Projektarbeit 2008

Zürcher Hochschule für Angewandte Wissenschaften

Studiengang Datenanalyse und Prozessdesign

Prognose der Schülerzahlen in der Stadt Zürich

Verfasserin

Sina Rüeger
Lindbergstrasse 21
8404 Winterthur
ruegesin@students.zhaw.ch

Betreuer

Prof., Dr. sc. Math. ETH Andreas Ruckstuhl
Zürcher Hochschule für Angewandte
Wissenschaften, Winterthur

Projektarbeit: 18. Februar bis 23. Mai 2008

Inhaltsverzeichnis

Zusammenfassung

Zusammenfassung auf deutsch

Abstract

The abstract abstract.

1. Einleitung

Für die Planung der Schulräume in der Stadt Zürich müssen die Schülerzahlen[1] für die kommenden Jahre geschätzt werden. Diese Aufgabe fällt in der Stadt Zürich dem Schul- und Sportdepartement zu, welche mit der bisherigen Methode stets zu hohe Schülerzahlen prognostiziert hat. Um exaktere Schülerzahlprognosen zu erhalten, wurde Statistik Stadt Zürich angefragt, die bisherige Prognosemethode zu überprüfen und bei Bedarf eine neue zu Methode zu entwickeln.

1.1. Rahmenbedingungen

Das von Statistik Stadt Zürich lancierte Projekt ist in verschiedene Etappen gegliedert worden.

1. Der erste Schritt realisiert eine Kurz- bis Mittelfristprognose von 4 Jahren. Unterteilt wird das Modell in Schulklassen und geografisch in Schulkreise.
2. In weiteren Schritten entsteht eine Langfristprognose, unterteilt in geografisch kleinere Einheiten, und die Berechnung der Anzahl Schulklassen.

Die Schulpflicht beträgt in der Stadt Zürich ab dem Schuljahr 2008/09 elf Jahre. Das Modell soll die zukünftige Entwicklung der Anzahl Schüler in den verschiedenen Schulkreisen der Stadt Zürich abbilden. Dabei besuchen die Schüler den Kindergarten oder eine Regelklasse (1. Klasse bis 9. Klasse). Der Kindergarten nimmt eine besondere Stellung ein, weil dieser im Schuljahr 2008/09 obligatorisch wird[2].
Zur Verfügung stehen effektive Schülerzahlen von vergangenen Jahren sowie Bevölkerungsregisterdaten der Stadt Zürich. Die Bautätigkeit der Stadt Zürich wird ebenfalls erfasst. Im statistischen Modell agieren Schülerzahlen von vergangenen Jahren als erklärende Grösse und als Zielgrösse Schülerzahlen, die in der Zukunft liegen. Zusätzliche erklärende Grössen sind die Einteilung in die obligatorische Schulpflicht und in Schulkreise, weil davon ausgegangen wird, dass nicht jeder Schulkreis und nicht jede Regelklasse nach denselben Mustern funktioniert.

1.2. Ziel der Projektarbeit

Das Ziel dieser Projektarbeit besteht darin, ein statistisches Modell für eine Kurz- bis Mittelfristprognose von 4 Jahren zu formulieren, die Schülerzahlen pro Schulkreis und Schulstufe

[1] Innerhalb dieser Projektarbeit wird für *Schülerinnen* einheitlich die männliche Form *Schüler* benutzt

[2] Ab dem Schuljahr 2008/09 ist der zweijährige Kindergartenbesuch für alle Kinder obligatorisch, die Schulpflicht erhöht sich von neun auf elf Jahre. Dadurch kann sich in einigen Schulkreisen die Anzahl Kindergartenkinder erhöhen.

zu prognostizieren und damit eine bessere Abschätzung als das Schul- und Sportdepartement zu erreichen - bestenfalls mit geringerem Rechenaufwand und weniger erklärenden Grössen.

2. Daten

2.1. Datenaufbereitung

Als Datenquelle dienten bisher erfasste Schülerzahlen des Schul- und Sportdepartementes. Zur Verfügung standen weiter das Bevölkerungsregister der Stadt Zürich und die Daten über die Bautätigkeit in der Stadt Zürich. Abbildung 2.1 illustriert das Vorgehen und die Quellen bei der Datenaufbereitung. Die Daten sind in der ersten Hälfte der Projektarbeit zusammengefügt worden.

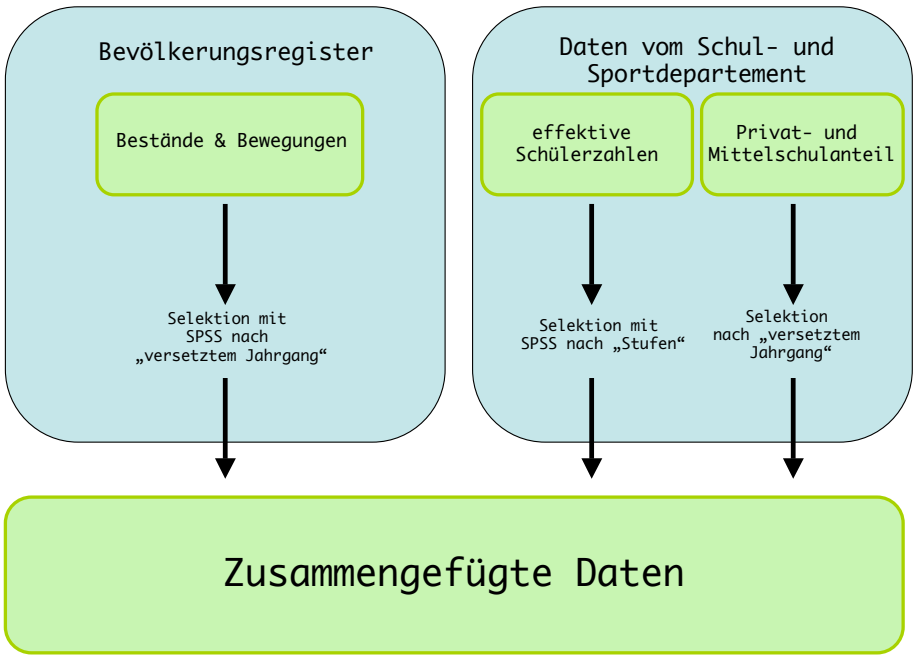

Abbildung 2.1.: Vorgehen bei der Datenaufbereitung

2.1.1. Daten Schul- und Sportdepartement

Die Daten des Schul- und Sportdepartements waren in Excel-Dateien gespeichert und in *effektive Schülerzahlen* und *Privat- und Mittelschulanteil* unterteilt. Die *effektive Schülerzahlen* waren pro Klasse und Schulkreis (resp. pro Kindergarten und Schulkreis) vorhanden. Im Gegensatz zu *Privat- und Mittelschulanteil*, wo die Anteile pro *versetztem* Jahrgang berechnet wurden. Relevant für die Projektarbeit sind die Daten der Jahre 2003 bis 2007. Die Jahrgänge kleiner wie 2003 enthielten Daten mit Codierungen, deren Aufbereitung unverhältnismässig viel Zeit in Anspruch genommen hätte.

2.1.2. Daten Statistik Stadt Zürich

Bestände und Bewegungen aus dem Bevölkerungsregister sind verwendet worden, um die Bevölkerung ausserhalb vom Schulrahmen abzubilden. Die Aggregation zu *versetzten* Jahrgängen dient dazu, die Bevölkerung vor der Einschulung beziffern zu können. Dabei bezeichnen die *versetzten* Jahrgänge nicht die herkömmlichen Jahrgänge, wie beispielsweise Jahrgang 2003, sondern - wie bei der Einschulung üblich - ein Jahrgang, der am 1. Mai des Jahres 2003 beginnt und am 30. April des Jahres 2004 endet[1]. Tabelle 2.1 zeigt ein Beispiel vom Schuljahr 2003/04. Im Anhang B.1.1 finden sich die *versetzten* Jahrgänge für alle in der Projektarbeit verwendeten Jahre.

***versetzter* Jahrgang**	**Datum von**	**Datum bis**
Jahrgang 0	1. Mai 2003	31.Dezember 2003
Jahrgang 1	1. Mai 2002	30. April 2003
Jahrgang 2	1. Mai 2001	30. April 2002
Jahrgang 3	1. Mai 2000	30. April 2001
Jahrgang 4	1. Mai 1999	30. April 2000

Tabelle 2.1.: Beispiel des *versetzten* Jahrgangs für das Schuljahr 2003/04

2.1.3. Zusammenführung der Daten

Zum Schluss sind die aufbereiteten Datensätze des Schul- und Sportdepartementes und des Bevölkerungsregisters zusammengefügt worden. So entstand eine Rohdatei mit 160 Zeilen und 157 Spalten.

Die detaillierten Vorgänge bei der Datenaufbereitung können dem Anhang B.1 entnommen werden.

[1] In den nächsten sechs Jahren soll das Schuleintrittsalter um drei Monate gesenkt werden. Heute gilt als Stichtag der 30. April. Kinder, die bis zu diesem Tag vier Jahre alt wurden, werden nach den Sommerferien jeweils eingeschult und kommen in den Kindergarten. Dieser Stichtag soll nun auf den 31. Juli verschoben werden. Die Änderung soll schrittweise erfolgen, um den erwarteten Schülerzuwachs auf mehrere Jahre zu verteilen. In den nächsten sechs Jahren soll der Stichtag jährlich um einen halben Monat nach hinten verschoben werden.

2.2. Bedeutung der Variablen

2.2.1. Schulkreise - $SK^{(i)}$

Die Stadt Zürich gliedert sich geografisch in Schulkreise i.

i	**Schulkreis**
1	`Glattal`
2	`Letzi`
3	`Limmattal`
4	`Schwamendingen`
5	`Uto`
6	`Waidberg`
7	`Zürichberg`

Tabelle 2.2.: Variable Schulkreis i

Die Schulkreise wiederum sind in mehrere Quartiere gegliedert, diese spalten sich in Schuleinheiten auf. Eine Schuleinheit kann ein oder mehrere Schulhäuser umfassen.

2.2.2. Stufe $ST^{(k)}$

Innerhalb dieser Arbeit werden der Kindergarten, die Klassen und Vorschuljahrgänge einheitlich mit Stufe k bezeichnet. Diese Normierung wird benötigt, damit für das Modell eine Variable geschaffen werden kann, die einheitliche Levels beinhaltet [2]. Die Anzahl Personen, die noch nicht eingeschult sind, werden für die Prognose benötigt. Schüler, die im Jahr `2007` die `Stufe 3` (1. Klasse) besuchen, waren im Jahr `2003` noch nicht eingeschult mit `Stufe -1` (Jahrgang 3), siehe auch Kapitel 2.1.2.

[2] Für die Prognose müssen die Levels der Indikatorvariable (Kapitel 3.1) Stufe geändert werden. Diese Leveländerung bringt einige Komplikationen mit sich, die eine Eigenheit von Indikatorvariablen ist. Deshalb wurde die Indikatorvariable Stufe mit numerischen Werten versehen, die einfacher geändert werden können, wie Character

Stufe k	**Klasse**	**Schuletappe**
`Stufe 1`	1. Kindergarten	Kindergarten
`Stufe 2`	2. Kindergarten	Kindergarten
`Stufe 3`	1. Klasse	Unterstufe
`Stufe 4`	2. Klasse	Unterstufe
`Stufe 5`	3. Klasse	Unterstufe
`Stufe 6`	4. Klasse	Mittelstufe
`Stufe 7`	5. Klasse	Mittelstufe
`Stufe 8`	6. Klasse	Mittelstufe
`Stufe 9`	7. Klasse	Oberstufe
`Stufe 10`	8. Klasse	Oberstufe
`Stufe 11`	9. Klasse	Oberstufe

Tabelle 2.3.: Variable Stufe $i, i = 1, \ldots, 11$, eingeteilt in Schuletappen

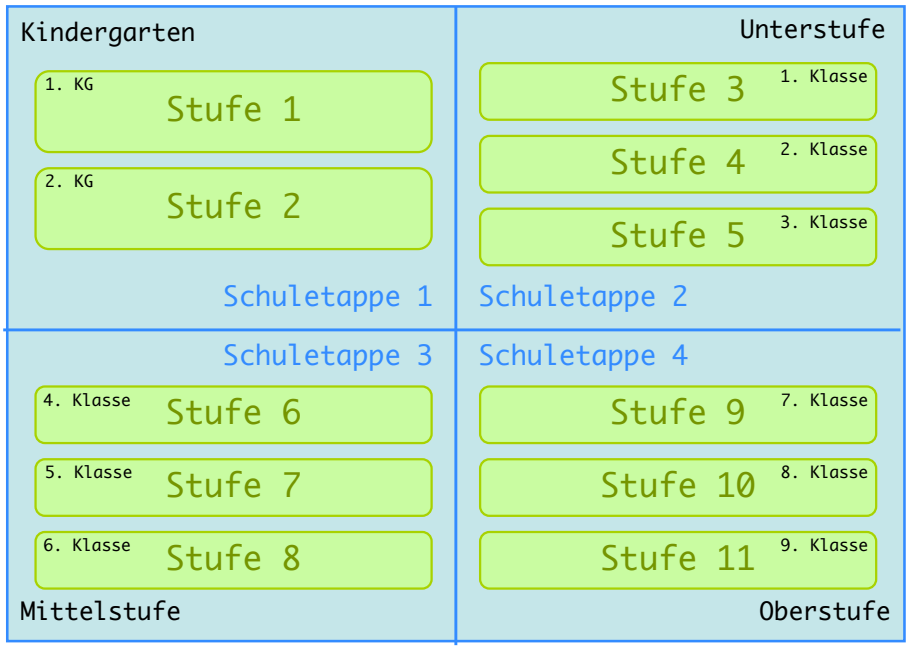

Abbildung 2.2.: Aufteilung der Schuletappen und Stufen

Stufe k	***versetzter* Jahrgang**
`Stufe -4`	Jahrgang 0
`Stufe -3`	Jahrgang 1
`Stufe -2`	Jahrgang 2
`Stufe -1`	Jahrgang 3
`Stufe  0`	Jahrgang 4

Tabelle 2.4.: Variable Schulkreis $i, i = -4, \ldots, 0$

2.2.3. Schuletappe θ

Eine Schuletappe umfasst 2 oder 3 Stufen und gliedert sich in Kindergarten, Unterstufe, Mittelstufe und Oberstufe. Die Abgrenzung der Stufen in Schuletappen wird für die Modelldifferenzierung benötigt.

