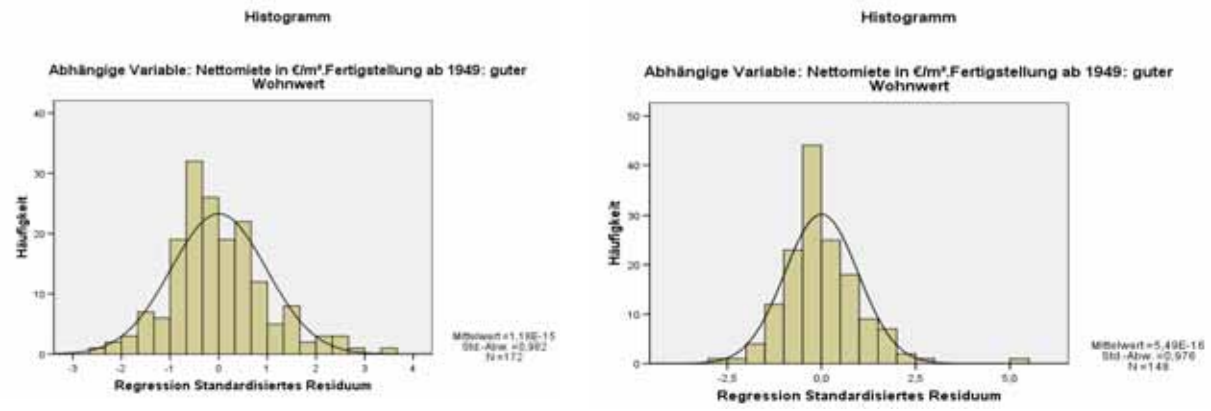


Abbildung 38: Histogramm der standardisierten Residuen (1. und 2. Teilstichprobe)

6.8.2 Vergleich der Modellgüte

Die Zufallsstichprobe mit 172 Fällen verbessert sich bei der Regressionsanalyse um 9,1% beim R^2 (64%) und um knapp 9% beim korrigierten R-Quadrat $\overline{R}^2$ (62,7%). Die zweite Teilstichprobe mit 168 Fällen verschlechtert sich beim R^2 um 4% auf 50,7%, das $\overline{R}^2$ verschlechtert sich um knapp 6% auf 48,2%. Die Fehlerquadratsummen können nicht mit der Gesamtstichprobe verglichen werden, da sich die Anzahl der Fälle halbiert hat.

Die Nullhypothese des F-Tests (in der Grundgesamtheit besteht kein linearer Zusammenhang zwischen Regressand und Regressoren) kann in beiden Teilstichproben zum Signifikanzniveau von 1% verworfen werden.

6.8.3 Vergleich der Koeffizienten

Die Modellstruktur in den Stichproben hat sich folgendermaßen verändert:

- Die ermittelten Koeffizienten haben sich betragsmäßig erhöht oder verringert. Einzig der Koeffizient des Kaufkraftindexes ist bei beiden Teilstichproben relativ stabil geblieben mit Werten von 0,032 bzw. 0,038 (Gesamtstichprobe: 0,033).
- Bei der 50%-Zufallsstichprobe mit den 172 Fällen wird die Variable „Sozialversicherungspflichtig Beschäftigte je Einwohner in 30km Umkreis“ nicht mehr signifikant. Der „Gemeindetyp 6“ trat in der Teilstichprobe nicht auf und wurde somit ebenfalls nicht mehr signifikant.
- Bei der zweiten 50%-Zufallsstichprobe mit 148 Fällen werden die „Arbeitslosenquote“ und „Anzahl Übernachtungsgäste pro Einwohner“ nicht mehr signifikant.

Das Regressionsmodell hat sich als nicht sehr stabil erwiesen, was mit der geringen Fallzahl in den Teilstichproben zusammenhängen kann. Variablen, die in der Gesamtstichprobe in einer angemessenen Anzahl vertreten waren, weisen in den Teilstichproben eine zu geringe Besetzung auf. „Gemeindetyp 6“ ist beispielsweise in der ersten 50%-Zufallsstichprobe überhaupt nicht enthalten, wodurch er natürlich auch nicht mehr signifikant werden kann. Dadurch, dass weniger Variablen signifikant geworden sind, lassen sich die Veränderungen bei den Koeffizienten-Höhen erklären.

7 Auswertung der Untersuchungsergebnisse

Das Untersuchungsziel dieser Ausarbeitung ist zu ermitteln, welche der untersuchten Regionalfaktoren einen signifikanten – also einen nicht zufälligen – Einfluss auf die Miethöhe ausüben. In diesem Kapitel sollen nun die Ergebnisse aus den Analysen der beiden Datensätze ausgewertet werden. Dazu wird zuerst wieder der Mietspiegel-Datensatz betrachtet und im Anschluss daran die Ergebnisse des IVD-Datensatzes ausgewertet.

7.1 Auswertung der Ergebnisse des Mietspiegel-Datensatzes

Vergleicht man die beiden Regressionsmodelle miteinander, so fällt auf, dass mit der eigentlichen Regressionsanalyse eine höhere Modellgüte erzielt werden kann. Das R^2 nimmt im ersten Modell einen Wert von 47,7% an, während im zweiten Modell lediglich ein R^2 von 43% erreicht wird. Auch das korrigierte R-Quadrat ($\overline{R}^2$) fällt bei Modell 1 mit einem Wert von 46,7% deutlich besser aus als in Modell 2 mit 41,8%. Die Regressionsgerade des ersten Modells bildet somit die abhängige Variable besser ab. Diese Annahme bestätigt auch der Vergleich der Standardfehler der Schätzer. Dieser nimmt im ersten Modell mit 14,7% (in Bezug auf den Mittelwert der abhängigen Variablen) einen geringeren Wert an, als im zweiten Modell mit 15,4%.

Das schlechtere Ergebnis unter Einbezug einer Faktorenanalyse kann zum einen an den Ausgangsdaten liegen, da die Korrelationsmatrix nach der Image-Analyse von Gutmann nicht für die Faktorenanalyse geeignet ist. Dadurch konnten die Korrelationen bei der Faktorenanalyse nicht besonders gut reproduziert werden und das Modell die Originaldaten nicht optimal darstellen (41% der Residuen wiesen einen Wert größer 0,05 auf). Die extrahierten drei Faktoren wiederum waren inhaltlich gut zu interpretieren und erklärten einen Gesamtvarianzanteil von 77%. In der anschließenden Regressionsanalyse wurden alle drei Faktoren signifikant, d.h. sie üben keinen zufälligen Einfluss auf die Miethöhe aus.

Den stärksten positiven Einfluss auf die Nettomiete in Modell 1 übt die Variable „Kaufkraftindex“ aus mit einem standardisierten Koeffizienten von 0,584.

Als problematisch bei den durchgeführten Analysen erwiesen sich die wenigen Fälle in den Datengrundlagen. Des Weiteren sind die Mietspiegel-Daten von Beginn an mit starken Schwankungen belegt gewesen. Diese entstehen durch die großen Abweichungen in den Erstellungsmethoden. Mit der zusätzlichen Einschränkung durch die Standardisierung eines Wohnungstyps ist eine zusätzliche Streuung in den Daten entstanden. Beispielsweise variierte es von Mietspiegel zu Mietspiegel, ob ein Modernisierungszuschlag für die 3-Zimmer-Wohnung erhoben werden konnte. In anderen Mietspiegeln wurde die Sanierung überhaupt nicht berücksichtigt oder es gab die Möglichkeit, die Wohnung in eine neuere Baualtersklasse einzuordnen.

Daher wäre es sinnvoll und wünschenswert, wenn es Vereinheitlichungen für Mietspiegel geben würde, da dann eine bessere Vergleichbarkeit gewährleistet wäre.

Werden anhand der aufgestellten Regressionsgleichungen die Nettomieten für alle Städte und Gemeinden in Deutschland berechnet, ergeben sich die folgenden Bilder:

- Bei beiden Modellen lassen sich eindeutig die Großstädte und Ballungsräume identifizieren.
- Die Unterschiede der Mietpreise zwischen Stadt und Land werden in Modell 1 deutlicher sichtbar.
- Modell 2 überschätzt vor allem in den Städten mit einer hohen Einwohnerzahl (München, Hamburg, Berlin, s. Abbildung 41)

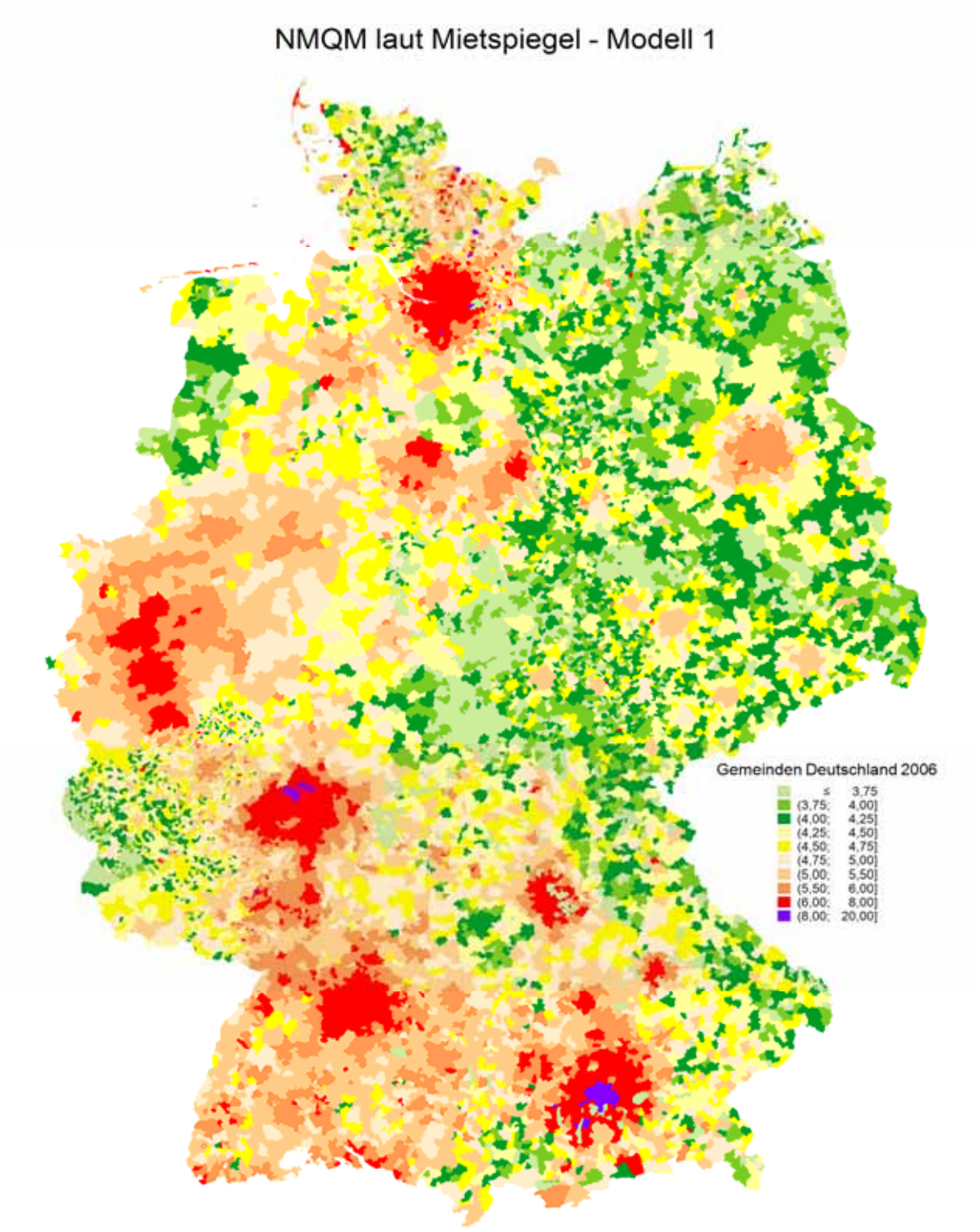


Abbildung 39: Mietpreise in Deutschland laut Mietspiegel[115] - Modell 1

[115] Quelle: Eigene Darstellung; Berechnung der Mietpreise erfolgte auf Basis von Modell 1