2.2.4. Zeit

Die Zeit ist keine Variable im Modell. Sie spielt zur Beschreibung der Schülerzahl Y_{ikt} eine Rolle.
Als Zeit kommen die Schuljahre 2003/04, 2004/05, 2005/06, 2006/07, 2007/08 vor, welche zur Vereinfachung als `2003, 2004, 2005, 2006, 2007` bezeichnet werden.

Zeit t	**Schuljahr**
2003	2003/04
2004	2004/05
2005	2005/06
2006	2006/07
2007	2007/08

Tabelle 2.5.: Variable Zeit

2.2.5. Schülerzahlen Y_{ikt}

Der Schulkreis i, die Stufe k und die Zeit t beschreiben eine Schülerzahl, $j = 1, ..., n$. Die Zeit t verändert sich je nachdem, welcher Wert Δ_t annimmt (Tabelle 2.6).

2.2.6. Schülerzahlen $Y_{i(k-\Delta_t)(t-\Delta_t)}$

Der Schulkreis i, die Stufe $(k - \Delta_t)$, die Zeit $(t - \Delta_t)$ beschreiben eine Schülerzahl. Die Zeit t verändert sich je nachdem, welcher Wert Δ_t annimmt (Tabelle 2.6).

Δ_t	**t**
1	2004,2005,2006,2007
2	2005,2006,2007
3	2006, 2007
4	2007

Tabelle 2.6.: Konstellation der Indizes t, Δ_t

2.3. Beschreibende Statistik der Daten

Wird die Variable $Y_{i(k-\Delta_1)(t-\Delta_1)}$ auf der x-Achse und auf der y-Achse Y_{ikt} aufgetragen, streuen die Punkte einem Band entlang. Idealerweise wäre das Band eine Gerade mit 45° Winkel, dann würden sich die Schülerzahlen in der gesamten Stadt Zürich gleich verhalten: im Jahr t und in der Stufe k hätte es in jedem Schulkreis genau gleich viel Schüler wie im Jahr $(t-\Delta_1)$ und in der Stufe $(k-\Delta_1)$.
Die Annahme, dass Schülerzahlen sich in jedem Schulkreis und in jeder Stufe verschieden verhalten, verdeutlicht sich in der Abbildung 2.3. Die Daten sind zusätzlich nach Schuletappen aufgeteilt. Durch diese Aufteilung wird ersichtlich, dass die Punkte eine unterschiedliche Bandbreite in der Streuung aufweisen. Die Mittelstufe ist im Vergleich mit der Oberstufe viel "kompakter". Weitere Abbildungen mit $t = 2004, 2005, 2006, 2007$ und $\Delta_t = 1, 2, 3, 4$ finden sich in Anhang B.3.1.

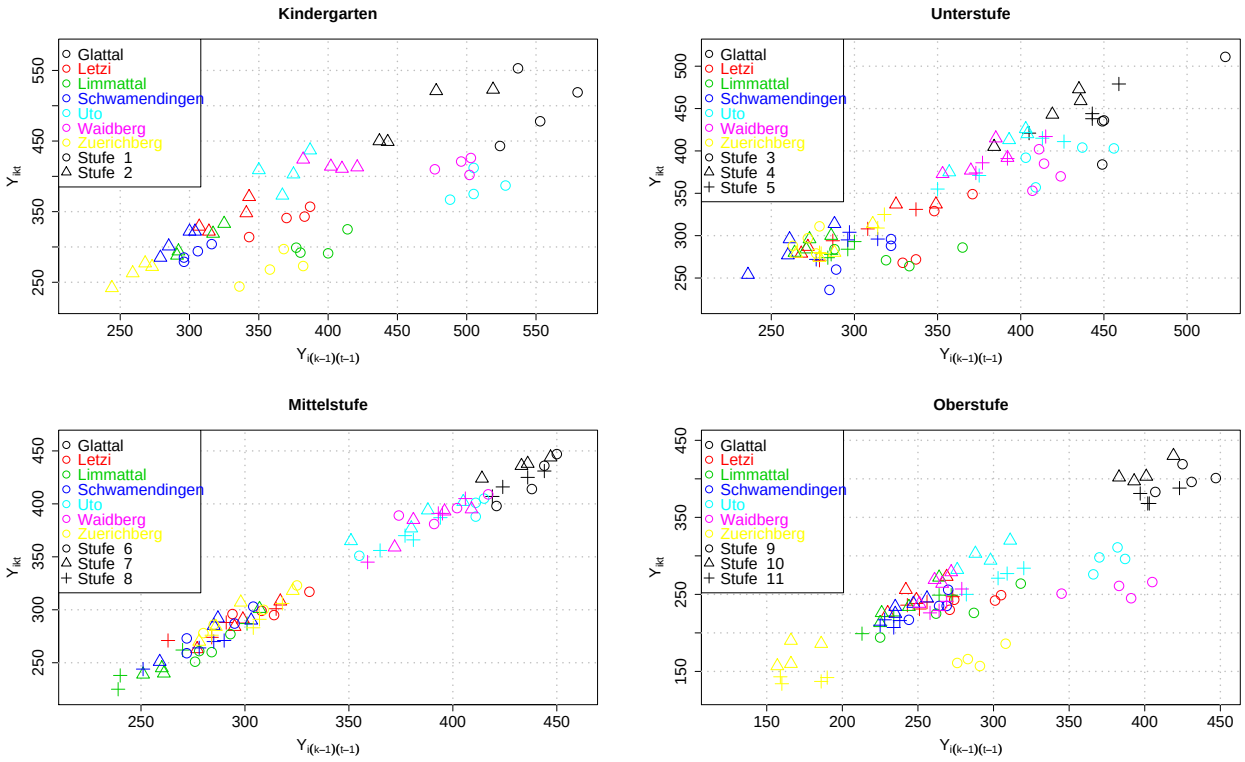

Abbildung 2.3.: Schülerzahlen der Stadt Zürich aufgeteilt in 4 Gruppen: auf der x-Achse die Variable $Y_{i(k-1)(t-1)}$, auf der y-Achse ist die Variable Y_{ikt} aufgetragen.

3. Statistische Methoden und Ergebnisse

3.1. Statistischer Hintergrund

Um die Verständlichkeit von Indikatorvariablen und Wechselwirkungen zu verbessern, sind in diesem Kapitel die Hintergründe dazu verfasst. Der Aufbau von statistischen Tests für die Modellwahl findet sich ebenfalls in diesem Kapitel.

Multiple lineare Regression

In der Regressionsrechnung wird der Zusammenhang zwischen erklärenden Grössen und der Zielgrösse untersucht. Das dazugehörige Modell lautet:

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + ... + \beta_m x_i^{(m)} + E_i \tag{3.1}$$

Das bedeutet: die i-te Beobachtung der Zielvariable Y_i wird durch die i-te Beobachtungen der k-ten Grösse der erklärenden Variablen $x_i^{(k)}$ beschrieben. $\beta_0, \beta_1, ..., \beta_m$ sind die unbekannten Parameter. E_i bezeichnet die zufälligen Abweichungen, bei denen die Annahme ist, das sie (näherungsweise) normalverteilt sind (Ruckstuhl, (2007)).

Indikatorvariable

Eine erklärende Variable $x_j^{(k)}$ aus der Formel 3.1 kann metrische Grössen enthalten - beispielsweise Gewicht eines Apfels oder Anzahl Schüler. Möglich sind aber auch kategorielle Grössen - zum Beispiel Sorte eines Apfels oder Schulkreise der Stadt Zürich. In diesem Zusammenhang spricht man auch von **Levels**. Wenn die Grösse von Äpfel untersucht werden, die von 20 Sorten stammen, hat die Indikatorvariable 20 Levels. 7 Schulkreise gibt es in der Stadt Zürich; die Indikatorvariable enthält 7 Levels. Eine Schülerzahl oder ein Apfel muss eindeutig einem Level zugeordnet werden. Dies geschieht mittels binären Grössen mit den Werten $\{0, 1\}$ wie in Formel 3.2 oder anders gesagt, mit einem on/off Schalter.

Die Indikatorvariable $F_j^{(k)}$ hat m Levels, $k = 1, 2, ..., m$.

$$F_j^{(k)} = \begin{cases} 1 & \text{falls j-te Beobachtung aus } k\text{-tem Level} \\ 0 & \text{sonst} \end{cases} \tag{3.2}$$

Wechselwirkungen

Wenn für die Beschreibung der Zielvariablen nicht nur die erklärenden Variablen alleine nötig sind, sondern auch Kombinationen untereinander, spricht man von Wechselwirkungen. $x_i^{(k)}$ und $x_i^{(k+1)}$ multiplizieren sich und bilden so eine neue erklärende Grösse.

Beispiel

Ein Beispiel am Thema Schülerzahlen soll Wechselwirkungen so anschaulich wie möglich erklären:

$$Y_j = \beta_0 + \beta_1 x_j^{(1)} + \gamma_1 ST_j^{(9)} + \gamma_2 ST_j^{(10)} + \varphi_1 SK_j^{(A)} + \varphi_2 SK_j^{(B)} + E_j \tag{3.3}$$

$\gamma_1 = \varphi_1 = 0$
In der Formel 3.3 sind $x_j^{(2)}$ und $x_j^{(3)}$ von der Formel 3.1 durch die Indikatorvariablen ST_j (Stufe) und SK_j(Schulkreis) ersetzt worden. Y_j sind die Anzahl Schüler, $x_j^{(1)}$ bezeichnet die Schülerzahlen im vorhergegangenen Jahr.

Die Indikatorvariable $ST_j^{(k)}$ hat 2 Levels $\{9, 10\}$.

$$ST_j^{(k)} = \begin{cases} 1 & \text{falls j-te Beobachtung aus } k\text{-tem Level} \\ 0 & \text{sonst} \end{cases} \tag{3.4}$$

Die Indikatorvariable $SK_j^{(i)}$ hat 2 Levels $\{A, B\}$.

$$SK_j^{(i)} = \begin{cases} 1 & \text{falls j-te Beobachtung aus } i\text{-tem Level} \\ 0 & \text{sonst} \end{cases} \tag{3.5}$$

γ_1 und φ_1 werden gleich null gesetzt, damit die Parameter für die Variablen $ST_j^{(k)}$ und $SK_j^{(i)}$ eindeutig identifizierbar (eindeutig schätzbar sind), weiterführend dazu Müller, (2006).

Die Indikatorvariablen $ST_J^{(k)}$ und $SK_j^{(i)}$ enthalten je 2 Levels. Dies ergibt 4 Kombinationsmöglichkeiten zwischen zwischen $ST_J^{(k)}$ und $SK_j^{(i)}$. Eine Zusammenstellung der 4 Fälle kann der Tabelle 3.1 entnommen werden.