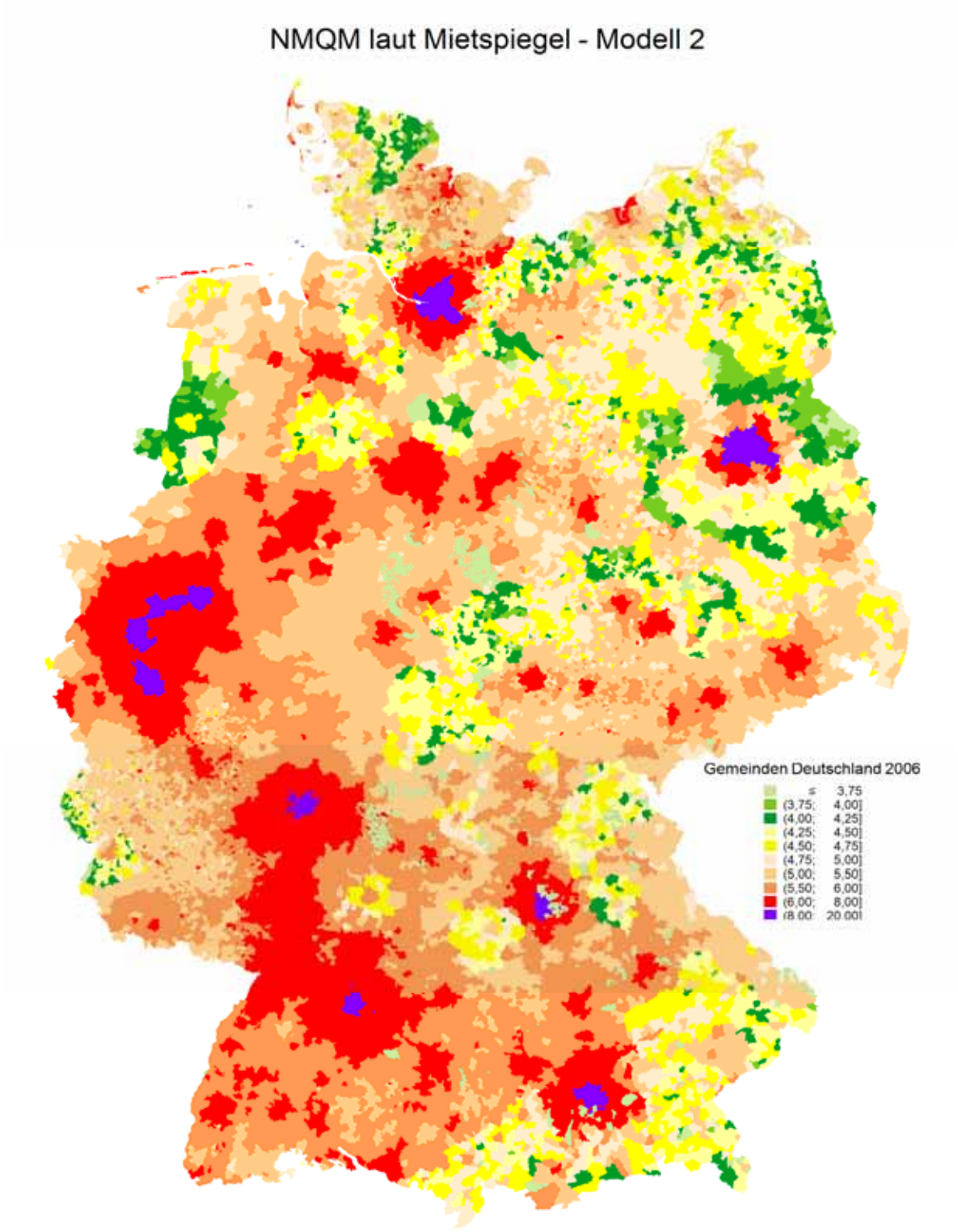


Abbildung 40: Mietpreise in Deutschland laut Mietspiegel[116]-Modell 2

[116] Quelle: Eigene Darstellung; Berechnung der Mietpreise erfolgte auf Basis von Modell 2

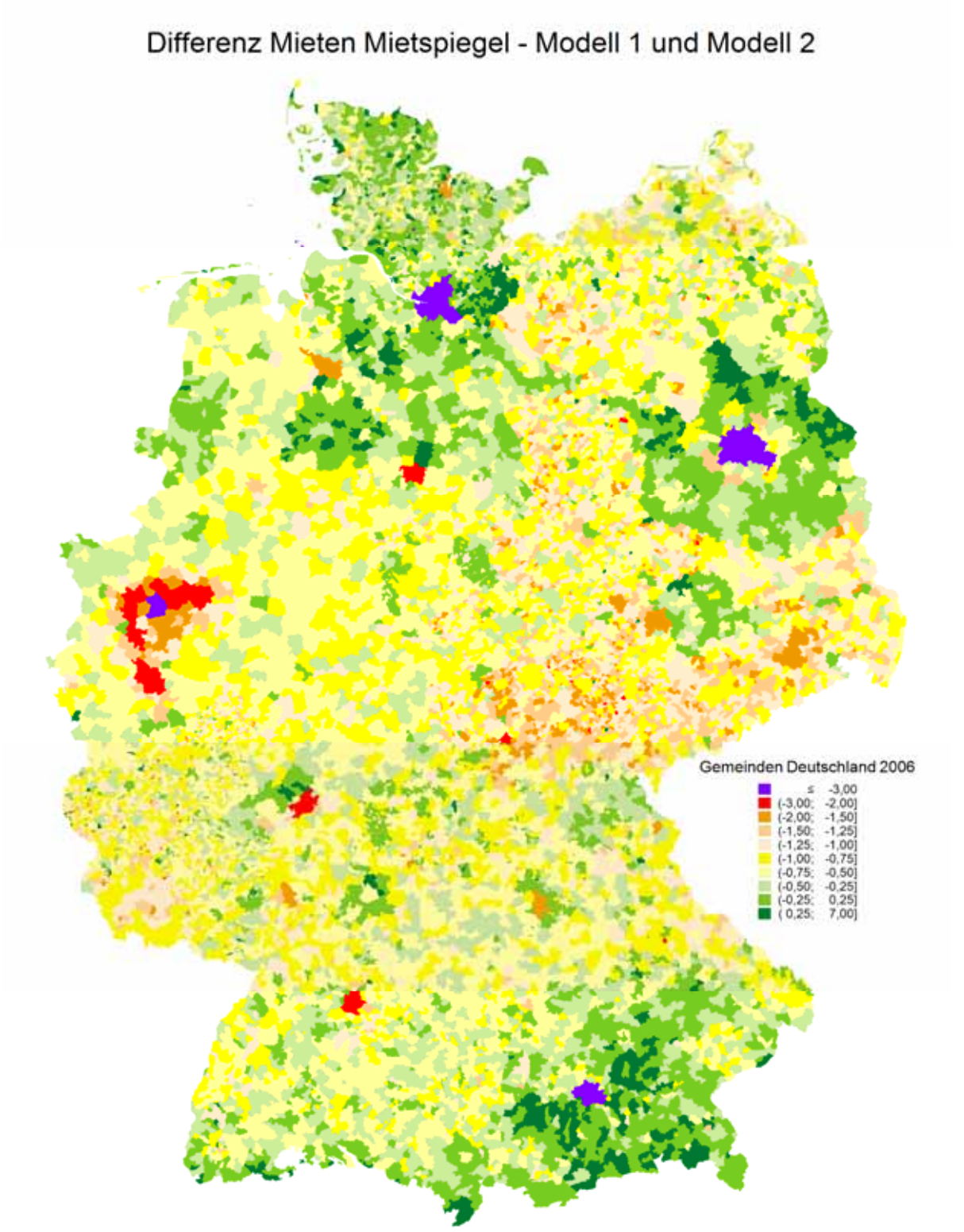


Abbildung 41: Differenz Mieten Mietspiegel[117] - Modell 1 und 2

[117] Quelle: Eigene Darstellung

7.2 Auswertung der Ergebnisse des IVD-Datensatzes

Bei der Auswertung der Ergebnisse aus den Analysen des IVD-Datensatzes ergibt sich ein ähnliches Bild wie bei den Mietspiegel-Modellen.

Mit Hilfe der eigentlichen Regressionsanalyse bekommt man eine bessere Schätzung der Nettomieten, als wenn die Faktorenanalyse mit einbezogen wird. Dies kann auch hier an den Ausgangsdaten liegen, da die Korrelationsmatrix nach der Image-Analyse von Gutmann ebenfalls nicht für die Faktorenanalyse geeignet ist. Dadurch konnten die Korrelationen bei der Faktorenanalyse nicht besonders gut reproduziert werden und das Modell die Originaldaten nicht optimal darstellen (58% der Residuen wiesen einen Wert > 0,05 auf). Die extrahierten drei Faktoren wiederum waren inhaltlich gut zu interpretieren und erklärten einen Gesamtvarianzanteil von knapp 77%. In der anschließenden Regressionsanalyse wurden alle drei Faktoren signifikant, d.h. sie üben keinen zufälligen Einfluss auf die Miethöhe aus. Trotzdem lässt sich mit der reinen Regressionsanalyse eine höhere Modellgüte erzielen. Hier lag des R^2 bei 54,9%, bei vorgeschalteter Faktorenanalyse war nur ein R^2 von 50,3% zu erzielen. Die Standardfehler der Schätzer lagen im ersten Modell (in Bezug auf den Mittelwert der abhängigen Variablen) bei 14,4% und im zweiten Modell bei 15,1%.

Das Ergebnis von Modell 1 zeigt, dass die Variablen „Kaufkraftindex“ und „Anzahl der Übernachtungsgäste pro Einwohner“ einen hohen Einfluss auf die Nettomiete haben. Die Bevölkerungsdichte lässt sich als „Stadtindikator“ interpretieren, da sie nur einen geringen Einfluss in den ländlichen Regionen hat, aber einen hohen Einfluss in den Städten.

Bsp.

Dichte	Zuschlag in Cent/m²
100	3,9
200	7,8
300	11,7
1000	39

Als problematisch bei den Untersuchungen erweist sich natürlich die geringe Datengrundlage, die mit 350 Fällen zu niedrig liegt. Dadurch sind verschiedene Gemeindetypen deutlich unterrepräsentiert, was zu hohen Abweichungen bei der Berechnung der Nettomiete führt. Des Weiteren sind die Daten von Anfang an mit starken Schwankungen belegt. Für die IVD-Daten, die von den für den IVD tätigen Maklern erhoben werden, gibt es keine Informationen darüber, wie diese Erhebungen genau durchgeführt werden und welcher Wohnungstyp unter „guten Wohnwert“ fällt.

Da nicht für alle Städte und Gemeinden in Deutschland Zahlen zu den Übernachtungsgästen vorliegen, können die Nettomieten nicht für alle Städte und Gemeinden berechnet werden. In den nachfolgenden Grafiken sind diese Gebiete deshalb weiß unterlegt. Werden die Netto-

mieten (auf Basis der bestimmten Regressionsgleichungen aus Modell 1 und 2) für alle Städte und Gemeinden in Deutschland berechnet, so ergeben sich die folgenden Darstellungen:

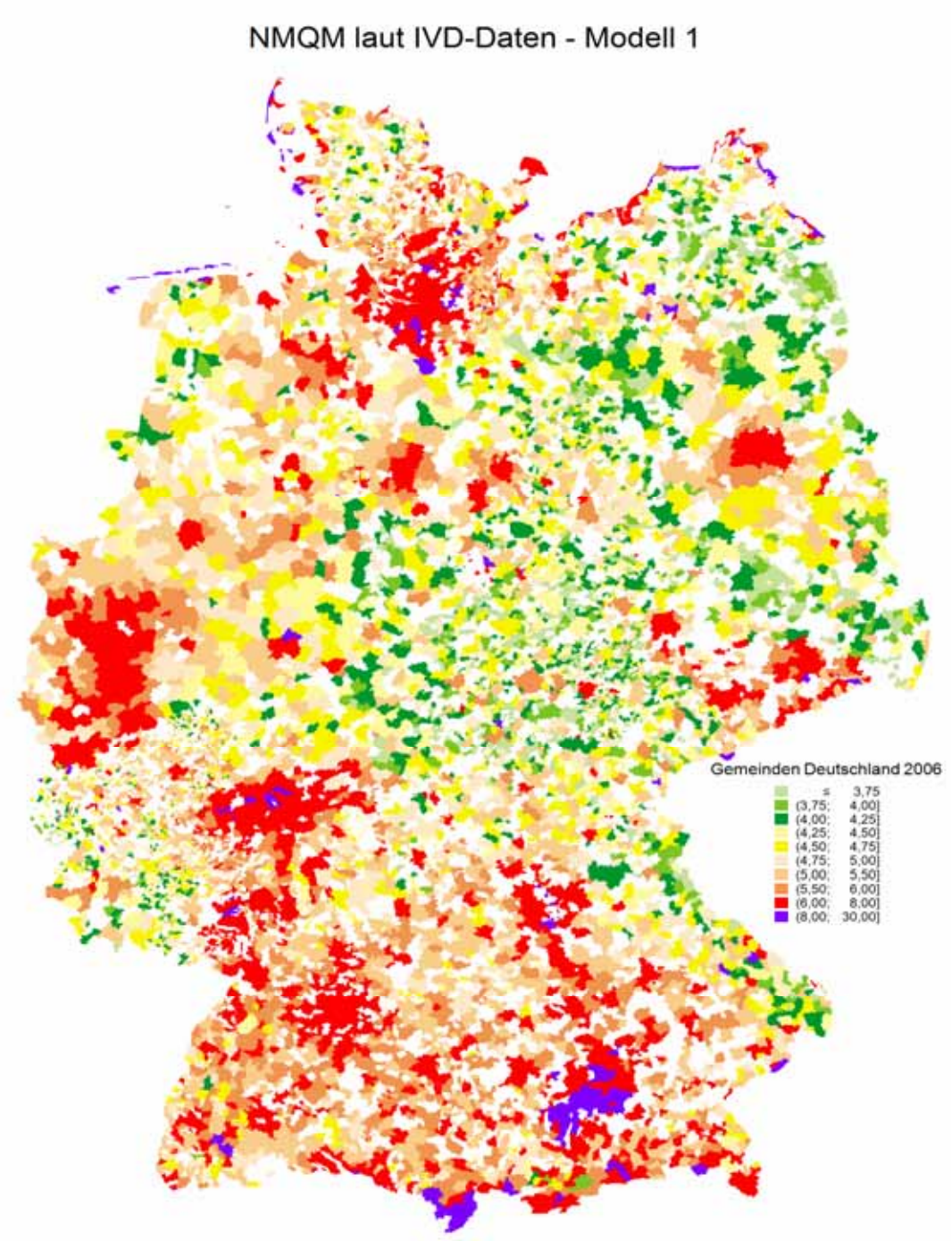


Abbildung 42: Mietpreise in Deutschland laut IVD-Modell[118] 1

[118] Quelle: Eigene Darstellung; Berechnung der Mietpreise erfolgte auf Basis von Modell 1