ST	*SK*		
9	A	$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \gamma_1 ST_j^{(9)} + \varphi_1 SK_j^{(A)}$	$\gamma_1 = \varphi_1 = 0$
9	B	$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \gamma_1 ST_j^{(9)} + \varphi_2 SK_j^{(B)}$	$\gamma_1 = 0$
10	A	$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \gamma_2 ST_j^{(10)} + \varphi_1 SK_j^{(A)}$	$\varphi_1 = 0$
10	B	$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \gamma_2 ST_j^{(10)} + \varphi_2 SK_j^{(B)}$	

Tabelle 3.1.: 4 Möglichkeiten, die Levels der Indikatorvariablen des Modells 3.3 zu kombinieren

Es gibt stets dieselbe Steigung β_1. Was sich von Fall zu Fall verändert, ist der Achsenabschnitt. Im ersten Fall ist er β_0 (da $\gamma_1 = \varphi_1 = 0$). Wenn die Stufe beibehalten wird und der Schulkreis von A zu B wechselt, erhöht sich der Achsenabschnitt um φ_2. Gleiches geschieht, wenn der Schulkreis beibehalten wird und die Stufe von 9 nach 10 wechselt. Dann erhöht sich der Achsenabschnitt um γ_2. Dies ist Abbildung 3.1 ersichtlich.
Was bedeutet das in Bezug auf die Schülerzahlen? Die **Differenz** zwischen der Schülerzahl im Schulkreis i in der Stufe k und der Schülerzahl im Schulkreis i in der Stufe $(k + \Delta)$ ist gleich wie die Differenz zwischen der Schülerzahl im Schulkreis $(i + 1)$ in der Stufe k und der Schülerzahl im Schulkreis $(i + 1)$ in der Stufe $(k + \Delta)$. **Differenz** $=\gamma_2$
Oder wenn die andere Kombination gewählt wird: Die **Differenz** zwischen der Schülerzahl im Schulkreis i in der Stufe k und der Schülerzahl im Schulkreis $(i + 1)$ in der Stufe k ist gleich wie die **Differenz** zwischen der Schülerzahl im Schulkreis i in der Stufe $(k + \Delta)$ und der Schülerzahl im Schulkreis $(i + 1)$ in der Stufe $(k + \Delta)$. **Differenz** $=\varphi_2$

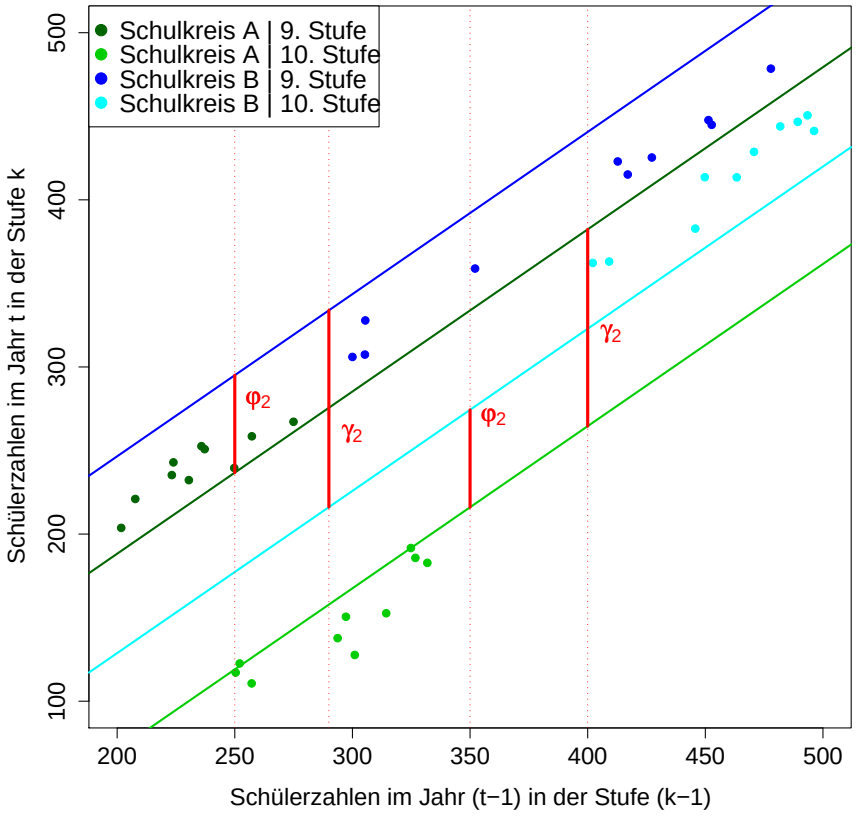

Abbildung 3.1.: aufgezeichnete Regressionsgeraden des Modells 3.3; analog zur Tabelle 3.1 sind 4 parallele Geraden aufgezeichnet

Nicht in jedem Schulkreis und nicht in jeder Stufe ist der Anteil der Schüler, die im Sommer

in die nächst höhere Klasse wechseln, gleich gross. Der Übergang von der 6. Klasse in das Gymnasium ist ein anschauliches Beispiel. Im Schulkreis Zürichberg gehen Ende der sechsten Klasse anteilsmässig mehr Schüler ins Gymnasium wie im Schulkreis Glattal (Anhang B.19). Die Differenz zwischen der 7. Klasse im Schulkreis Zürichberg und jener im Schulkreis Glatttal hat sich gegenüber der Differenz bei der 6. Klasse gegenüber erhöht. Deshalb muss versucht werden, diese Situation abzubilden. **Wechselwirkungen** führen neue Variablen ein die eine Kombination zwischen den Variablen ermöglichen. Die Geraden verlaufen weiterhin parallel. Wechselwirkungen erlauben aber, dass die Abstände zwischen den Geraden unterschiedlich sein können. Dies ist in Abbildung 3.2 ersichtlich.

Wenn in der Formel 3.3 Wechselwirkungen zugelassen werden, entsteht Formel 3.6.

$$Y_j = \beta_0 + \beta_1 x_j^{(1)} + \gamma_1 ST_j^{(9)} + \gamma_2 ST_j^{(10)} + \varphi_1 SK_j^{(A)} + \varphi_2 SK_j^{(B)} + \nu_1 ST_j^{(9)} SK_j^{(A)}$$
$$+\nu_2 ST_j^{(9)} \cdot SK_j^{(B)} + \nu_3 ST_j^{(10)} SK_j^{(A)} + \nu_4 ST_j^{(10)} SK_j^{(B)} + E_j \quad (3.6)$$

$$\gamma_1 = \varphi_1 = \nu_1 = \nu_2 = \nu_3 = 0$$

$\gamma_1, \varphi_1, \nu_1, \nu_2$ und ν_3 werden gleich null gesetzt, damit die Parameter für die Variablen $ST_j^{(k)}$, $SK_j^{(i)}$ und $ST_j^{(k)} \cdot SK_j^{(i)}$ eindeutig identifizierbar (eindeutig schätzbar sind), weiterführend dazu Müller, (2006).

Eine Zusammenstellung der 4 Fälle kann der Tabelle 3.2 entnommen werden.

ST	*SK*		
9	A	$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \gamma_1 ST_j^{(9)} + \varphi_1 SK_j^{(A)} + \nu_1 SK_j^{(9)} \cdot SK_j^{(A)}$	$\gamma_1 = \varphi_1 = \nu 1 = 0$
9	B	$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \gamma_1 ST_j^{(9)} + \varphi_2 SK_j^{(B)} + \nu_2 ST_j^{(9)} \cdot SK_j^{(B)}$	$\gamma_1 = \nu_2 = 0$
10	A	$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \gamma_2 ST_j^{(10)} + \varphi_1 SK_j^{(A)} + \nu_3 ST_j^{(10)} \cdot SK_j^{(A)}$	$\varphi_1 = \nu_3 = 0$
10	B	$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \gamma_2 ST_j^{(10)} + \varphi_2 SK_j^{(B)} + \nu_4 ST_j^{(10)} \cdot SK_j^{(B)}$	

Tabelle 3.2.: 4 Möglichkeiten, die Levels der Indikatorvariablen des Modells 3.6 zu kombinieren

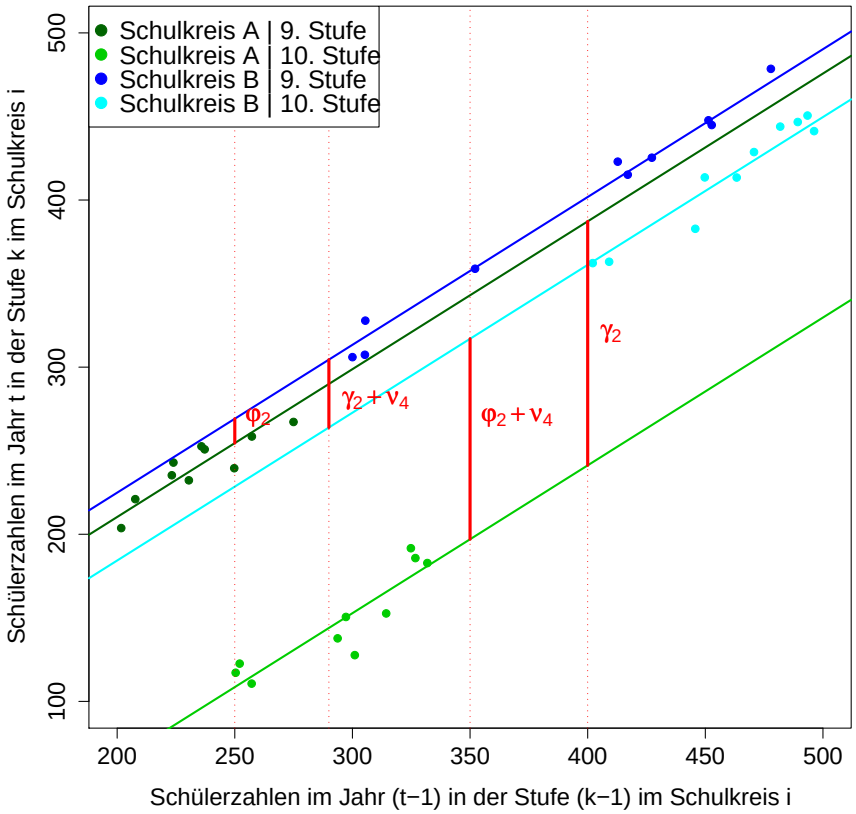

Abbildung 3.2.: aufgezeichnete Regressionsgeraden des Modells 3.6 mit Wechselwirkungen; analog zur Tabelle 3.2 sind 4 parallele Geraden aufgezeichnet

Devianz-Teststatistik

Um ein Modell mit p Parametern und ein Modell mit q Parametern zu vergleichen - wobei $p > q$ - kann eine Teststatistik die auf Residuen-Devianzen basiert gemacht werden. $D\langle y; \widehat{\beta}^r\rangle$ ist die Residuen-Devianz des kleineren Modells, $\widehat{\phi}$ ist die Schätzung des Dispersionsparameters, im Fall der Poissonverteilung also 1.
H_0: Das grössere Modell ergibt keine signifikante Verbesserung gegenüber dem kleineren Modell

$$T_D = \frac{D\langle y; \widehat{\beta}^r\rangle - D\langle y; \widehat{\beta}\rangle}{\widehat{\phi}},\ T_D \overset{as}{\sim} \chi^2_{p-q} \tag{3.7}$$

$T_D \overset{as}{\sim} \chi^2_{p-q}$, wenn das kleinere Modell stimmt (Ruckstuhl (2007)).

Goodness-of-fit

$$T = \sum_{k=1}^{n} \frac{(Y_i - \widehat{\mu})^2}{\widehat{\mu}} = \sum_{k=1}^{n} (R_i^{(P)})^2,\ T \sim \chi_n^2 \tag{3.8}$$

Als Y_i werden Out-of-sample-Daten verwendet. H_0: Das Modell beschreibt die zukünftigen Werte
H_0 ablehnen, falls $T > \chi^2_{0.95,n}$

Prognoseintervalle

Für poissonverteilte Zielvariablen ist das $100(1-\alpha)$ Prognoseintervall

$$\widehat{\mu} \pm 2\sqrt{\widehat{\mu} + se^2_{\widehat{\mu}}} \tag{3.9}$$

Der Parameter für poissonverteilte Grössen lautet λ. Der Erwartungswert der Zielvariable ist gleich dem Parameter und identisch mit der Varianz der Zielvariable ($E\langle Y\rangle = Var\langle Y\rangle = \lambda$). Deshalb lässt sich das Prognoseintervall wie in Formel 3.9 schreiben.

3.2. Modellwahl-Verfahren

Die Idee ist die Zielgrösse Y_{ikt} mit der erklärenden Grösse $Y_{i(k-\Delta_t)(t-\Delta_t)}$ zu beschreiben, abstrakt veranschaulicht das Abbildung 3.3. Als Indikatorvariablen stehen $ST_j^{(k)}$ und $SK_j^{(i)}$ zur Verfügung. Weil Schülerzahlen von 5 Jahre vorhanden sind, werden 4 Prognosehorizonte ermöglicht. Grundsätzlich kann zwischen zwei Verfahren unterschieden werden: **Verfahren A**, das mit einem Modell ohne Wechselwirkungen startet und eine Variablenselektion durchläuft. Dieses Verfahren wird in Anhang B.2 beschrieben. **Verfahren B** startet mit einem Modell mit Wechselwirkungen und durchläuft ebenfalls eine Variablenselektion, dies wird in Kapitel 3.2.1 beschrieben. Auf Grund der Devianz-Teststatistik (und allenfalls der Residuenanalyse) wird eine Modellwahl getroffen. Zusätzlich wird eine Out-of-sample-Prognose gemacht, der Goodness-of-fit gemessen und zwischen den beiden Verfahren verglichen. Im Idealfall bestätigen die Schlussfolgerungen der Out-of-sample-Prognose die Modellwahl.

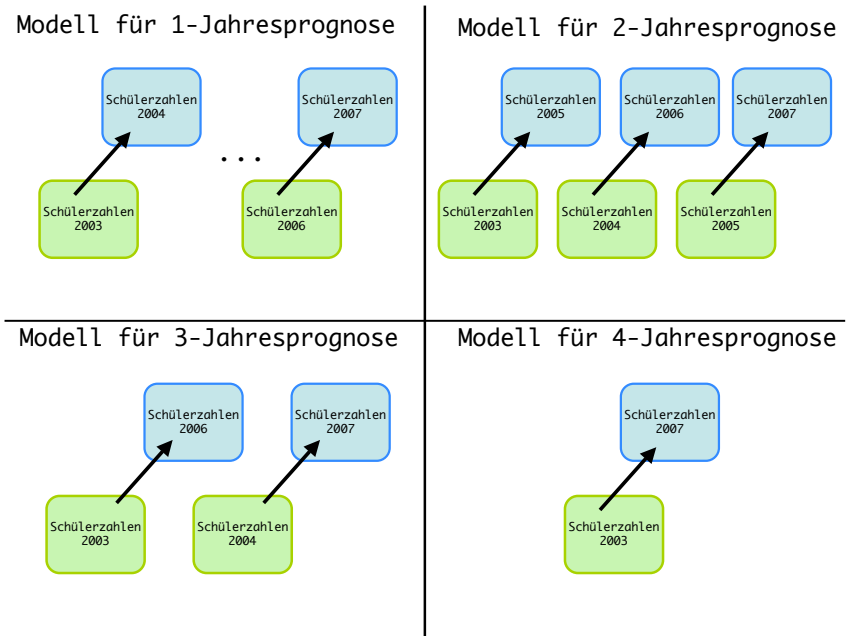

Abbildung 3.3.: Modelle für 4 Prognosehorizonte

3.2.1. Modell aus Verfahren B

Verteilungsannahme

Die Verteilung der Zielvariable Y_{ikt} gehört der Poissonverteilung an (Zähldaten) mit $E\langle Y_{ikt}\rangle = \mu_{ikt}$. Die Zielgrössen Y_{ikt} sind unabängig.

Strukturannahme

Der Erwartungswert μ_{ikt} der Zielvariable Y_{ikt} lässt sich durch eine nichtlineare Funktion $h\langle\rangle = \exp\langle\rangle$ mit dem linearen Prädiktor in Beziehung setzen (Ruckstuhl (2007)).

Unter diesen Voraussetzungen lautet die Modell, nach Schuletappen gegliedert

- Kindergarten

$$E\langle Y_{ikt}\rangle = exp\langle\beta_0+\beta_1 \ln\langle Y_{i(k-\Delta_t)(t-\Delta_t)}\rangle+\sum_{k=1}^{2}\gamma_k ST^{(k)}+\sum_{i=1}^{7}\varphi_i SK^{(i)}+\sum_{k=1}^{2}\sum_{i=1}^{7}\nu_{ki} STK^{(ki)}\rangle \tag{3.10}$$

- Unterstufe

$$E\langle Y_{ikt}\rangle = exp\langle\beta_0+\beta_1 \ln\langle Y_{i(k-\Delta_t)(t-\Delta_t)}\rangle+\sum_{k=3}^{5}\gamma_k ST^{(k)}+\sum_{i=1}^{7}\varphi_i SK^{(i)}+\sum_{k=3}^{5}\sum_{i=1}^{7}\nu_{ki} STK^{(ki)}\rangle \tag{3.11}$$

- Mittelstufe

$$E\langle Y_{ikt}\rangle = exp\langle \beta_0 + \beta_1 \ln\langle Y_{i(k-\Delta_t)(t-\Delta_t)}\rangle + \sum_{k=6}^{8}\gamma_k ST^{(k)} + \sum_{i=1}^{7}\varphi_i SK^{(i)} + \sum_{k=6}^{8}\sum_{i=1}^{7}\nu_{ki} STK^{(ki)}\rangle \tag{3.12}$$

- Oberstufe

$$E\langle Y_{ikt}\rangle = exp\langle \beta_0 + \beta_1 \ln\langle Y_{i(k-\Delta_t)(t-\Delta_t)}\rangle + \sum_{k=9}^{11}\gamma_k ST^{(k)} + \sum_{i=1}^{7}\varphi_i SK^{(i)} + \sum_{k=9}^{11}\sum_{i=1}^{7}\nu_{ki} STK^{(ki)}\rangle \tag{3.13}$$

$$ST^{(k)} = \begin{cases} 1 & \text{falls Beobachtung aus k-ter Stufe} \\ 0 & \text{sonst} \end{cases}$$

$$SK^{(i)} = \begin{cases} 1 & \text{falls Beobachtung aus i-tem Schulkreis} \\ 0 & \text{sonst} \end{cases}$$

$$STK^{(ki)} = SK^{(i)} \cdot ST^{(k)} = \begin{cases} 1 & \text{falls Beobachtung aus k-ter Stufe und i-tem Schulkreis} \\ 0 & \text{sonst} \end{cases}$$

Bemerkungen

- Modell aus Verfahren B enthält 16 Untermodelle - pro Schuletappe und Δ_t eines. Um den Lesefluss nicht zu hindern, wird im folgenden Text jeweils von *einem* Modell aus Verfahren B gesprochen.
- Die Parameter β_0, β_1 und $\gamma_i, i = 1, \ldots, 7$ sind bei jeder Schuletappe verschieden voneinander.
- Null gesetzt werden die Parameter γ_1, γ_3, γ_6, γ_9, φ_1 und ν_p, $p = 11, \ldots, 17, 21, 31, \ldots, 37, 41, 51, 61, \ldots, 67, 71, 81, 91, \ldots, 97, 101, 111$ (Kapitel :Wechselwirkungen).
- β_1 ist gleich 1.

Parameter β_1

β_1 wird gleich 1 gesetzt. Aus welchem Grund dies geschieht, wird im folgenden Abschnitt beschrieben. Das Modell aus Formel 3.11 kann auch ausgeschrieben werden

$$E\langle Y_{ikt}\rangle = e^{\beta_0} \cdot Y_{i(k-\Delta_t)(t-\Delta_t)}^{\beta_1} \cdot e^{\sum_{k=2}^{5}\gamma_k ST^{(k)}} \cdot e^{\sum_{i=1}^{7}\varphi_i SK^{(i)}} \cdot e^{\sum_{k=3}^{5}\sum_{i=1}^{7}\nu_{ki} STK^{(ki)}} \tag{3.14}$$

In Formel 3.14 wird β_1 als Exponenten dargestellt. Welche Werte für β_1 sind plausibel? Ist $\beta_1 < 1$, so wird angenommen, dass, je tiefere die Werte die Variable $Y_{i(k-\Delta_t)(t-\Delta_t)}$ annehmen

kann, desto tiefer sind die Werte für Y_{ikt}. Ist $\beta_1 > 1$, so wird angenommen, dass, je höher die Werte, die Variable $Y_{i(k-\Delta_t)(t-\Delta_t)}$ annehmen kann, desto höher sind die Werte für Y_{ikt} (vgl. Abbildung 3.4). Diese Annahmen können bei Schülerzahlen nicht bestätigt werden *(Beweis noch angeben)*.

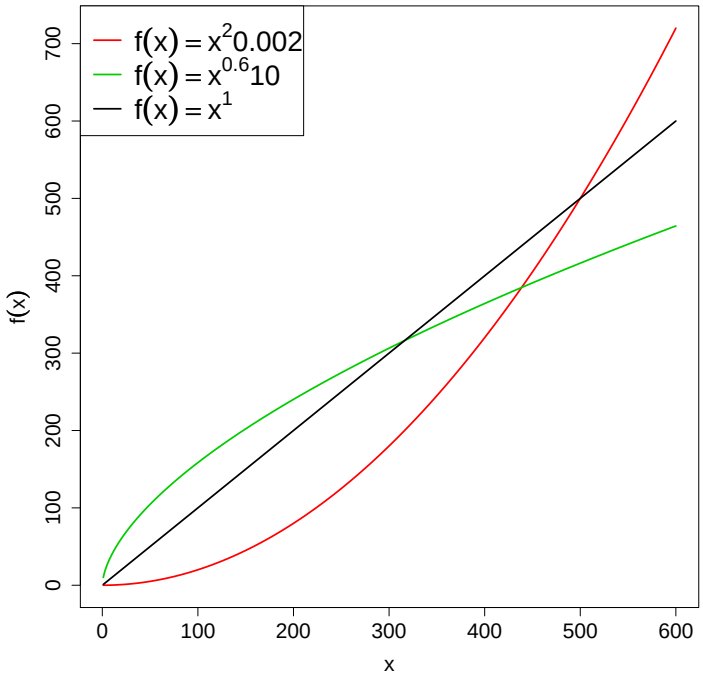

Abbildung 3.4.: Potenzfunktion mit Exponent>1 (rot), mit Exponent<1 (grün) und Exponent $= 1$ (schwarz)

3.2.2. Variablenselektion

Die Variablenselektion ist mit dem Akaike's information criterion AIC durchgeführt worden. Das Modell aus Verfahren B (Formel 3.10 bis 3.13) reduziert sich in einigen Fällen, wobei (+) bedeutet, dass die Variable beibehalten wurde, und (-) das Gegenteil (Tabelle 3.3 bis 3.6). Die Variable $Y_{i(k-\Delta_t)(t-\Delta_t)}$ wurde stets beibehalten. Mit Modell aus Verfahren A ist ebenfalls eine Variablenselektion durchgeführt worden (Tabelle B.1 bis B.4). In allen Fällen wo die Variablenselektion dem Modell von Verfahren B die Wechselwirkungen strich, ist das reduzierte Modell von Verfahren B identisch mit dem reduzierten Modell von Verfahren A.

Kindergarten			
Δ_t	$ST^{(k)}$	$SK^{(i)}$	$STK^{(ki)}$
Δ_1	+	+	+
Δ_2	+	+	-
Δ_3	-	+	-
Δ_4	-	+	-

Tabelle 3.3.: Modell Verfahren B (Kindergarten) nach Variablenselektion

Unterstufe			
Δ_t	$ST^{(k)}$	$SK^{(i)}$	$STK^{(ki)}$
Δ_1	+	+	+
Δ_2	+	+	+
Δ_3	+	+	+
Δ_4	+	+	-

Tabelle 3.4.: Modell Verfahren B (Unterstufe) nach Variablenselektion

Mittelstufe			
Δ_t	$ST^{(k)}$	$SK^{(i)}$	$STK^{(ki)}$
Δ_1	-	-	-
Δ_2	-	+	-
Δ_3	+	+	-
Δ_4	+	+	-

Tabelle 3.5.: Modell Verfahren B (Mittelstufe) nach Variablenselektion

Oberstufe			
Δ_t	$ST^{(k)}$	$SK^{(i)}$	$STK^{(ki)}$
Δ_1	+	+	+
Δ_2	+	+	+
Δ_3	+	+	-
Δ_4	+	+	-

Tabelle 3.6.: Modell Verfahen B (Oberstufe) nach Variablenselektion

3.2.3. Residuenanalyse

Im Anschluss an die Variablenselektion wird mit Hilfe der Residuen geprüft, ob die Modellannahmen stimmen. In Abbildung 3.5 sind links die Tukey-Anscombe-Diagramme pro Schu-

letappe aufgezeichnet. In der rechten Spalte sind Diagramme, um Strukturen in der Varianz der Residuen aufzudecken. Die Residuenanalyse entdeckt keine systematischen Abweichungen oder Unregelmässigkeiten in den Residuen. Die Daten streuen auf der x-Achse nicht ganz harmonisch - die Schulkreise Glattal, Uto und Waidberg weisen sehr hohe Schülerzahlen auf. Schwamedingen, Limmattal, Letzi und Zürichberg eher tiefe. Dazwischen entsteht ein Graben, wo der Lowess-Glätter einen Knick macht (und deshalb die Faustregel $Wert \pm 4^{1/4}$ teilweise verletzt wird). Hebelpunkte sind nach dem Kriterium $2 * AnzahlParameter/n$ keine vorhanden. In Anhang B.3.2 sind die restlichen Abbildungen der Residuenanalye abgebildet.