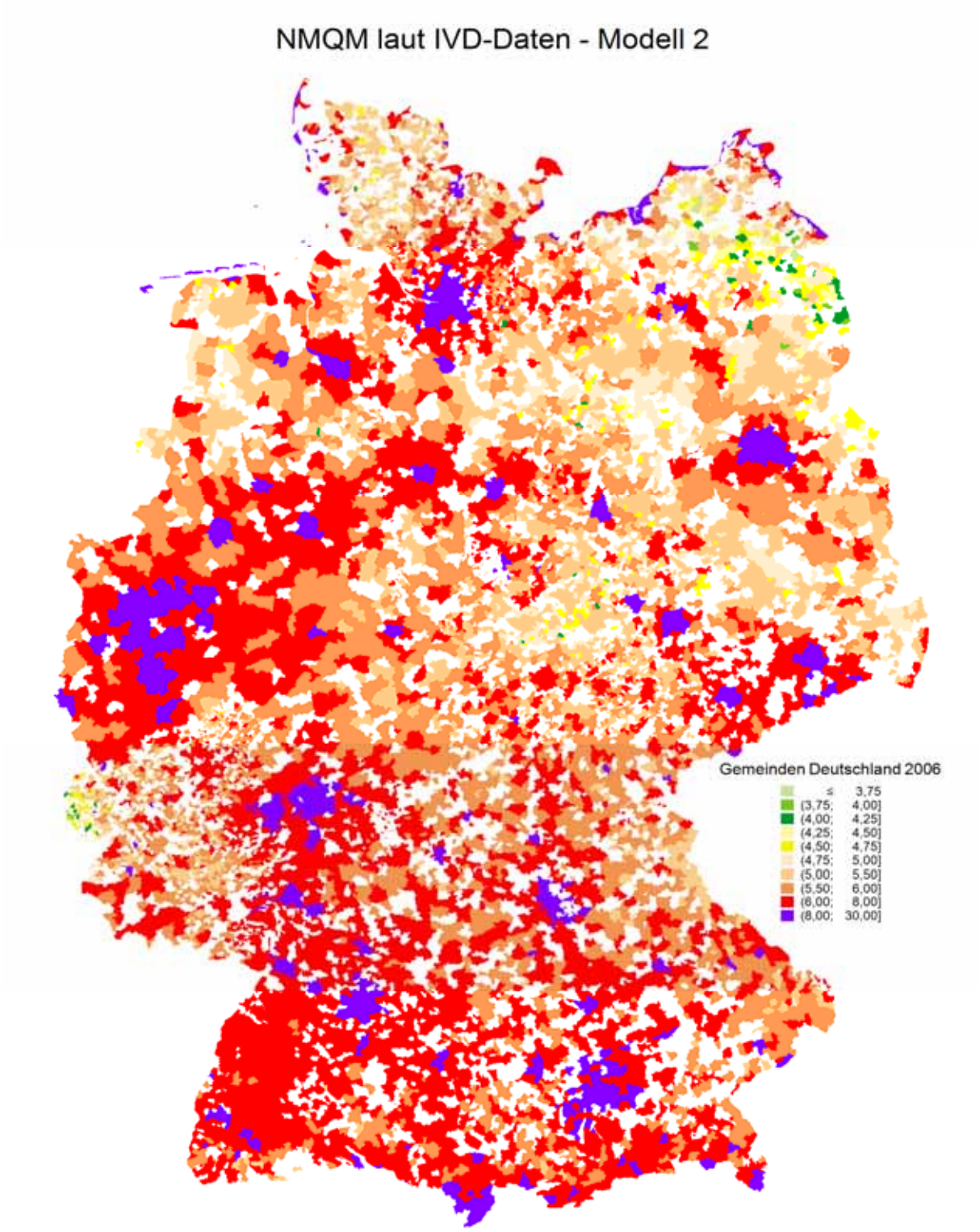


Abbildung 43: Mietpreise in Deutschland laut IVD-Modell[119] 2

[119] Quelle: Eigene Darstellung; Berechnung der Mietpreise erfolgte auf Basis von Modell 2

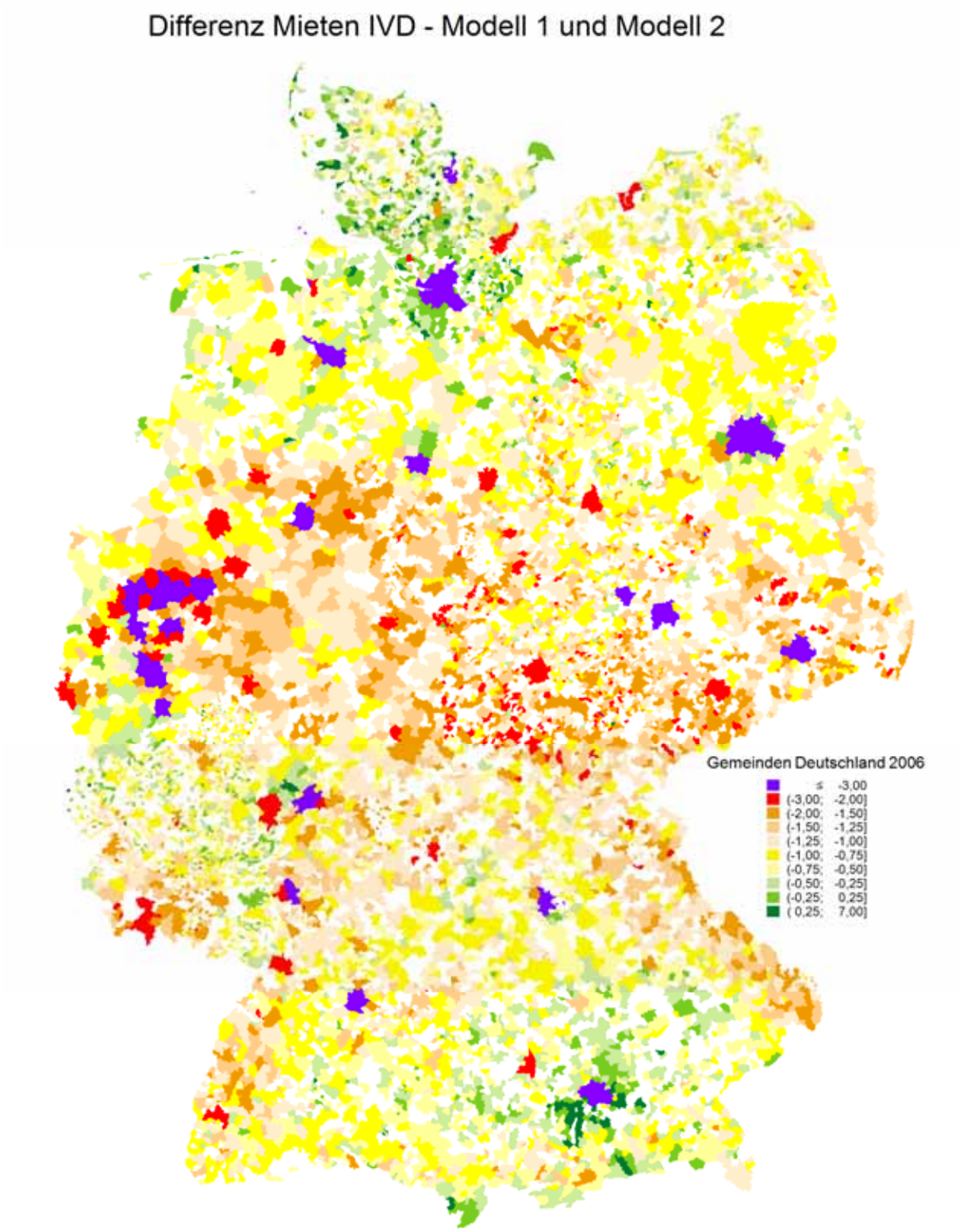


Abbildung 44: Differenz Mieten IVD-Modell[120] 1 und 2

[120] Quelle: Eigene Darstellung

8 Zusammenfassung und Ausblick

In der vorliegenden Diplomarbeit wurde der Einfluss regionaler Faktoren auf den Wohnungsmietpreis in Deutschland untersucht.

Dazu wurden zwei verschiedene Datensätze analysiert: Der erste Datensatz enthielt durch Auswertung von Mietspiegeln ermittelte Wohnungsmieten; der zweite Datensatz basierte auf Mietpreiserhebungen des Immobilienverband Deutschland. Als regionale Einflussfaktoren wurden vom Statistischen Bundesamt, der Bundesagentur für Arbeit u.a. Daten über Einwohnerzahlen, Arbeitslosenquoten, Kaufkraftindex etc. für alle deutsche Städte und Gemeinden erhoben, die dann als unabhängige Variablen in die Analysen mit eingeflossen sind. Die abhängige Variable, die erklärt werden sollte, trat in allen Modellen als „monatliche Nettomiete in Euro pro m² (nmqm)“ auf.

Beide Datensätze wurden mit Hilfe von Regressionsanalysen untersucht. Für den Mietspiegel-Datensatz ließ sich damit ein Bestimmtheitsmaß von 47,7% erzielen. In den IVD-Datensatz wurde eine weitere Variable mit aufgenommen, wodurch sich ein Bestimmtheitsmaß von 54,9% erzielen ließ. Dabei hat sich herausgestellt, dass vor allem der Kaufkraftindex einen hohen positiven Einfluss auf den Mietpreis hat. Dagegen hat der Anteil der Wohnungen in Zweifamilienhäusern einen negativen Effekt. Die Bevölkerungsdichte hat sich als reiner Stadt-Indikator identifizieren lassen. Ihr Einfluss ist in den ländlichen Regionen sehr gering, in den Großstädten dagegen sehr stark.

Die Regressionsmodelle wurden jeweils anhand einer 50%-igen Zufallsstichprobe getestet. Hierbei stellte sich heraus, dass die Regressionsgleichungen in Hinblick auf die Reduzierung der Fallzahl keinen stabilen Eindruck hinterlassen haben. Die signifikanten Variablenstrukturen unterschieden sich im Vergleich zu den Gesamtmodellen, was auf eine zufallsbedingte Schätzung der Koeffizienten und ihrer Signifikanzen hindeuten kann. Die Ursache hierfür ist aber eher in den Stichproben selbst zu sehen, die zum einen nicht repräsentativ waren und zum anderen eine viel zu geringe Fallzahl aufweisen. Die Ergebnisse der Gesamtmodelle erwiesen sich durchaus als brauchbar.

Die durchgeführten Faktorenanalysen mit anschließender Regression der extrahierten Faktoren, führten dagegen nicht zu den gewünschten Ergebnissen. Zwar ließen sich die extrahierten Faktoren inhaltlich gut interpretieren, es ergaben sich jedoch in den anschließenden Regressionsanalysen deutlich schlechtere Modellanpassungen. Dies ist aber mit großer Wahrscheinlichkeit wieder auf die geringen Datengrundlagen zurückzuführen. Auch die Ergebnisse der Faktorenanalyse wurden anhand von 50%-Zufallsstichproben überprüft. Die Resultate erwiesen sich bis auf kleinere Abweichungen als relativ stabil.