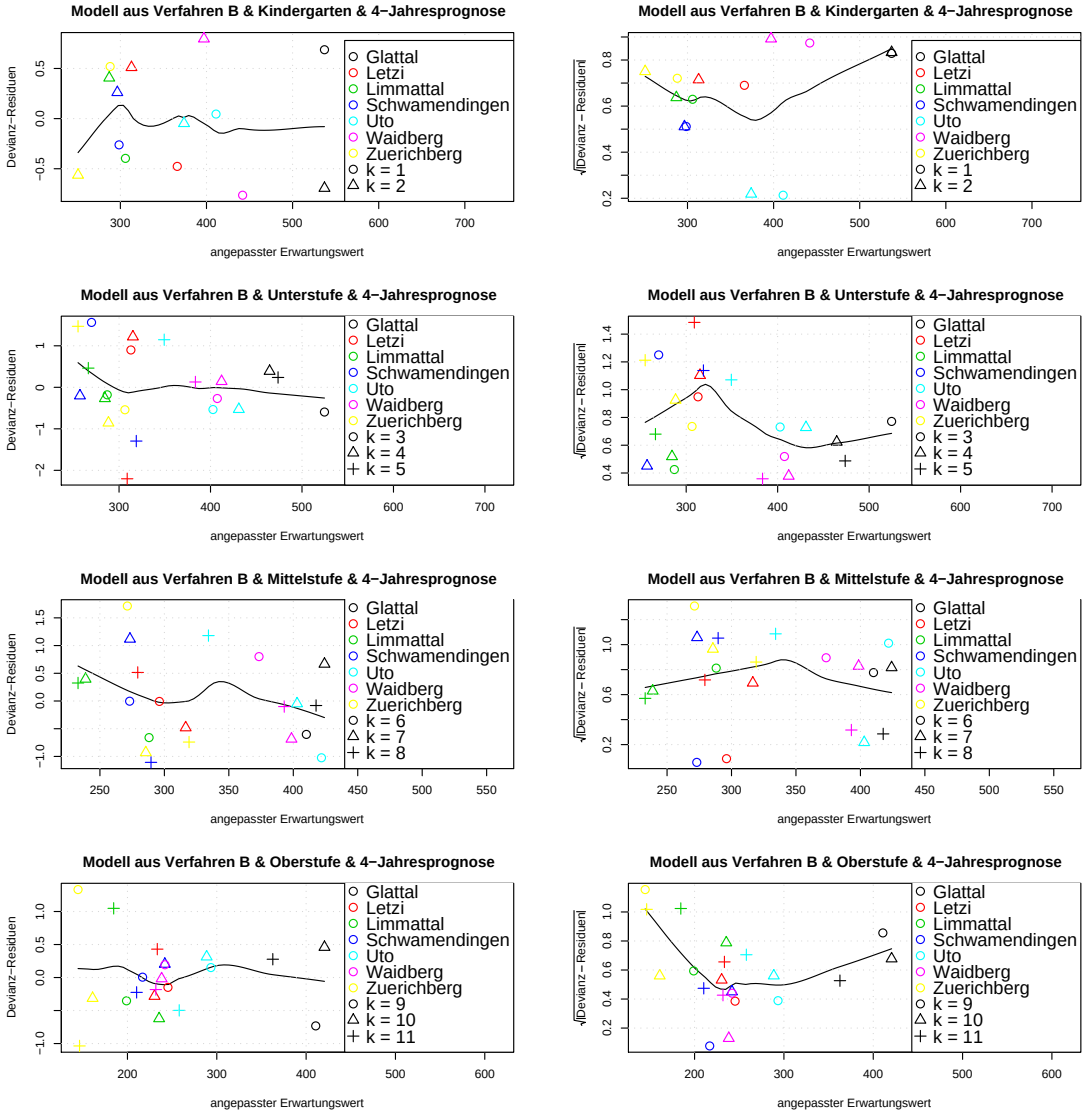

Abbildung 3.5.: Residuenanalyse des Modells von Verfahren B 3.10 bis 3.13 mit $\Delta_t = 4$ nach Variablenselektion

Bei einem Poissonmodell ist der Dispersionsparameter gleich 1 und muss nicht geschätzt werden. Mit der Residuen-Devianz D kann mit der Faustregel $D < (n-p) + 2\sqrt{2(n-p)}$ (Ruck-

stuhl, (2007)) abgeschätzt werden, ob die Abweichungen zwischen der geschätzen Zielgrösse und den echten Werten Zielgrösse unter dem Modell wahrscheinlich sind. D ist bei Modell B.θ_3 (Mittelstufe) bei der 1- und 2-Jahresprognose etwas tief mit $D = 14.869$, df= 62 und $D = 7.8285$, df= 35, was auf eine *Underdispersion* deutet. Das hängt damit zusammen, dass die Modelle (zu) gut angepasst werden und es in diesen Modellen meist nur zwei Parameter und damit viele Freiheitsgrade hat.

3.3. Modellwahl

Mit der der Devianz-Teststatistik (Formel 3.7) wird geprüft, ob die Differenz der Residuen-Devianzen vom Modell aus Verfahren A und Modell aus Verfahren B (beide nach Variablenselektion) gross genug ist, das ein grösseres Modell die Anpassung verbessert.
In allen Fällen, wo Modell B mehr Parameter wie Modell A beinhaltet, ergibt sich mit Modell B eine signifikante Verbesserung gegenüber Modell A.

3.4. Out-of-sample

Bei einer Prognose in die Zukunft sind die zukünftigen Werte unbekannt. Um die Güte von Verfahen A und Verfahren B zu beurteilen und die Modellwahl zu von Kapitel 3.3 zu verifizieren, wurden die Daten für das Jahr 2007 vom Datensatz separiert, $\Delta_t = 4$ fällt weg. Aus diesem Grund gibt es 3 statt 4 Prognosehorizonte.
Die Out-of-sample-Modelle wurden analog zu Verfahren A und B aufgebaut.

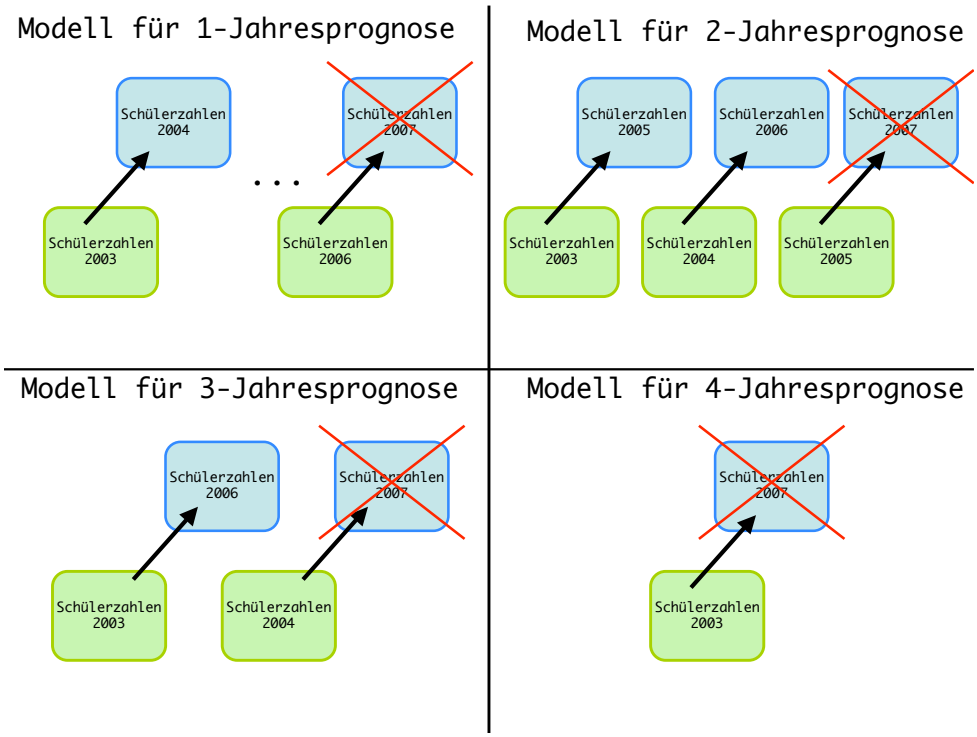

Abbildung 3.6.: Out-of-sample-Modelle für 3 Prognosehorizonte - die 4-Jahresprognose fällt weg. Die Daten aus dem Jahr 2007 werden zur Überprüfung verwendet.

Variablenselektion

Das Resultat der Variablenselektion vom Modell aus Verfahren B der Out-of-sample-Daten (Tabellen 3.10 bis 3.10) stimmt mit jenem von Modell aus Verfahren B in Kapitel 3.2.2 überein. Die Variablenwahl ist dieselbe - ausser bei der Unterstufe, Δ_3, wo bei der Out-of-sample-Version auch die Wechselwirkungen wegfallen. Analog zu Kapitel 3.2.2: In allen Fällen wo die Variablenselektion dem Modell von Verfahren B die Wechselwirkungen strich, ist das reduzierte Modell von Verfahren B identisch mit dem reduzierten Modell von Verfahren A.

Kindergarten			
Δ_t	$ST^{(k)}$	$SK^{(i)}$	$STK^{(ki)}$
Δ_1	+	+	+
Δ_2	+	+	-
Δ_3	-	+	-

Tabelle 3.7.: Out-of-sample: Modell aus Verfahren B (Kindergarten) nach Variablenselektion

Unterstufe			
Δ_t	$ST^{(k)}$	$SK^{(i)}$	$STK^{(ki)}$
Δ_1	+	+	+
Δ_2	+	+	+
Δ_3	+	+	-

Tabelle 3.8.: Out-of-sample: Modell aus Verfahren B (Unterstufe) nach Variablenselektion

Mittelstufe			
Δ_t	$ST^{(k)}$	$SK^{(i)}$	$STK^{(ki)}$
Δ_1	-	-	-
Δ_2	-	+	-
Δ_3	+	+	-

Tabelle 3.9.: Out-of-sample: Modell aus Verfahren B (Mittelstufe) nach Variablenselektion

Oberstufe			
Δ_t	$ST^{(k)}$	$SK^{(i)}$	$STK^{(ki)}$
Δ_1	+	+	+
Δ_2	+	+	+
Δ_3	+	+	-

Tabelle 3.10.: Out-of-sample: Modell aus Verfahren B (Oberstufe) nach Variablenselektion

Goodness-of-fit

Mit Modell aus Verfahren B und dem Modell aus Verfahren A wird eine Prognose gemacht - vom Jahr 2006 eine 1-Jahresprognose, vom Jahr 2005 eine 2-Jahresprognose und vom Jahr 2004 eine 3-Jahresprognose. Somit gibt es 3 Prognosen für das Jahr 2007. Um die Güte des Modells zu messen, kann das Mass *Goodness-of-fit* (Formel 3.8) verwendet werden. Damit werden die Abstände zwischen der Prognose und den realen Werten vom Jahr 2007 quadriert, durch die Prognose dividiert und aufsummiert. Der Goodness-of-fit wird auch von den Prognosen des Schul-und Sportdepartements durchgeführt. Diese drei Goodness-of-fit-Werte können verglichen werden. Im Vergleich zählt: je tiefer die Summe mit den quadrierten Werten, desto besser. Abbildung 3.7 bis 3.9 zeigt, dass das Modell aus Verfahren B (grüner Balken) etwas besser oder mindestens gleich gut abschneidet wie das Modell aus Verfahren A. Wenn beide Balken gleich hoch sind, werden dieselben Modellstrukturen verwendet (identisches Modell). Nur in einem Fall, 2-Jahresprognose, Unterstufe, ist Modell A leicht besser wie Modell B. Vom Schul- und Sportdepartement gibt es keine Prognosen für den Kindergarten.

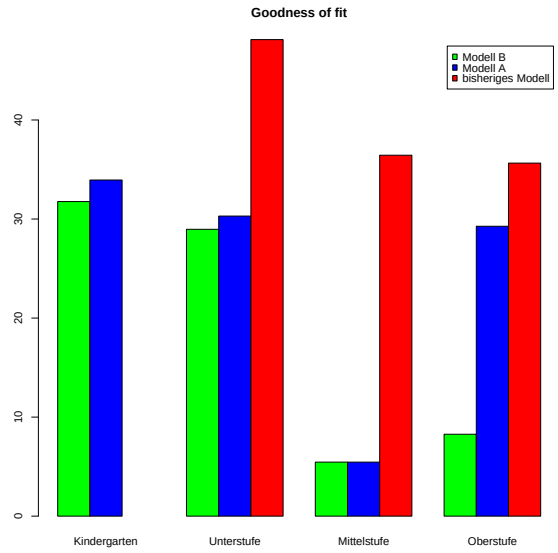

Abbildung 3.7.: Goodness-of-fit der 1-Jahresprognosen für 2007: Prognose mit Modell aus Verfahren A (grün), Prognose mit Modell aus Verfahren B (blau), Prognose des Schul- und Sportdepartements (rot)

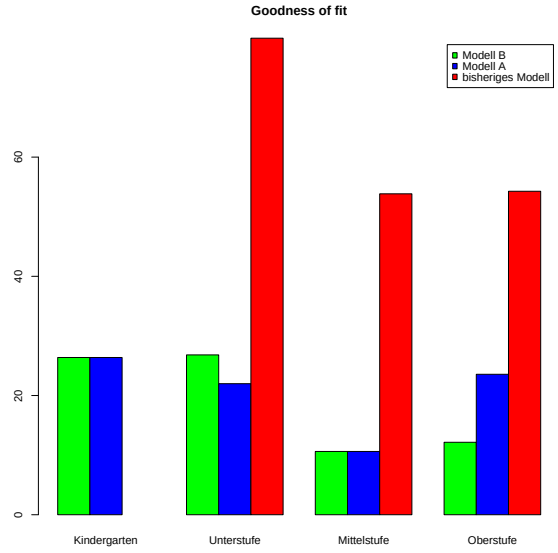

Abbildung 3.8.: Goodness-of-fit der 2-Jahresprognosen für 2007: Prognose mit Modell aus Verfahren A (grün), Prognose mit Modell aus Verfahren B (blau), Prognose des Schul- und Sportdepartements (rot)

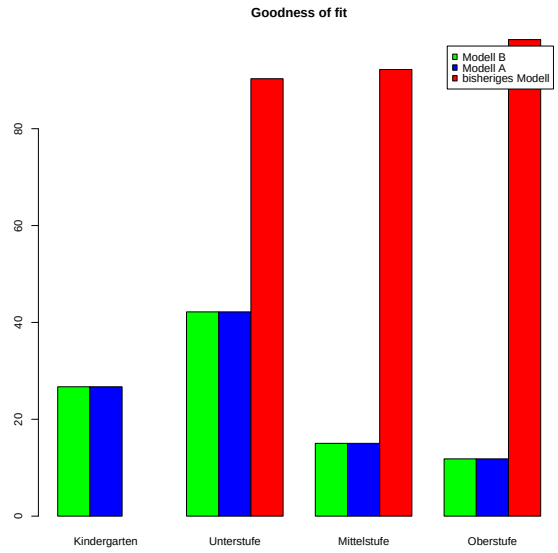

Abbildung 3.9.: Goodness-of-fit der 3-Jahresprognosen für 2007: Prognose mit Modell aus Verfahren A (grün), Prognose mit Modell aus Verfahren B (blau), Prognose des Schul- und Sportdepartements (rot)

In Anhang B.3.4 sind die ausführlichen Resultate der Out-sample-Prognose.