Die Hochrechnung der Mieten für alle Städte und Gemeinden Deutschlands ergab, dass mit den Modellen die Mieten in den Städten tendenziell überschätzt werden, in den ländlichen Gebieten dagegen eher unterschätzt. In einem nächsten Analyseschritt wäre z.B. zu untersuchen, für welche Gemeindegrößen die Hochrechnungen überhaupt sinnvoll sind, da es sich um keine repräsentativen Stichproben handelte. Durch die Auswertung von Mietspiegeln

(die hauptsächlich in größeren Städten und Gemeinden (über 20.000 Einwohner) erstellt werden) ergibt sich ein verzerrter Ausschnitt aus allen Gemeinden und vermutlich können die kleineren Gemeinden nicht adäquat abgebildet werden. Auch die Erhebungen des IVD werden tendenziell für größere Städte und Gemeinden durchgeführt.

Welches der beiden Modelle besser geeignet ist, um eine Hochrechnung der Mieten für alle Städte und Gemeinden durchzuführen, lässt sich abschließend nicht pauschal beurteilen. Die Auswertung des IVD-Datensatzes mit Hilfe einer Regressionsanalyse führte zu einem deutlich höheren R^2 (54,9%), als es beim Mietspiegel-Datensatz der Fall war (dort betrug das R^2 47,7%). Allerdings ist die Modellierung der Mieten anhand des IVD-Datensatzes ebenfalls in Teilsegmenten verbesserungsfähig und könnte weiterentwickelt werden. Beispielsweise werden unter dem IVD-Datensatz in sehr kleinen Gemeinden mit hohen Übernachtungszahlen extrem hohe Mieten berechnet, was nicht der Wirklichkeit entspricht.

Ein weiterer Punkt der zu analysieren wäre, sind die Ausreißer im Datensatz. Diese wurden in den vorgenommenen Untersuchungen nicht weiter berücksichtigt. Da eindeutig Ausreißer zu identifizieren sind, könnte mit tiefer gehenden Analysen eine Modelloptimierung erreicht werden.

9 Literaturverzeichnis

9.1 Bücher

BACH Bach, C. (2008): Unterlagen zur Vorlesung Statistik 2 aus dem Sommersemester 2008, Hochschule Darmstadt

BACK Backhaus, K., Erichson, B., Plinke, W., Weiber, R.(2003): Multivariate Analysemethoden - Eine anwendungsorientierte Einführung; Springer: Berlin, Heidelberg 2006, 11. Auflage

BBR Bundesamt für Bauwesen und Raumordnung (2004): Wohnungsmärkte in Deutschland, Ausgabe 2004, Bundesamt für Bauwesen und Raumordnung (Hrsg.), Bonn 2004

BGB BGB – Bürgerliches Gesetzbuch (2006), §§558 ff.; DTV-Beck: München 2006, 58. Aufl.

BMVBS Bundesministerium für Verkehr, Bau- und Wohnungswesen (2002): Hinweise zur Erstellung von Mietspiegeln, Bundesministerium für Verkehr, Bau- und Wohnungswesen (Hrsg.), Berlin

BORTZ Bortz, J. (1993): Statistik für Sozialwissenschaftler; Springer: Berlin, Heidelberg 2005, 6. Auflage

DESTATIS Statistisches Bundesamt Deutschland (2008): DVD Statistik lokal; Ausgabe 2008, Statistisches Bundesamt Deutschland (Hrsg.), Wiesbaden 2008

DESTATIS2 Statistisches Bundesamt Deutschland (2009): Zuhause in Deutschland - Ausstattung und Wohnsituation privater Haushalte - Ausgabe 2009, Statistisches Bundesamt Deutschland (Hrsg.), Wiesbaden, März 2009

DESTATIS3 Statistisches Bundesamt Deutschland (2006): Bevölkerung und Erwerbstätigkeit – Struktur der sozialversicherungspflichtig Beschäftigten – Ausgabe 2006, Statistisches Bundesamt Deutschland (Hrsg.), Wiesbaden 2007

FAHR Fahrmeir, L., Kneib, T. Lang, S. (2007): Regression – Modelle, Methoden und Anwendungen; Springer: Berlin, Heidelberg 2007

FAKTOR SPSS 12.0 Installations-CD/Algorithms/factor.pdf

IVD Immobilienverband Deutschland (IVD) Bundesverband (2006): Wohnpreisspiegel 2006/2007, Immobilienverband Deutschland (Hrsg.), Berlin 2006

PFEI Pfeifer, A., Schuchmann, M. (1996): Datenanalyse mit SPSS für Windows; Oldenbourg: München Wien 1996, 2. neu bearbeitete Auflage

SACHS Sachs, L. (1999): Angewandte Statistik- Anwendung statistischer Methoden; Springer: Berlin, Heidelberg 1999, 9. Auflage

SPSS Schübo, W., Uehlinger, H.-M., Perleth, Ch., Schröger, E., Sierwald, W. (1991): SPSS Handbuch der Programmversionen 4.0 und SPSS-X 3.0; Gustav Fischer: Stuttgart 1991

STIER Stier, W. (1999): Empirische Forschungsmethoden; Springer: Berlin, Heidelberg 1999, 2. Auflage

VOSS Voß, W. et al. (2004): Taschenbuch der Statistik; Carl Hanser Verlag: München Wien 2004, 2. verbesserte Auflage

WIRTZ Wirtz, M., Nachtigall, Ch. (2008):Deskriptive Statistik – Statistische Methoden für Psychologen, Teil 1; Juventa Verlag: Weinheim und München 2008, 5. überarbeitete Auflage

9.2 Internetquellen

[1] Bundeszentrale für politische Bildung, http://www.bpb.de/wissen/GLSOS3,0,0,Bev%F6lkerung_und_Haushalte.html (Abgerufen: 26.05.2009)

[2] Wikipedia, Die freie Enzyklopädie. Bearbeitungsstand: 02. Mai 2009 http://de.wikipedia.org/wiki/Kaufkraft_(Konsum) (Abgerufen: 26.05.2009)

[3] Bundesamt für Bauwesen und Raumordnung (BBR), Raumabgrenzungen. http://www.bbr.bund.de/cln_016/nn_103086/BBSR/DE/Raumbeobachtung/Werkzeuge/Raumabgrenzungen/SiedlungsstrukturelleGebietstypen/Gemeindetypen/Downloadangebote.html (Abgerufen: 26.05.2009)

[4] Bundesamt für Bau-, Stadt- und Raumforschung (BBSR), Raumabgrenzungen. http://www.bbsr.bund.de/cln_016/nn_103086/BBSR/DE/Raumbeobachtung/Werkzeuge/Raumabgrenzungen/SiedlungsstrukturelleGebietstypen/gebietstypen.html (Abgerufen: 26.05.2009)

[5] Bundesamt für Bau-, Stadt- und Raumforschung (BBSR), Wohnungsmarktbeobachtung. http://www.bbsr.bund.de/cln_005/nn_22382/BBSR/DE/Fachthemen/Wohnungswesen/Wohnungsmarkt/Wohnungsmarktbeobachtung/Mietspiegel/Mietspiegel.html (Abgerufen: 28.05.2009)

[6] Applied Factor Analysis by R. J. Rummel, S. 213ff http://books.google.de/books?id=g_eNa_XzyEIC&printsec=frontcover&source=gbs_navlinks_s#v=onepage&q=&f=false (Abgerufen: 17.08.2009)

Anhang

Mietspiegelverbreitung nach Erstellungsform

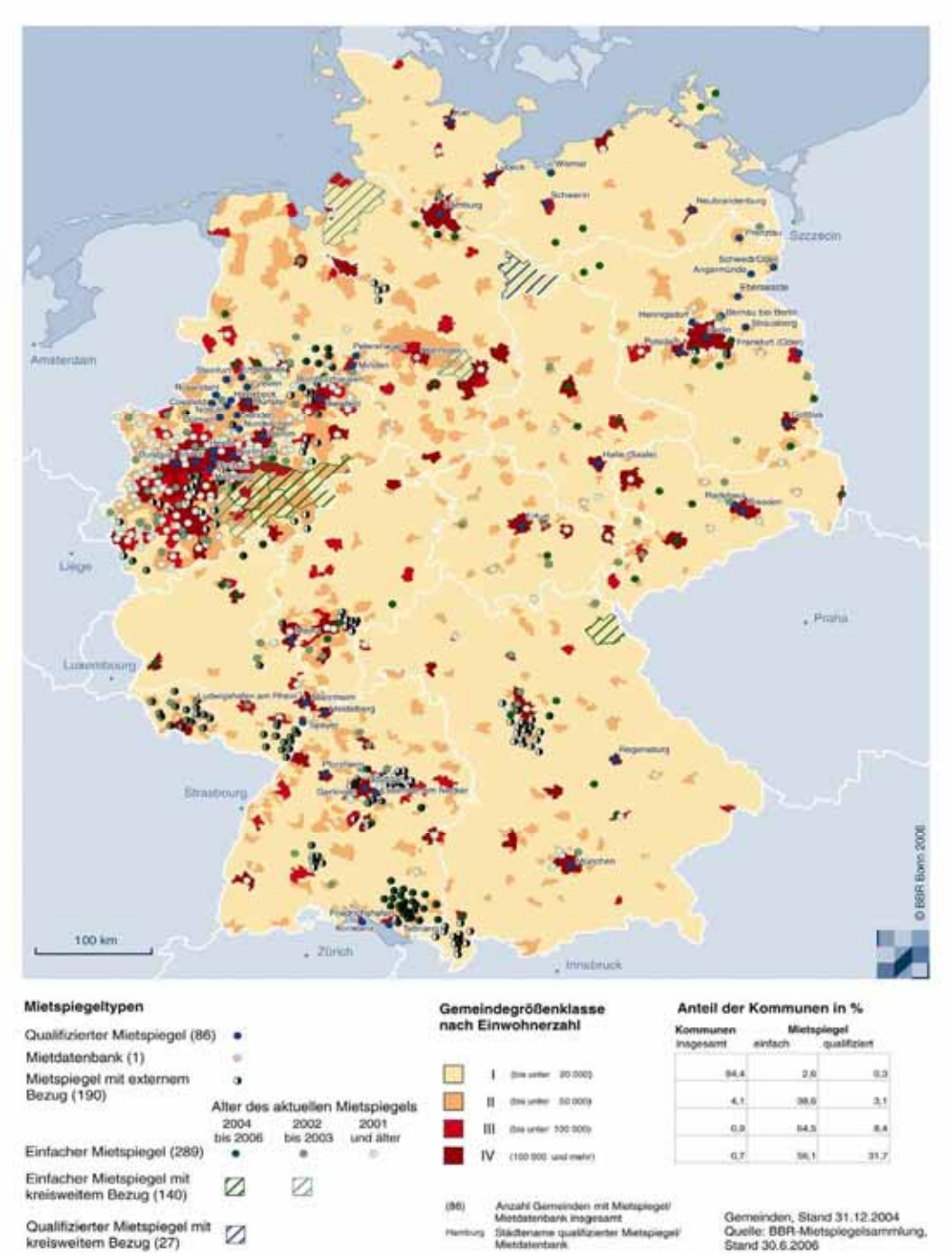


Abbildung 45: Mietspiegelverbreitung[121]

[121] Quelle: [5]

Grafische Darstellung der erklärenden Variablen

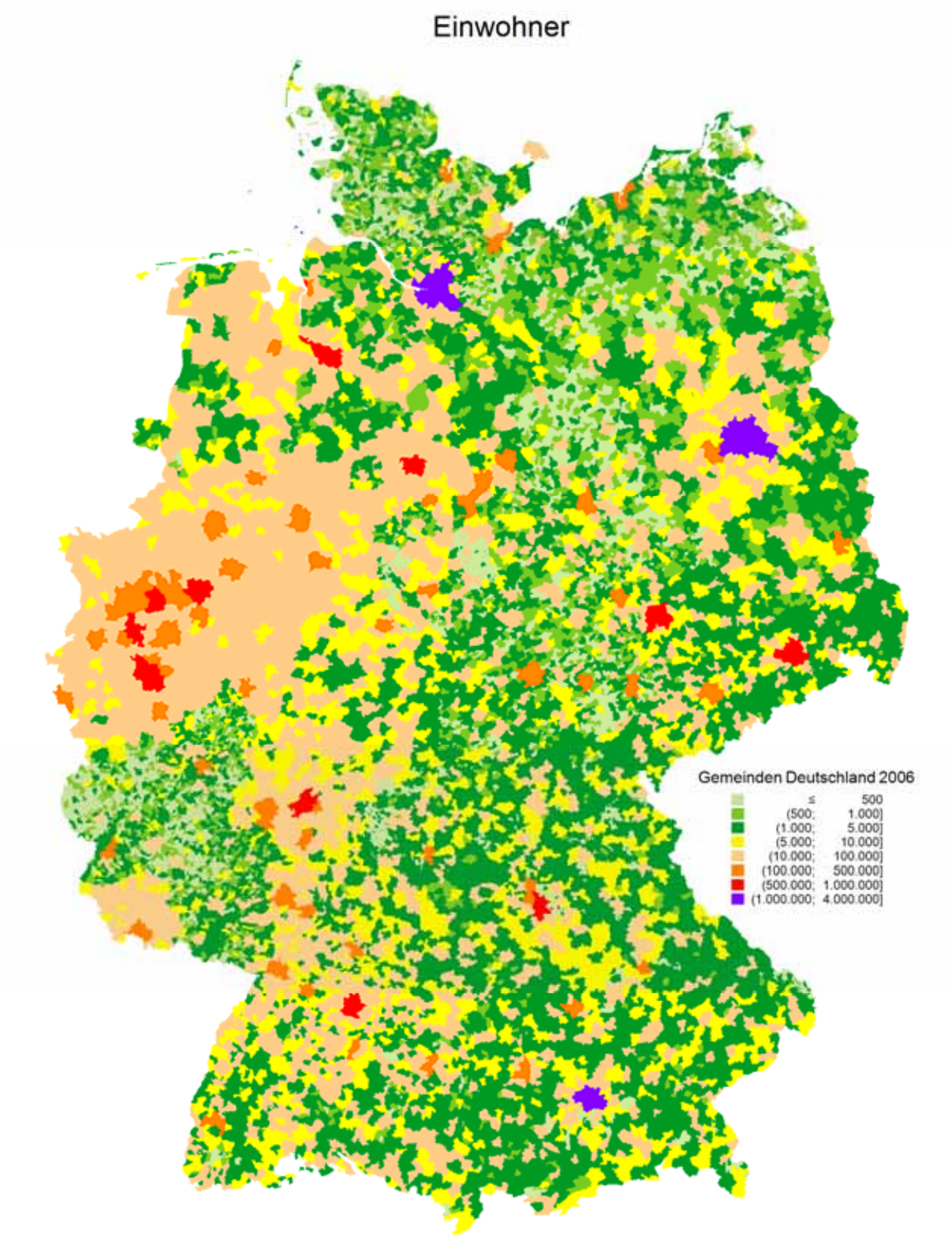


Abbildung 46: Einwohnerzahl (Stand: 31.12.2006)

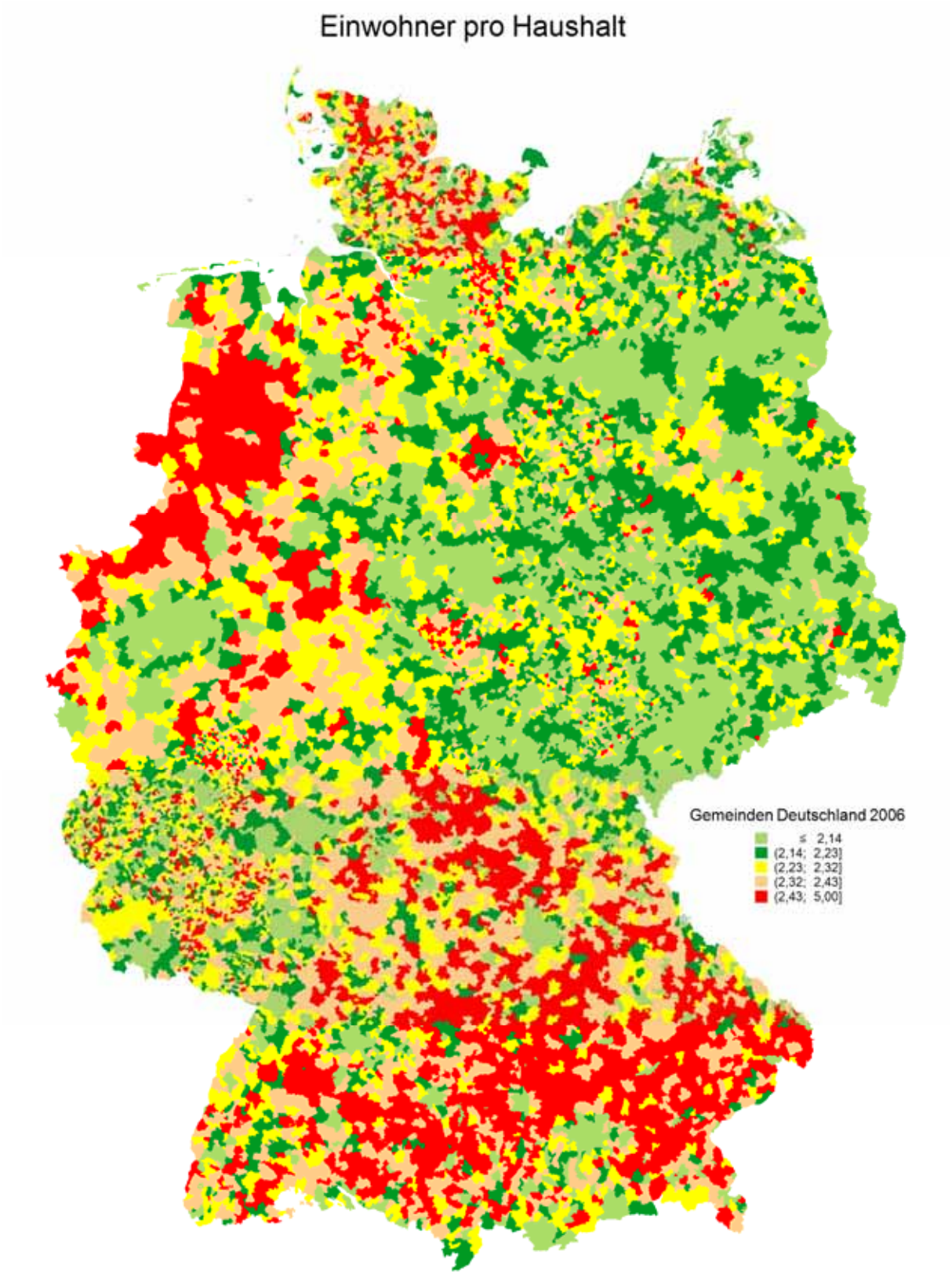


Abbildung 47: Einwohnerzahl pro Haushalt (Stand 31.12.2006)

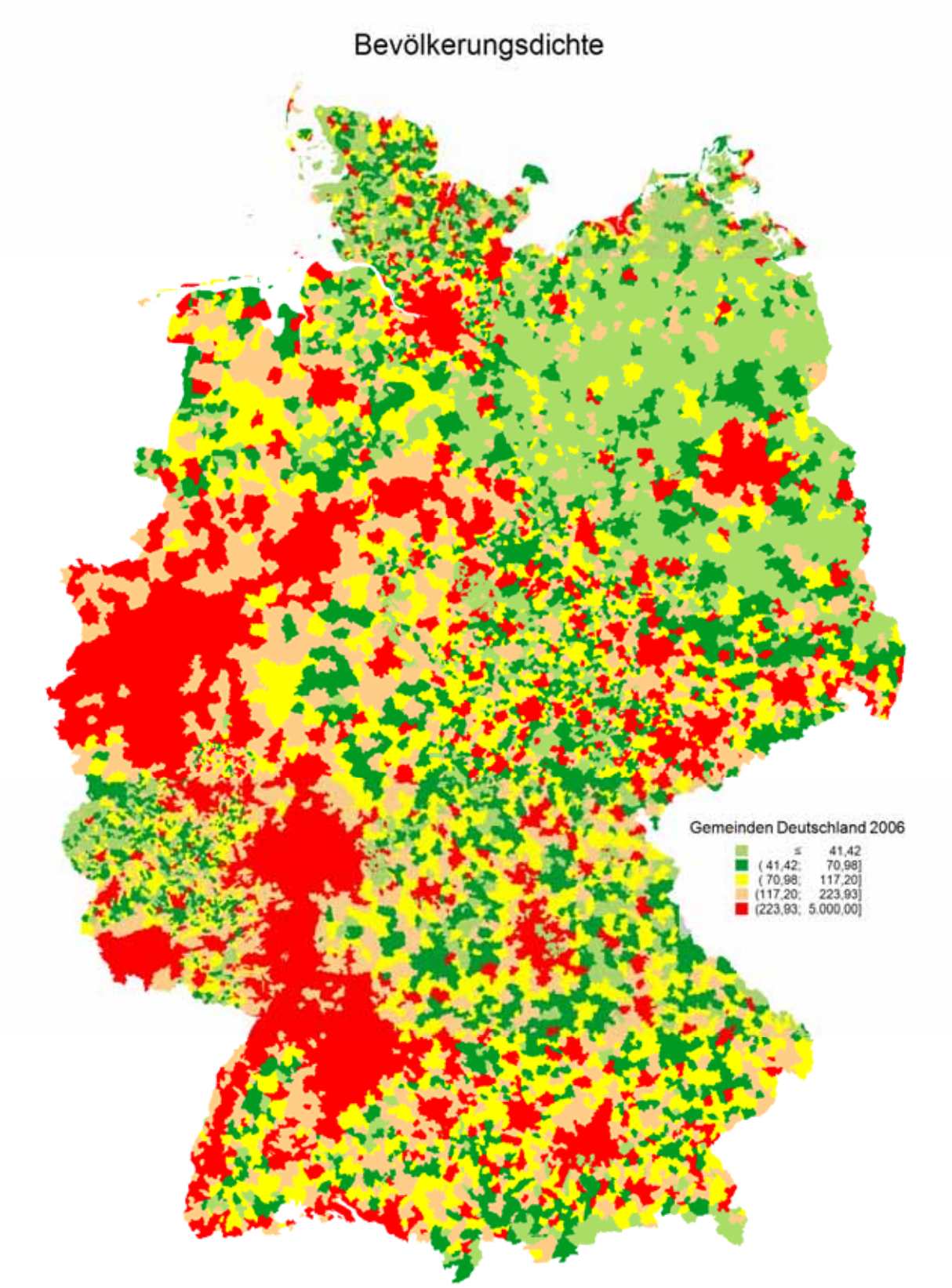


Abbildung 48: Einwohner je km² (Stand: 31.12.2006)

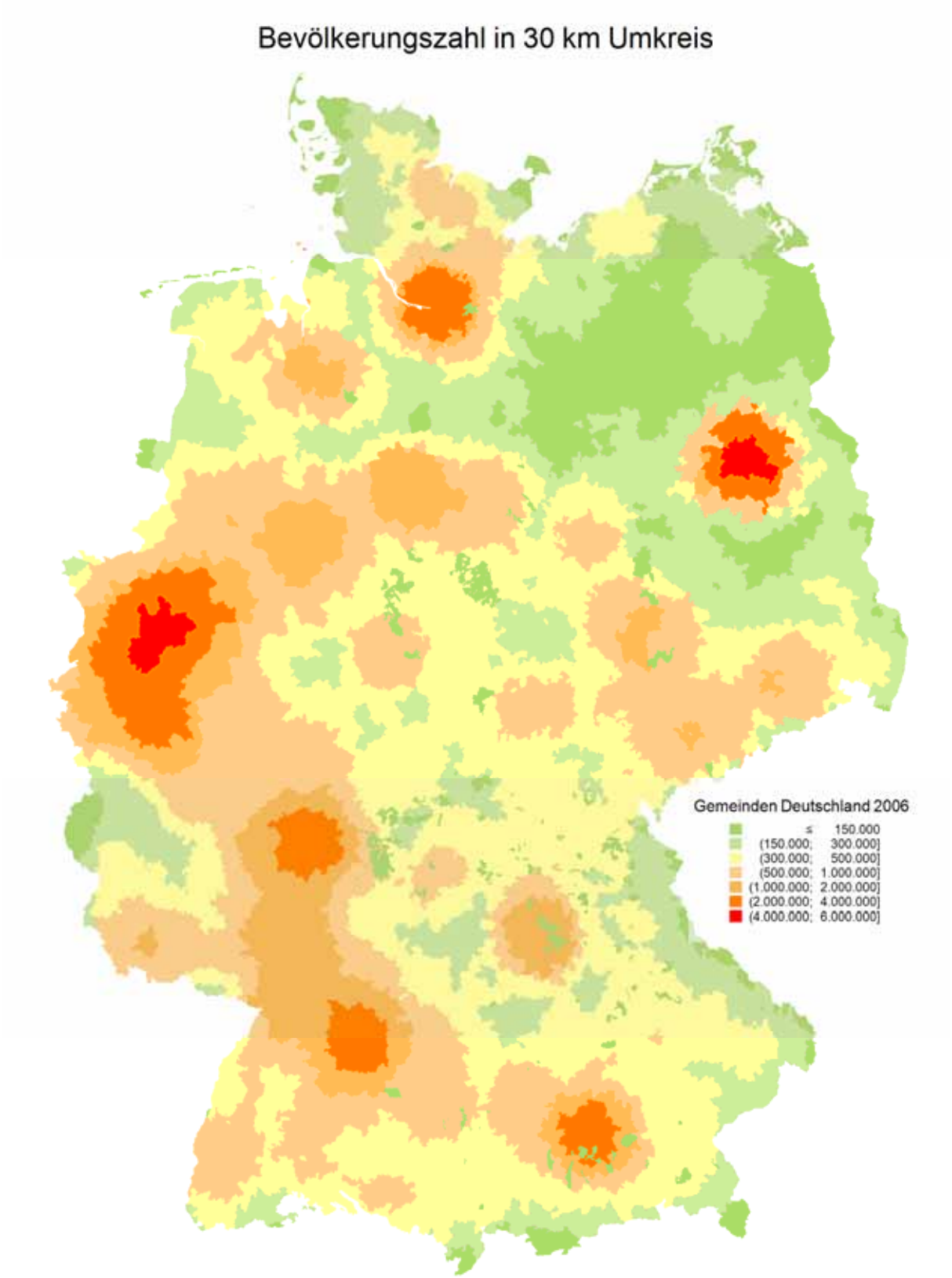


Abbildung 49: Einwohner in 30km Umkreis (Stand 31.12.2006)