3.5. Ergebnisse

Verfahren BDer Devianztest in Kapitel 3.3 hat gezeigt, dass Wechselwirkungen - und damit mehr Parameter - eine signifikante Verbesserung des Modells bewirken. Werden aus dem Datensatz die Schülerzahlen vom Jahr 2007 hinausgenommen (Kapitel 3.4) und mit den Out-of-sample-Daten die Modelle A und B geschätzt, so zeigt sich, dass Modell B im einen kleineren Goodness-of-fit aufweist. Deshalb wird Modell B (Formel 3.10 bis 3.13) gewählt und eine Variablenselektion durchgeführt.

4. Diskussion und Ausblick

4.1. Zusammenfassung und Interpretation der Resultate

Das Ziel der Projektarbeit aus Kapitel 1.2

1. ein statistisches Modell für eine Kurz- bis Mittelfristprognose von 4 Jahren
2. die Schülerzahlen pro Schulkreis und Schulstufe zu prognostizieren
3. bessere Abschätzung als das Schul- und Sportdepartement
4. weniger erklärenden Grössen

Die Abbildungen zu den Prognosen der Jahre `2008`, `2009`, `2010` und `2011` finden sich in Anhang B.3.

4.2. Ausblick auf offene Fragen

Bautätigkeit

θ_1 mit KG und 1.Klasse?

Prognosehorizont

Erweiterung des Modells Wenn $\Delta_t > 4$

Detaillierungsgrad Geografie

Nicht Schulkreise, sondern Schuleinheiten

Verteilung auf Klassen

4.3. weitere Auswertemöglichkeiten

A. Literaturverzeichnis

Müller, M. (2006). Experimental Design, Unveröffentliches Vorlesungsskript, Zürcher Hochschule für Angewandte Wissenschaften.

Ruckstuhl, A. (2006). Statistisches Modellieren I: Statistische Regressionsrechnung und ihre Anwendung, unveröffentlichtes Vorlesungsskript, Zürcher Hochschule für Angewandte Wissenschaften.

Ruckstuhl, A. (2007). Statistisches Modellieren II: Generalisierte lineare Modelle und ihre Anwendung, unveröffentlichtes Vorlesungsskript, Zürcher Hochschule für Angewandte Wissenschaften.

B. Anhang

B.1. Daten

B.1.1. detaillierte Datenaufbereitung

versetzter Jahrgang

Jahr	versetzter Jahrgang	Datum bis [YYMMDD]	Datum von [YYMMDD]
2007	JG0	20070501	20071231
2007	JG1	20060501	20070430
2007	JG2	20050501	20060430
2007	JG3	20040501	20050430
2007	JG4	20030501	20040430
2006	JG0	20060501	20061231
2006	JG1	20050501	20060430
2006	JG2	20040501	20050430
2006	JG3	20030501	20040430
2006	JG4	20020501	20030430
2005	JG0	20050501	20051231
2005	JG1	20040501	20050430
2005	JG2	20030501	20040430
2005	JG3	20020501	20030430
2005	JG4	20010501	20020430
2004	JG0	20040501	20041231
2004	JG1	20030501	20040430
2004	JG2	20020501	20030430
2004	JG3	20010501	20020430
2004	JG4	20000501	20010430
2003	JG0	20030501	20031231
2003	JG1	20020501	20030430
2003	JG2	20010501	20020430
2003	JG3	20000501	20010430
2003	JG4	19990501	20000430

B.1.2. SPSS

Neben Excel ist mit SPSS gearbeitet worden. Neucodierungen funktionieren mit SPSS beispielsweise besser, wie auch das Selektieren und Aggregieren von Daten.

B.2. weitere Modellverfahren

B.2.1. Modell aus Verfahren A

Verteilungsannahme

Die Verteilung der Zielvariable Y_{ikt} gehört der Poissonverteilung an (Zähldaten) mit $E\langle Y_{ikt,j}\rangle = \mu_{ikt}$. Die Zielgrössen Y_{ikt} sind unabängig.

Strukturannahme

Der Erwartungswert μ_{ikt} der Zielvariable Y_{ikt} lässt sich durch eine nichtlineare Funktion $h\langle\rangle = \exp\langle\rangle$ mit dem linearen Prädiktor in Beziehung setzen (Ruckstuhl (2007)).

Unter diesen Voraussetzungen lautet das Modell , nach Schuletappen gegliedert

- Kindergarten

$$E\langle Y_{ikt}\rangle = \exp\langle\beta_0 + \beta_1 \ln\langle Y_{i(k-\Delta_t)(t-\Delta_t)}\rangle + \sum_{k=1}^{2} \gamma_k ST^{(k)} + \sum_{i=1}^{7} \varphi_i SK^{(i)}\rangle \tag{B.1}$$

- Unterstufe

$$E\langle Y_{ikt}\rangle = \exp\langle\beta_0 + \beta_1 \ln\langle Y_{i(k-\Delta_t)(t-\Delta_t)}\rangle + \sum_{k=3}^{5} \gamma_k ST^{(k)} + \sum_{i=1}^{7} \varphi_i SK^{(i)}\rangle \tag{B.2}$$

- Mittelstufe

$$E\langle Y_{ikt}\rangle = \exp\langle\beta_0 + \beta_1 \ln\langle Y_{i(k-\Delta_t)(t-\Delta_t)}\rangle + \sum_{k=6}^{8} \gamma_k ST^{(k)} + \sum_{i=1}^{7} \varphi_i SK^{(i)}\rangle \tag{B.3}$$

- Oberstufe

$$E\langle Y_{ikt}\rangle = \exp\langle\beta_0 + \beta_1 \ln\langle Y_{i(k-\Delta_t)(t-\Delta_t)}\rangle + \sum_{k=9}^{11} \gamma_k ST^{(k)} + \sum_{i=1}^{7} \varphi_i SK^{(i)}\rangle \tag{B.4}$$

$$ST^{(k)} = \begin{cases} 1 & \text{falls Beobachtung aus k-ter Stufe} \\ 0 & \text{sonst} \end{cases}$$

$$SK^{(i)} = \begin{cases} 1 & \text{falls Beobachtung aus i-tem Schulkreis} \\ 0 & \text{sonst} \end{cases}$$

Bemerkungen

- Modell aus Verfahren A enthält 16 Untermodelle - pro Schuletappe und Δ_t eines. Um den Lesefluss nicht zu hindern, wird im folgenden Text jeweils von *einem* Modell aus Verfahren A gesprochen.
- Die Parameter β_0, β_1 und $\gamma_i, i = 1, \ldots, 7$ sind bei jeder Schuletappe verschieden voneinander.
- Null gesetzt werden die Parameter γ_1, γ_3, γ_6, γ_9 und φ_1 (Kapitel :Wechselwirkungen).
- β_1 ist gleich 1.

B.2.2. Variablenselektion

Die Variablenselektion wird mit dem *Akaike's information criterion* AIC durchgeführt. Die Modelle B.1 bis B.4 reduzieren sich wie in Tabelle B.1 bis B.4, wobei + bedeutet, dass die Variable beibehalten wurde, und - das Gegenteil. Die Variable $Y_{i(k-\Delta_t)(t-\Delta_t)}$ wurde stets beibehalten.

Kindergarten		
Δ_t	$ST_{it}^{(k)}$	$SK_{kt}^{(i)}$
Δ_1	+	+
Δ_2	+	+
Δ_3	-	+
Δ_4	-	+

Tabelle B.1.: Modell aus Verfahren A (Kindergarten) nach Variablenselektion

Unterstufe		
Δ_t	$ST_{it}^{(k)}$	$SK_{kt}^{(i)}$
Δ_1	+	+
Δ_2	+	+
Δ_3	+	+
Δ_4	+	+

Tabelle B.2.: Modell aus Verfahren A (Unterstufe) nach Variablenselektion

Mittelstufe		
Δ_t	$ST_{it}^{(k)}$	$SK_{kt}^{(i)}$
Δ_1	-	-
Δ_2	-	+
Δ_3	+	+
Δ_4	+	+

Tabelle B.3.: Modell aus Verfahren A (Mittelstufe) nach Variablenselektion

Oberstufe		
Δ_t	$ST_{it}^{(k)}$	$SK_{kt}^{(i)}$
Δ_1	+	+
Δ_2	+	+
Δ_3	+	+
Δ_4	+	+

Tabelle B.4.: Modell aus Verfahren A (Oberstufe) nach Variablenselektion

B.3. Abbildungen

B.3.1. ergänzende deskriptive Statistik

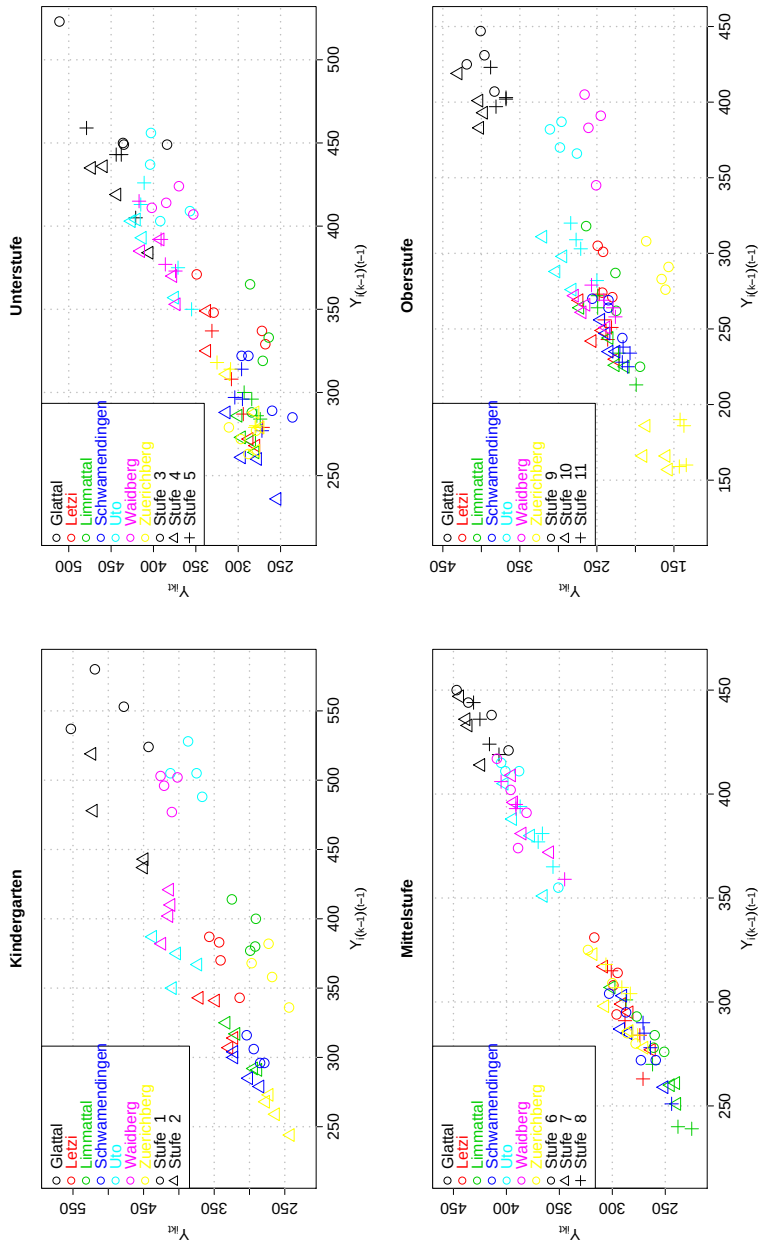

Abbildung B.1.: Schülerzahlen der Stadt Zürich aufgeteilt in 4 Gruppen: auf der x-Achse die Variable vom Kapitel 2.2.6 aufgetragen (mit Δ_1), auf der y-Achse ist die Variable vom Kapitel 2.2.5 aufgetragen.