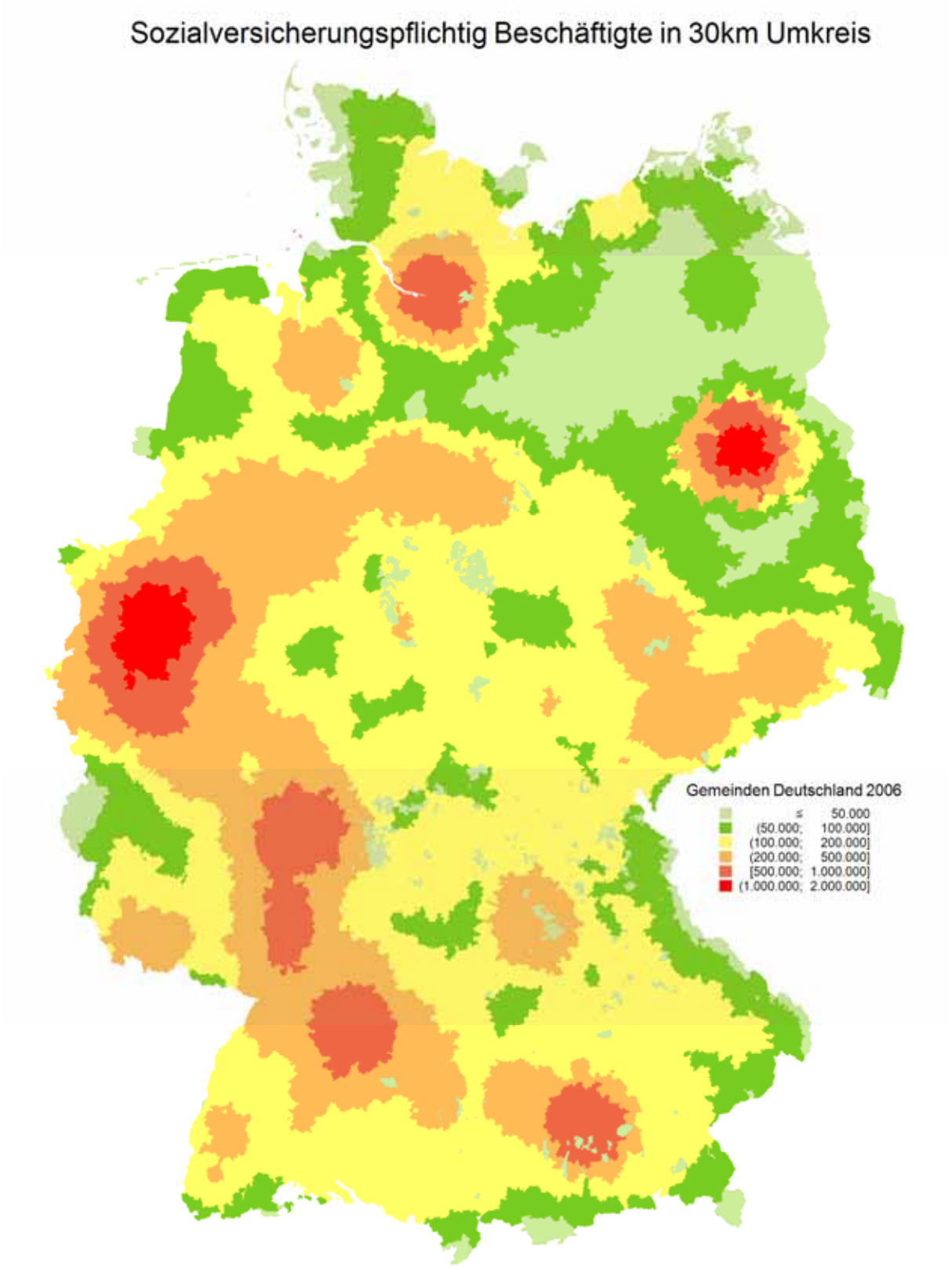


Abbildung 50: Sozialversicherungspflichtig Beschäftigte in 30km Umkreis (Stand 30.06.2006)

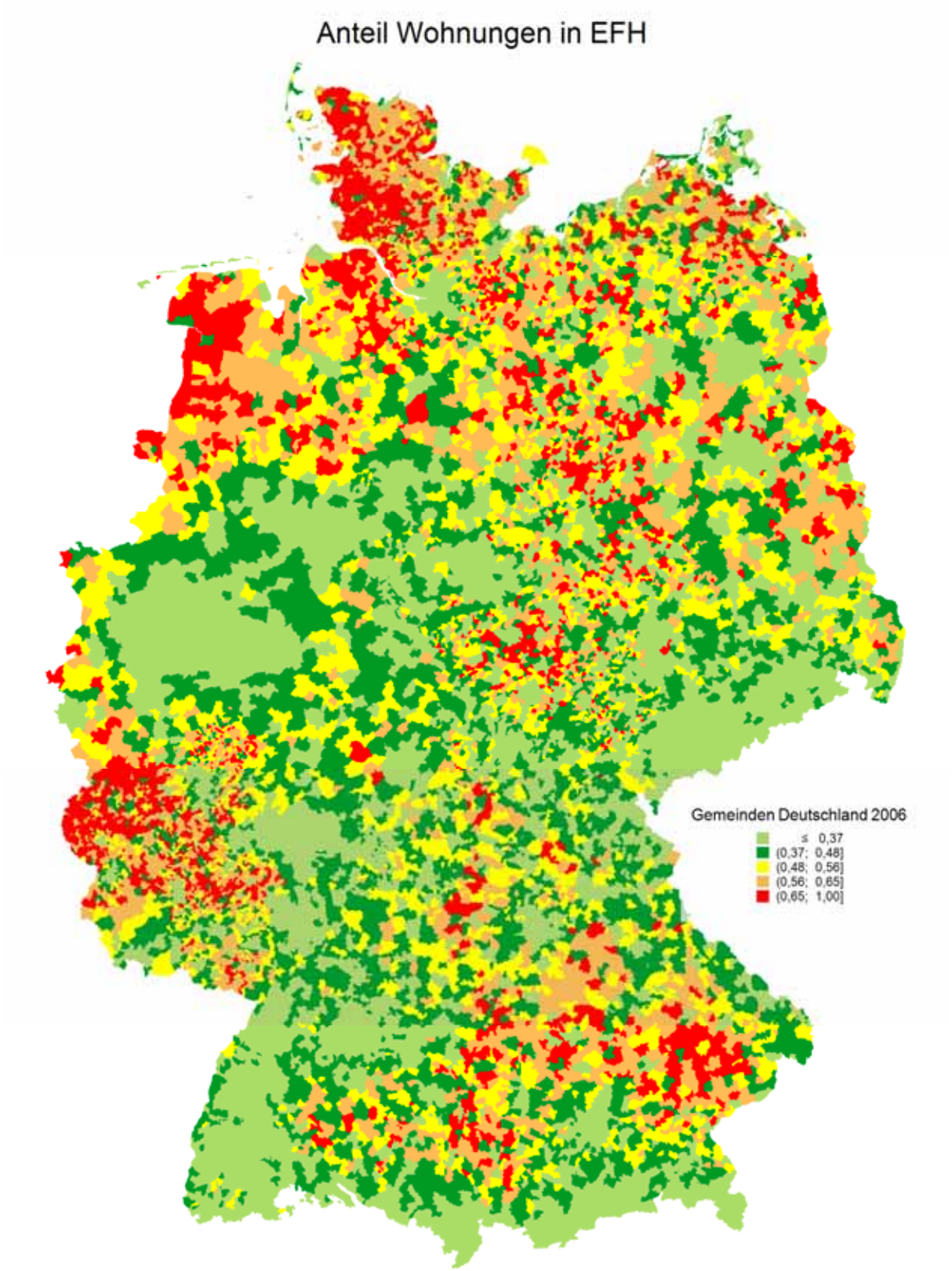


Abbildung 51: Anteil Wohnungen in Einfamilienhäusern (Stand: 31.12.2006)

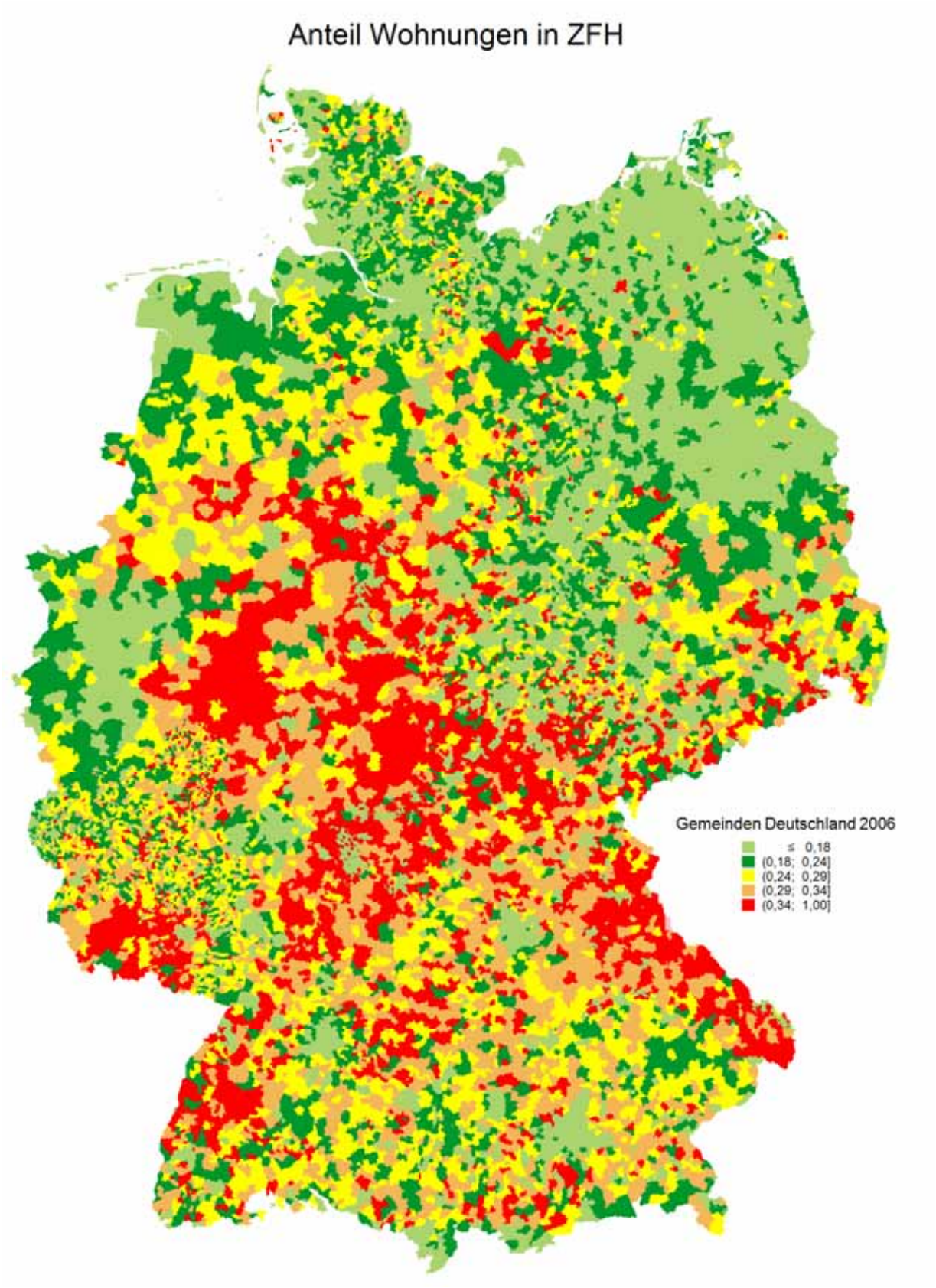


Abbildung 52: Anteil Wohnungen in Zweifamilienhäusern (Stand: 31.12.2006)

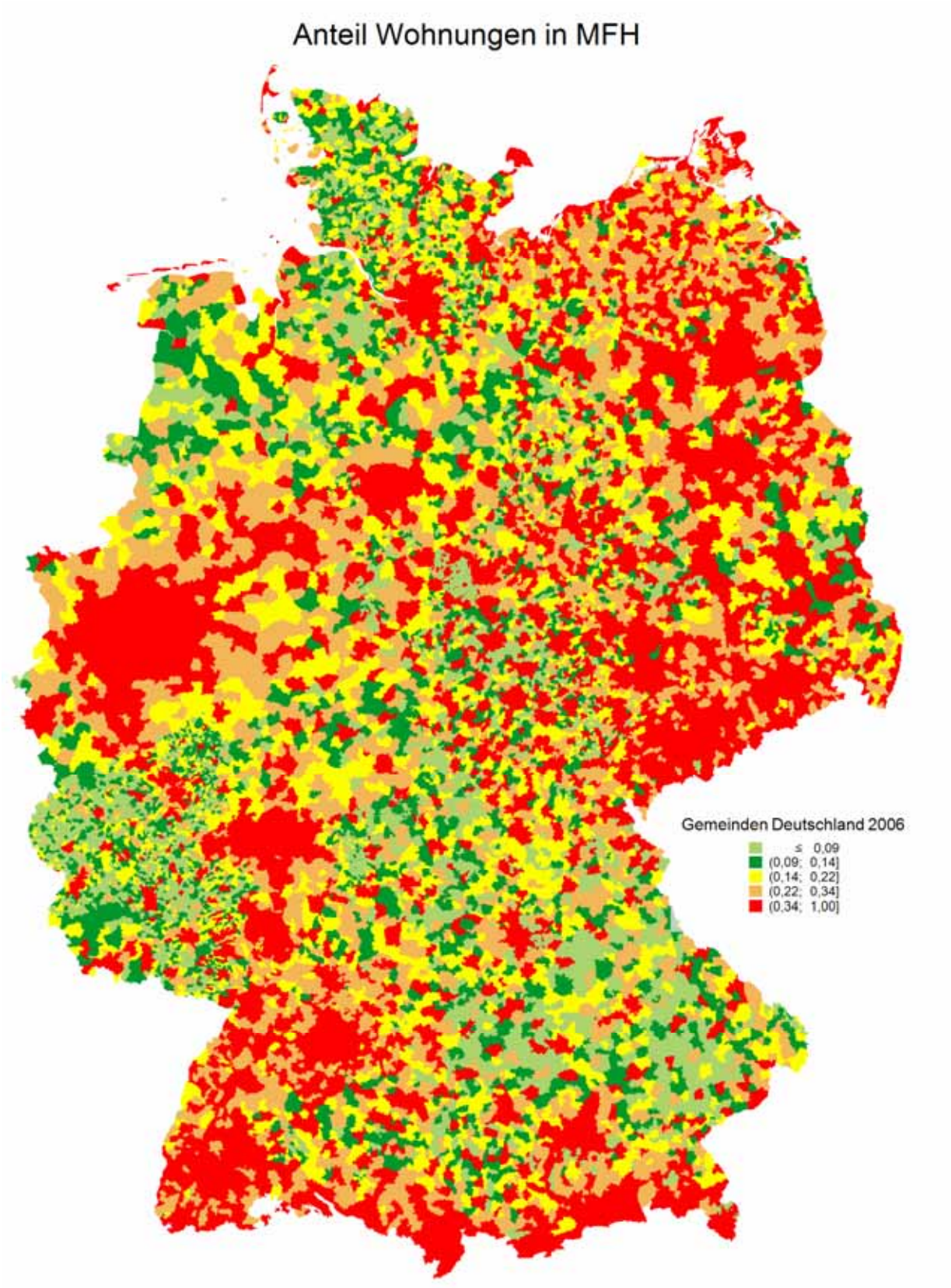


Abbildung 53: Anteil Wohnungen in Mehrfamilienhäusern (Stand: 31.12.2006)

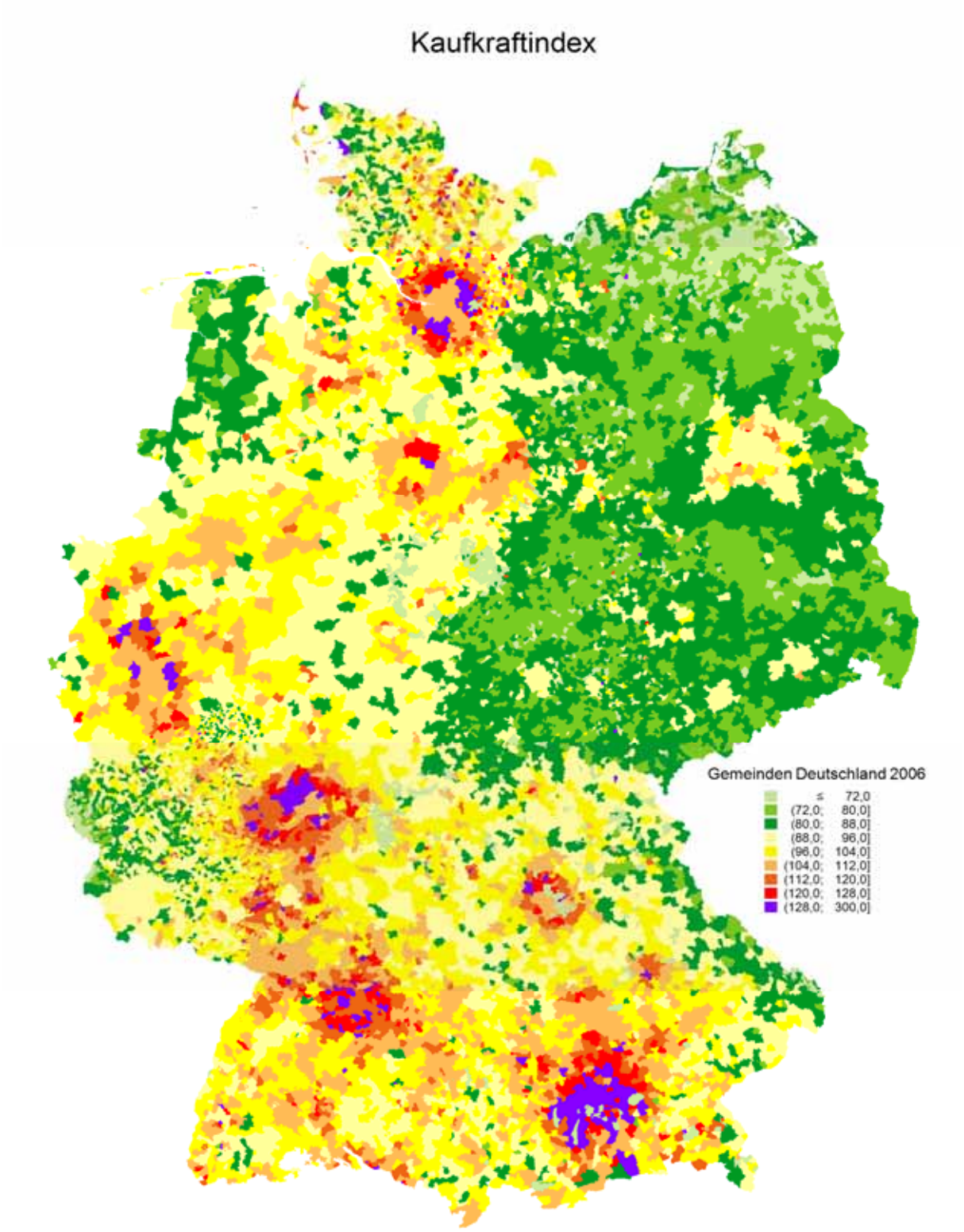


Abbildung 54: Kaufkraftindex[122] (Stand: 31.12.2006)

[122] Landesdurchschnitt = 100

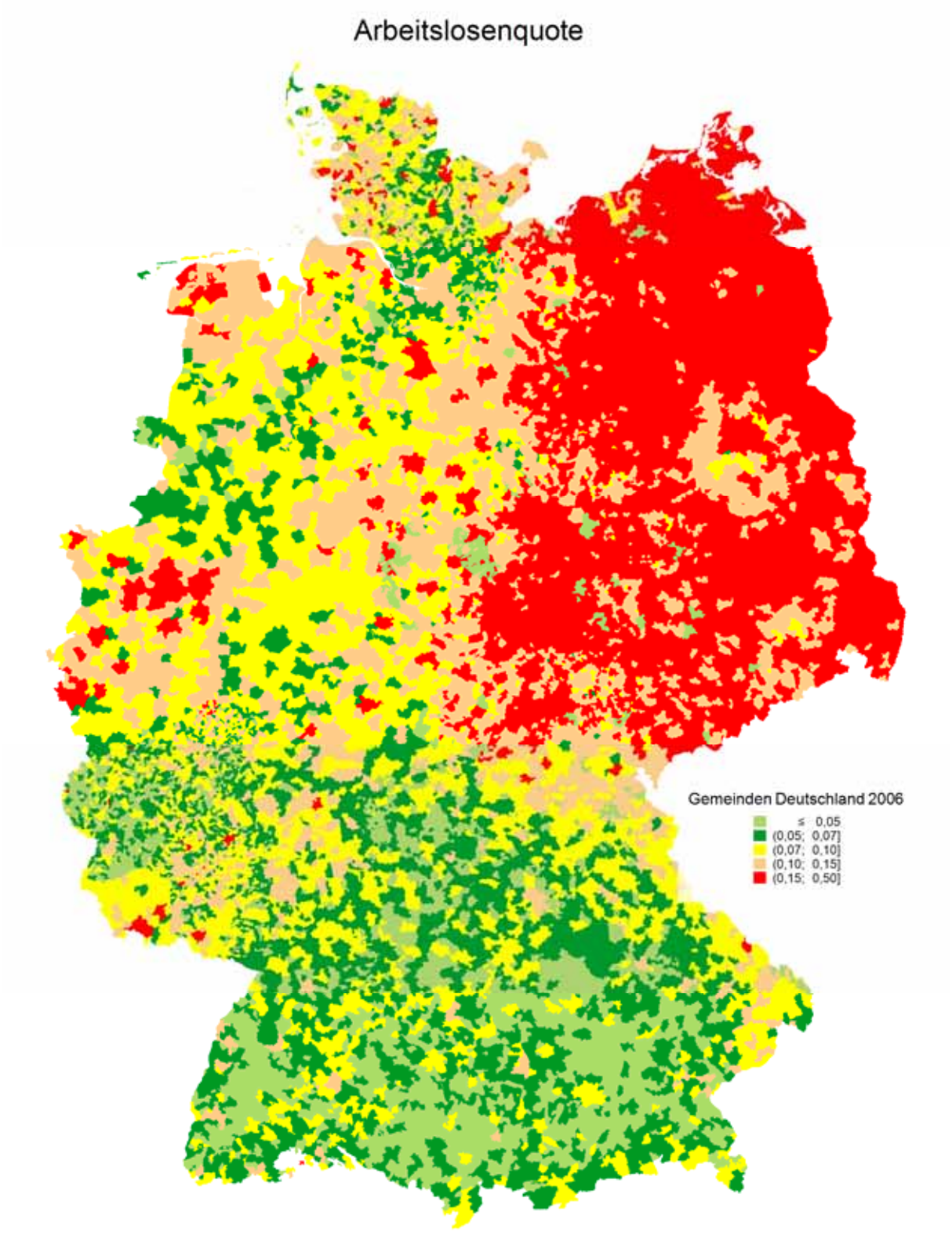


Abbildung 55: Arbeitslosenquote (Stand: 30.06.2006)