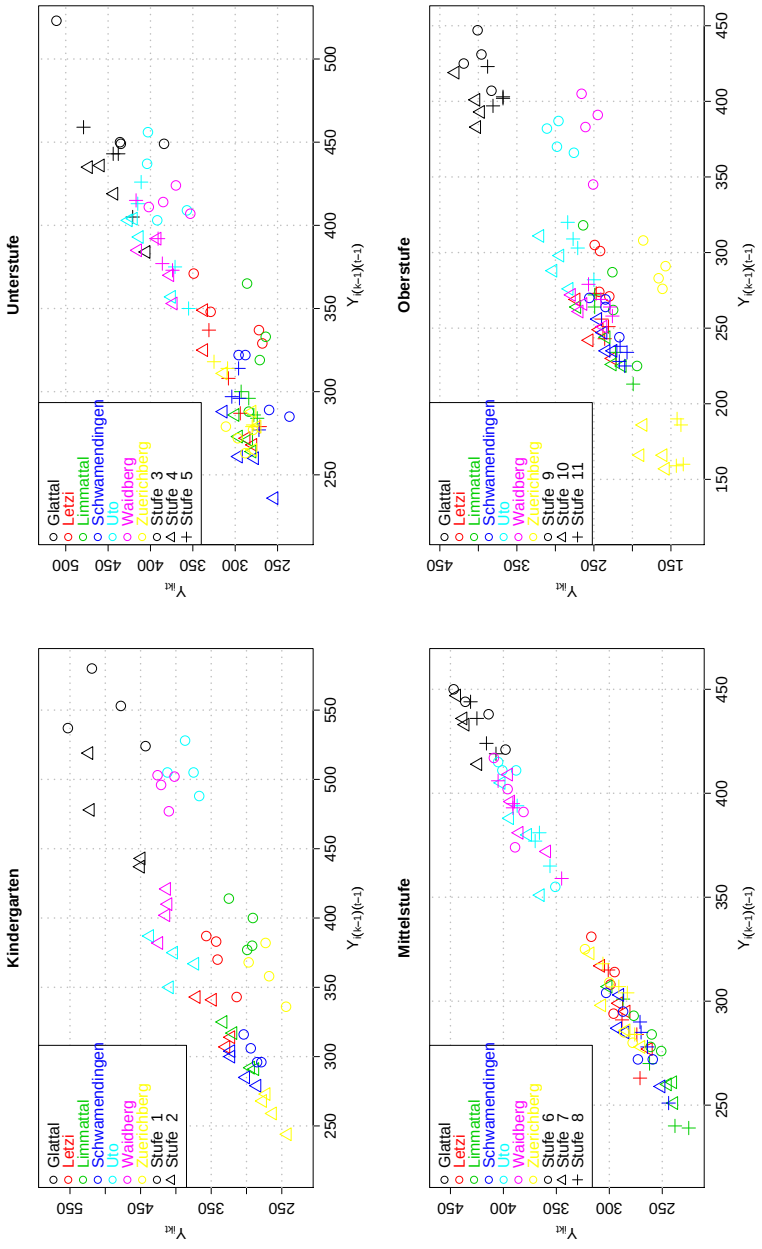

Abbildung B.2.: Schülerzahlen der Stadt Zürich aufgeteilt in 4 Gruppen: auf der x-Achse die Variable vom Kapitel 2.2.6 aufgetragen (mit Δ_2), auf der y-Achse ist die Variable vom Kapitel 2.2.5 aufgetragen.

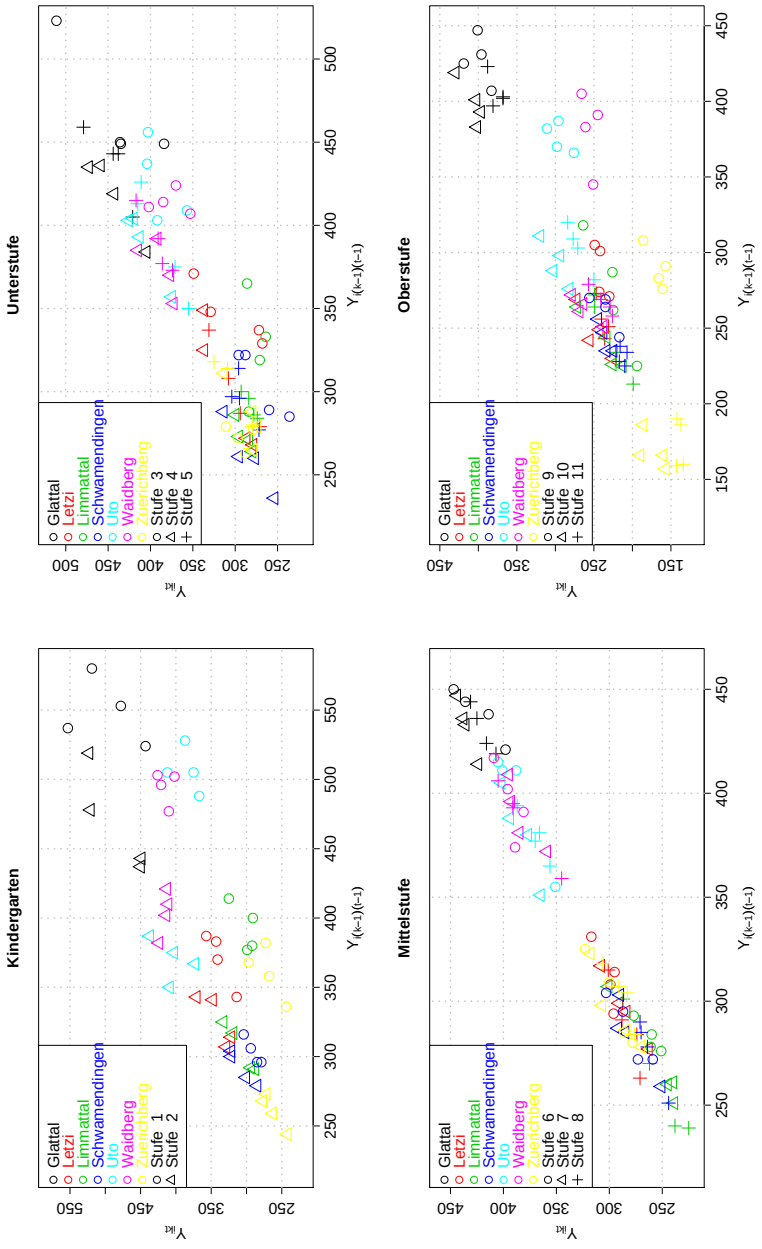

Abbildung B.3.: Schülerzahlen der Stadt Zürich aufgeteilt in 4 Gruppen: auf der x-Achse die Variable vom Kapitel 2.2.6 aufgetragen (mit Δ_3), auf der y-Achse ist die Variable vom Kapitel 2.2.5 aufgetragen.

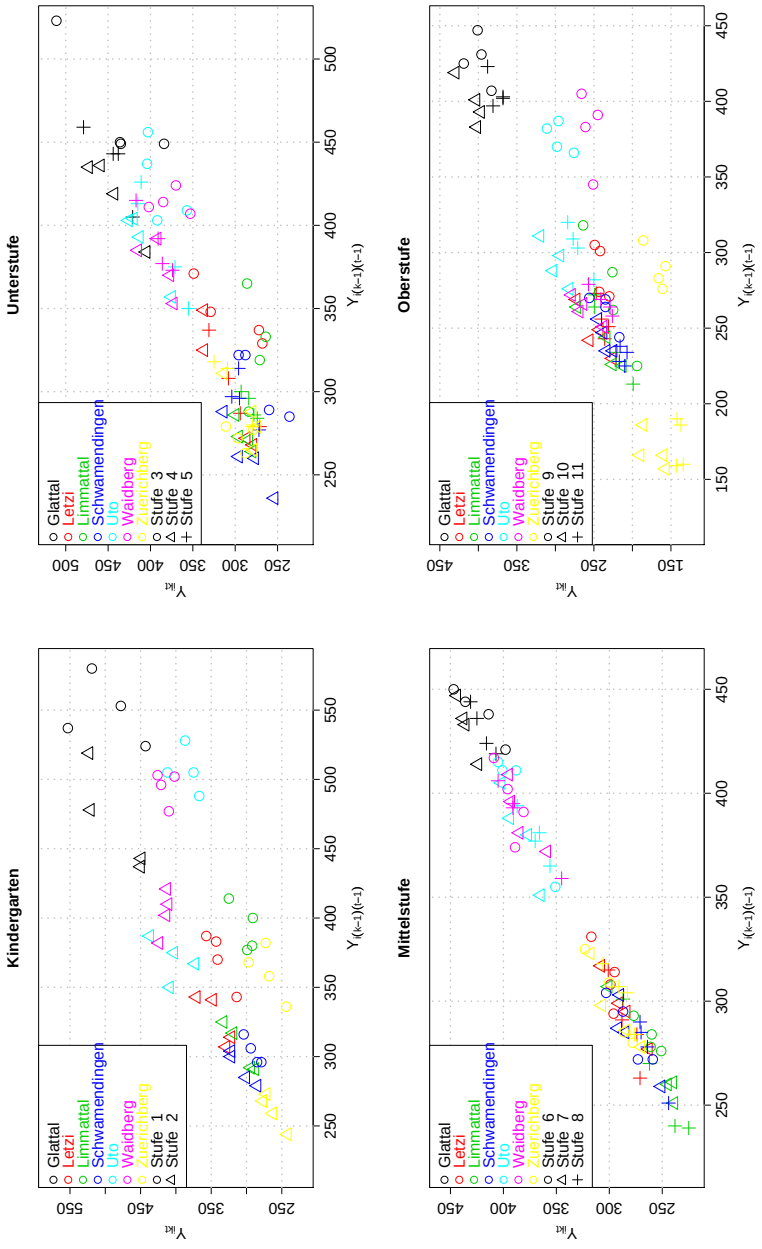

Abbildung B.4.: Schülerzahlen der Stadt Zürich aufgeteilt in 4 Gruppen: auf der x-Achse die Variable vom Kapitel 2.2.6 aufgetragen (mit Δ_4), auf der y-Achse ist die Variable vom Kapitel 2.2.5 aufgetragen.

B.3.2. Residuenanalyse Verfahren B

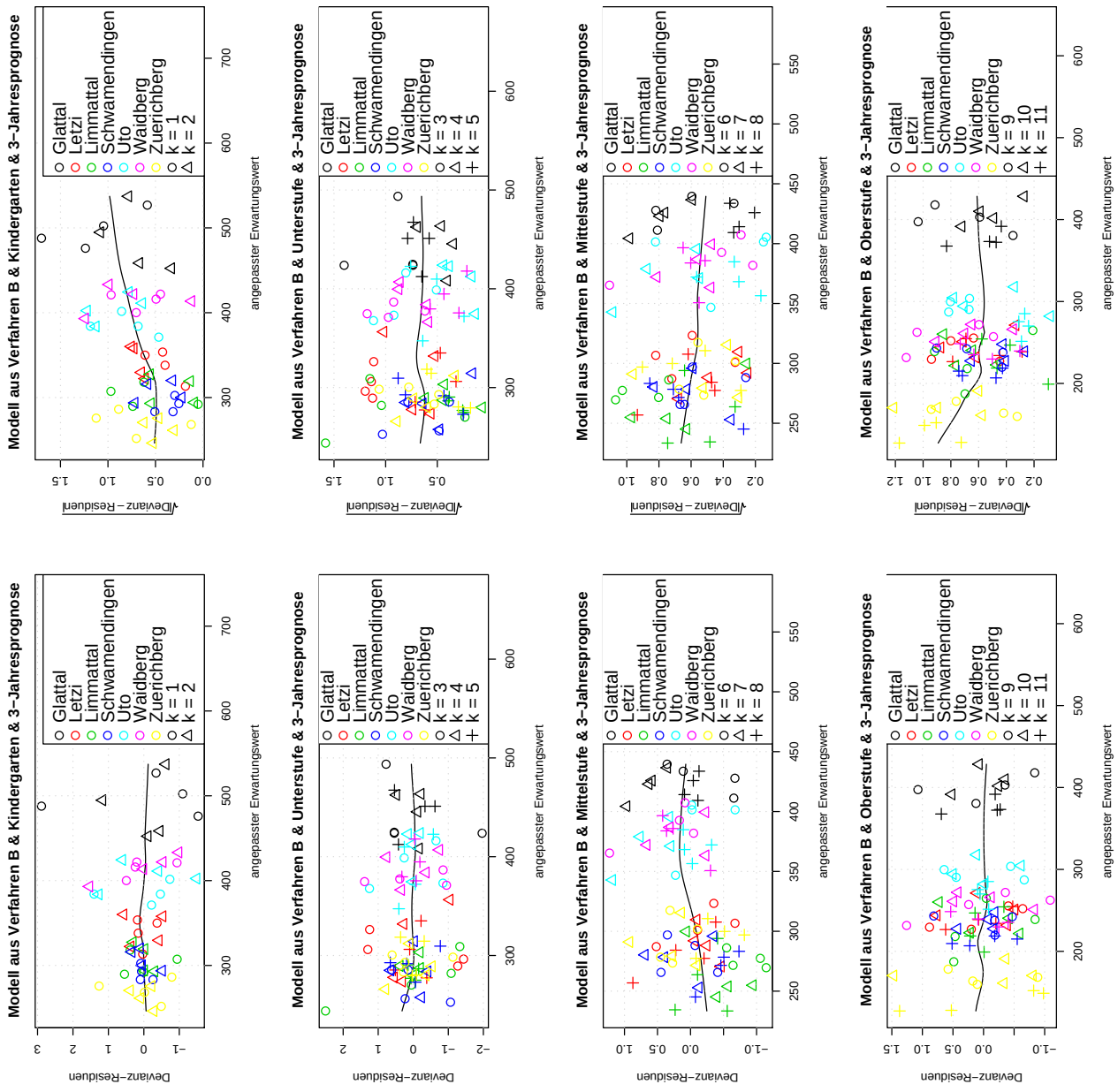

Abbildung B.5.: Residuenanalyse von Modell aus Verfahren B (Formel 3.10 bis 3.13 mit $\Delta_t = 1$, nach Variablenselektion)