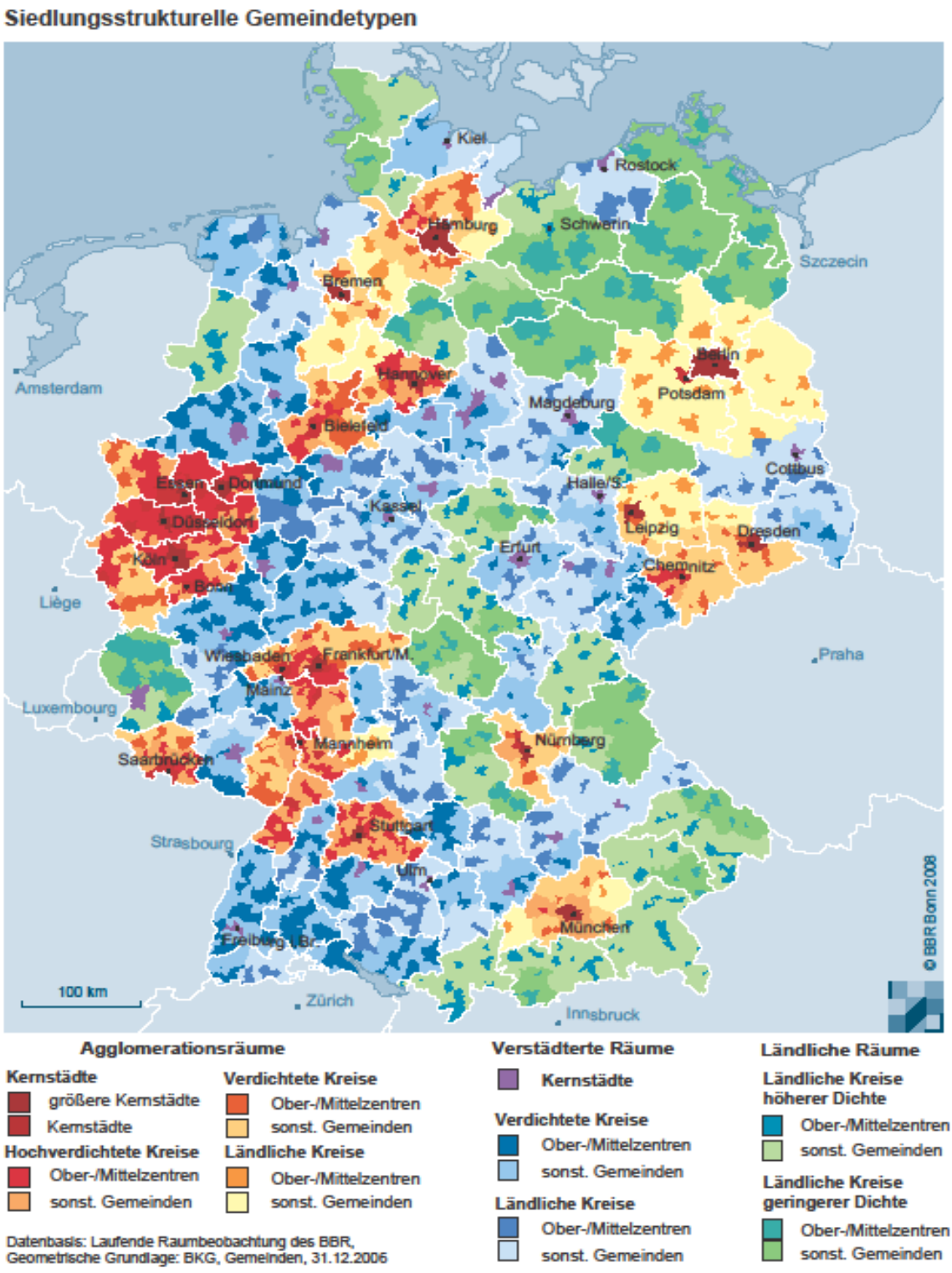


Abbildung 56: Siedlungsstrukturelle Gemeindetypen[123]

[123] Quelle: Karte Gemeindetypen 2006. Aus: [4]

Variablenliste 1: Datei Mietspiegel.sav

Name	*Label*
gkz	Gemeindekennziffer
gkz_regio	Gemeindekennziffer Regiograph
gemname	Gemeindename
gemtyp	Gemeindetyp
siedl_gemtyp	siedlungsstruktureller Gemeindetyp
nmqm_1Z	Nettomiete in Euro pro m² für 1-Z-Wohnung
nmqm_2Z	Nettomiete in Euro pro m² für 2-Z-Wohnung
nmqm_3Z	Nettomiete in Euro pro m² für 3-Z-Wohnung
nmqm_4Z	Nettomiete in Euro pro m² für 4-Z-Wohnung
nmqm_EFH	Nettomiete in Euro pro m² für Einfamilienhaus
nm_1Z	monatl. Nettomiete in Euro 1-Z-Wohnung
nm_2Z	monatl. Nettomiete in Euro 2-Z-Wohnung
nm_3Z	monatl. Nettomiete in Euro 3-Z-Wohnung
nm_4Z	monatl. Nettomiete in Euro 4-Z-Wohnung
nm_EFH	monatl. Nettomiete in Euro Einfamilienhaus
EW	Anzahl Einwohner zum 31.12.2006
HH	Anzahl Haushalte zum 31.12.2006
EW_HH	Einwohner pro Haushalt
Arblose_ins	Arbeitslose insgesamt zum 30.06.2006 lt. BA
SVB_ins_Wohn	Sozialverspflichtig Besch. am Wohnort zum 30.06.2006 lt. BA
GeB_ins_Wohn	Geringfügig Besch. am Wohnort zum 30.06.2006 lt. BA
SVB_pro_Einw	Sozialverspflichtig Besch. am Wohnort pro Einwohner
ALQ	Arbeitslosenquote
KKS	Kaufkraftsumme in Mio. €
KKI	Kaufkraftindex je EW
km30_SVBinsg_antlg	Soz.vers.pflichtig Besch. in 30km Umkreis

km30_EwGfK_antlg	Einwohnerzahl in 30km Umkreis
km30_KKSGfK_antlg	Kaufkraftsumme in 30km Umkreis
km30_HHGfK_antlg	Anzahl Haushalte in 30km Umkreis
SVB_EW_30km	Soz.vers.pfl. Besch. je EW in 30km Umkreis
qkm	Fläche in km²
Dichte	Bevölkerungsdichte
Geb_ins	Anzahl Gebäude
efh	Anzahl Einfamilienhäuser
zfh	Anzahl Zweifamilienhäuser
Anteil_efh	Anteil Einfamilienhäuser am Gebäudebestand
Anteil_zfh	Anteil Zweifamilienhäuser am Gebäudebestand
Whg_ins	Anzahl der Wohnungen insgesamt
Whg_in_efh	Anzahl Wohnungen in Einfamilienhäusern
Whg_in_zfh	Anzahl Wohnungen in Zweifamilienhäusern
Anteil_Whg_efh	Anteil Wohnungen in Einfamilienhäusern
Anteil_Whg_zfh	Anteil Wohnungen in Zweifamilienhäusern
Anteil_Whg_mfh	Anteil Wohnungen in Mehrfamilienhäusern
gtyp1	Kern- und Großstädte >= 500.000 E. in Agglomerationsräumen
gtyp2	Kern- und Großstädte unter 500.000 E. in Agglomerationsräumen
gtyp3	OZ/MZ in hochverd. Kreisen von Agglomerationsräumen
gtyp4	sonstige Gemeinden in hochverd. Kreisen von Agglomerations-räumen
gtyp5	OZ/MZ in verd. Kreisen von Agglomerationsräumen
gtyp6	sonstige Gemeinden in verd. Kreisen von Agglomerationsräumen
gtyp7	OZ/MZ in ländlichen Kreisen von Agglomerationsräumen
gtyp8	sonstige Gemeinden in ländlichen Kreisen von Agglomerations-räumen
gtyp9	Kern- und Großstädte in Verstädterten Räumen
gtyp10	OZ/MZ in verd. Kreisen von Verstädterten Räumen
gtyp11	sonstige Gemeinden in verd. Kreisen von Verstädterten Räumen

gtyp12	OZ/MZ in ländlichen Kreisen von Verstädterten Räumen
gtyp13	sonstige Gemeinden in ländlichen Kreisen von Verstädterten Räumen
gtyp14	OZ/MZ in Kreisen höherer Dichte von Ländlichen Räumen
gtyp15	sonstige Gemeinden in Kreisen höherer Dichte von Ländlichen Räumen
gtyp16	OZ/MZ in Kreisen geringerer Dichte von Ländlichen Räumen
gtyp17	sonstige Gemeinden in Kreisen geringerer Dichte von Ländlichen Räumen

Variablenliste 2: Datei IVD.sav

Name	*Label*
gkz_regio	Gemeindekennziffer Regiograph
gemname	Gemeindename
bundesland	Bundesland
gemtyp	Gemeindetyp
siedl_gemtyp	siedlungsstruktureller Gemeindetyp
nmqm48_eWW	Nettomiete in €/m²; Fertigstellung bis 1948: einfacher Wohnwert
nmqm48_mWW	Nettomiete in €/m²; Fertigstellung bis 1948: mittlerer Wohnwert
nmqm48_gWW	Nettomiete in €/m²; Fertigstellung bis 1948: guter Wohnwert
nmqm49_eWW	Nettomiete in €/m²; Fertigstellung ab 1949: einfacher Wohnwert
nmqm49_mWW	Nettomiete in €/m²; Fertigstellung ab 1949: mittlerer Wohnwert
nmqm49_gWW	Nettomiete in €/m²; Fertigstellung ab 1949: guter Wohnwert
nmqmnNeu_mWW	Nettomiete in €/m²; Neubau/Erstbezug mittlerer Wohnwert
nmqmNeu_gWW	Nettomiete in €/m²; Neubau/Erstbezug guter Wohnwert
nmqmtoplage	ca. Höchstmiete in Toplage
EW	Anzahl Einwohner zum 31.12.2006
HH	Anzahl Haushalte zum 31.12.2006
EW_HH	Einwohner pro Haushalt
Arblose_ins	Arbeitslose insgesamt zum 30.06.2006 lt. BA
SVB_ins_Wohn	Sozialverspflichtig Besch. am Wohnort zum 30.06.2006 lt. BA
GeB_ins_Wohn	Geringfügig Besch. am Wohnort zum 30.06.2006 lt. BA
SVB_pro_Einw	Sozialverspflichtig Besch. am Wohnort pro Einwohner
ALQ	Arbeitslosenquote
KKS	Kaufkraftsumme in Mio. €
KKI	Kaufkraftindex je EW

km30_SVBinsg_antlg	Soz.vers.pflichtig Besch. in 30km Umkreis
km30_EwGfK_antlg	Einwohnerzahl in 30km Umkreis
km30_KKSGfK_antlg	Kaufkraftsumme in 30km Umkreis
km30_HHGfK_antlg	Anzahl Haushalte in 30km Umkreis
SVB_EW_30km	Soz.vers.pfl. Besch. je EW in 30km Umkreis
qkm	Fläche in km²
Dichte	Bevölkerungsdichte
Geb_ins	Anzahl Gebäude
efh	Anzahl Einfamilienhäuser
zfh	Anzahl Zweifamilienhäuser
Anteil_efh	Anteil Einfamilienhäuser am Gebäudebestand
Anteil_zfh	Anteil Zweifamilienhäuser am Gebäudebestand
Whg_ins	Anzahl der Wohnungen insgesamt
Whg_in_efh	Anzahl Wohnungen in Einfamilienhäusern
Whg_in_zfh	Anzahl Wohnungen in Zweifamilienhäusern
Anteil_Whg_efh	Anteil Wohnungen in Einfamilienhäusern
Anteil_Whg_zfh	Anteil Wohnungen in Zweifamilienhäusern
Anteil_Whg_mfh	Anteil Wohnungen in Mehrfamilienhäusern
gtyp1	Kern- und Großstädte >= 500.000 E. in Agglomerationsräumen
gtyp2	Kern- und Großstädte unter 500.000 E. in Agglomerationsräumen
gtyp3	OZ/MZ in hochverd. Kreisen von Agglomerationsräumen
gtyp4	sonstige Gemeinden in hochverd. Kreisen von Agglomerationsräumen
gtyp5	OZ/MZ in verd. Kreisen von Agglomerationsräumen
gtyp6	sonstige Gemeinden in verd. Kreisen von Agglomerationsräumen
gtyp7	OZ/MZ in ländlichen Kreisen von Agglomerationsräumen
gtyp8	sonstige Gemeinden in ländlichen Kreisen von Agglomerationsräumen

gtyp9	Kern- und Großstädte in Verstädterten Räumen
gtyp10	OZ/MZ in verd. Kreisen von Verstädterten Räumen
gtyp11	sonstige Gemeinden in verd. Kreisen von Verstädterten Räumen
gtyp12	OZ/MZ in ländlichen Kreisen von Verstädterten Räumen
gtyp13	sonstige Gemeinden in ländlichen Kreisen von Verstädterten Räumen
gtyp14	OZ/MZ in Kreisen höherer Dichte von Ländlichen Räumen
gtyp15	sonstige Gemeinden in Kreisen höherer Dichte von Ländlichen Räumen
gtyp16	OZ/MZ in Kreisen geringerer Dichte von Ländlichen Räumen
gtyp17	sonstige Gemeinden in Kreisen geringerer Dichte von Ländlichen Räumen
Gast_EW	Übernachtungsgäste pro Einwohner im Jahr 2006