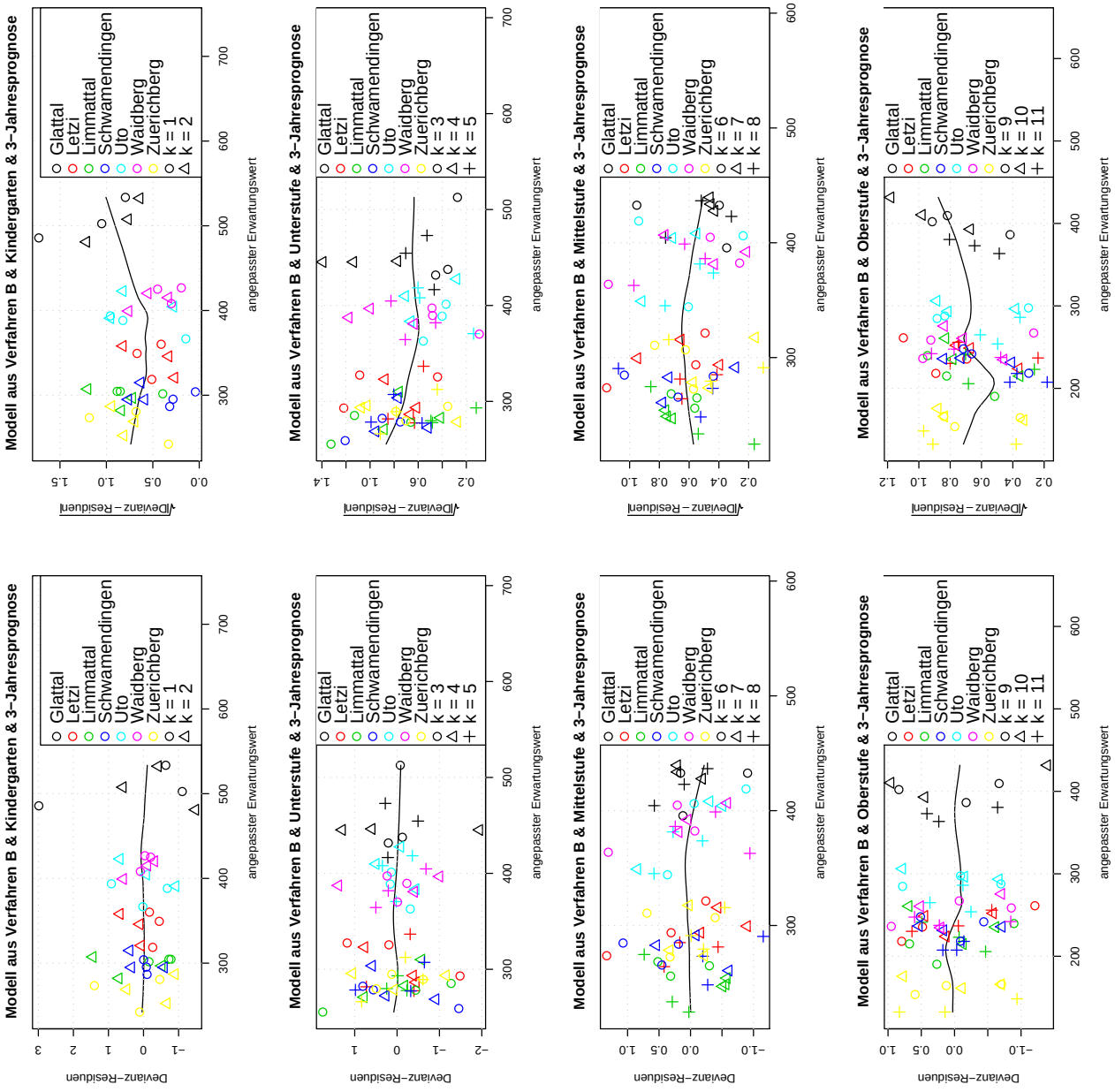

Abbildung B.6.: Residuenanalyse von Modell aus Verfahren B (Formel 3.10 bis 3.13 mit $\Delta_t = 2$, nach Variablenselektion)

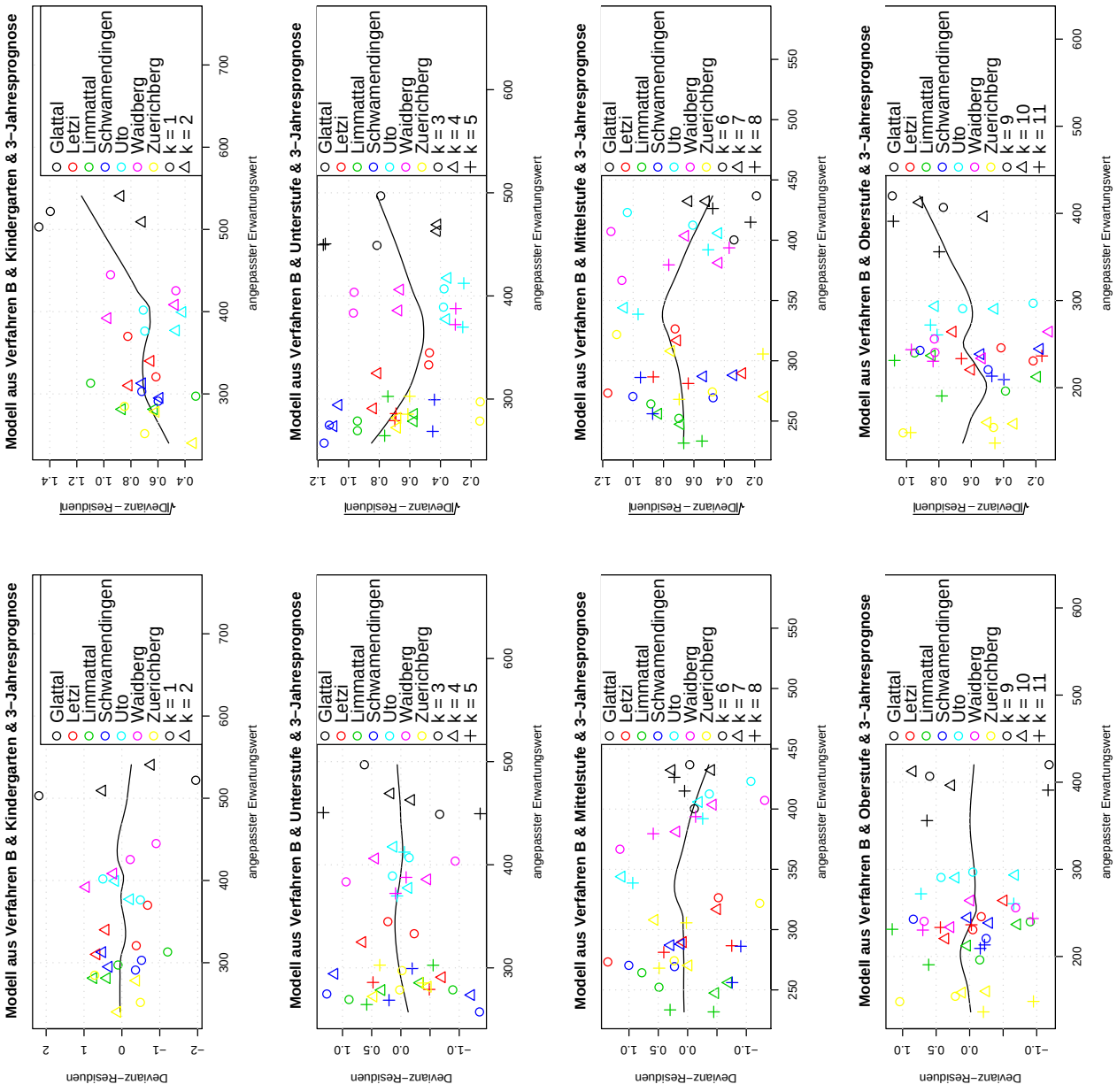

Abbildung B.7.: Residuenanalyse von Modell aus Verfahren B (Formel 3.10 bis 3.13 mit $\Delta_t = 3$, nach Variablenselektion)

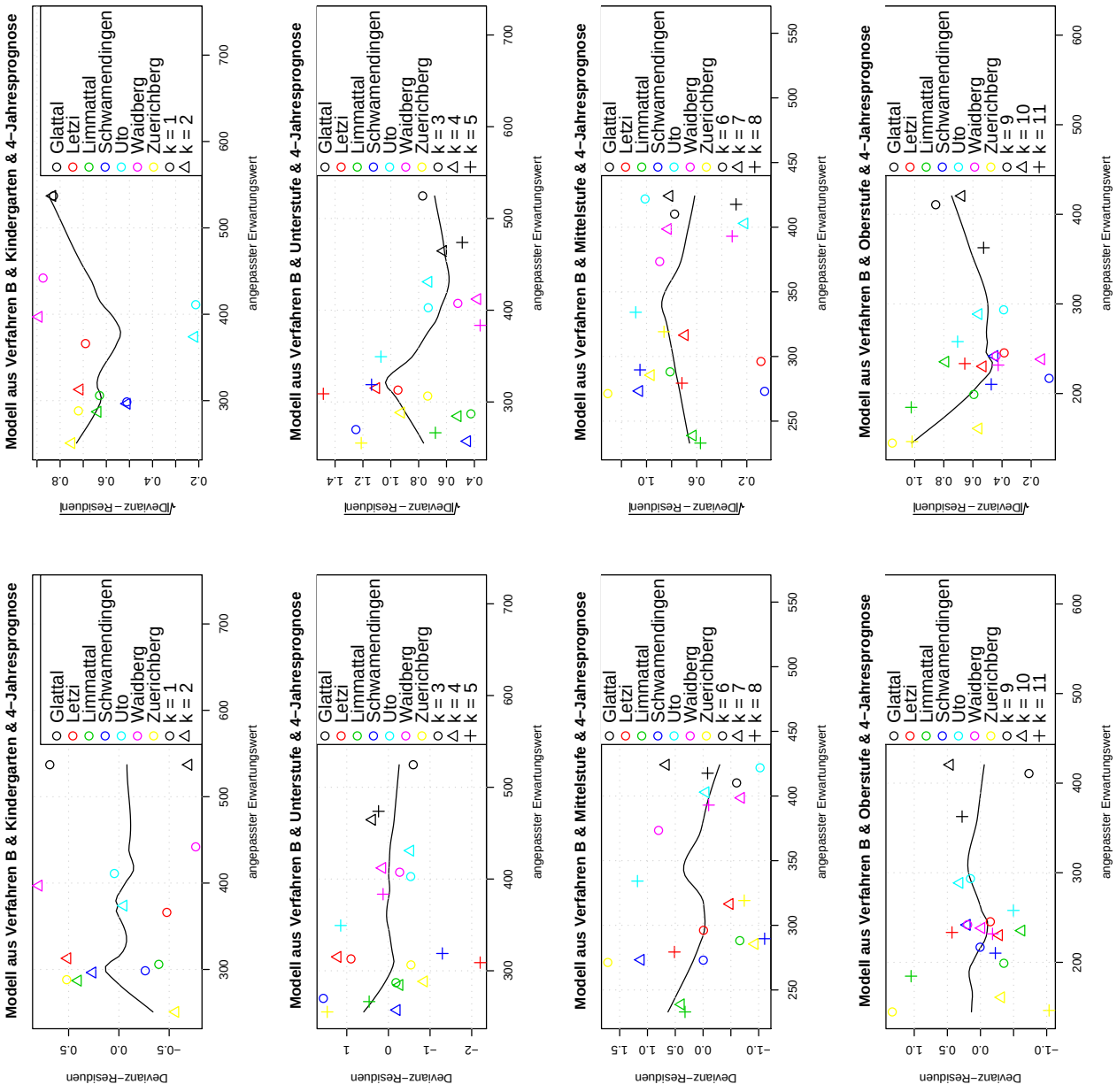

Abbildung B.8.: Residuenanalyse von Modell aus Verfahren B (Formel 3.10 bis 3.13 mit $\Delta_t = 4$, nach Variablenselektion)

B.3.3. Residuenanalyse Verfahren A

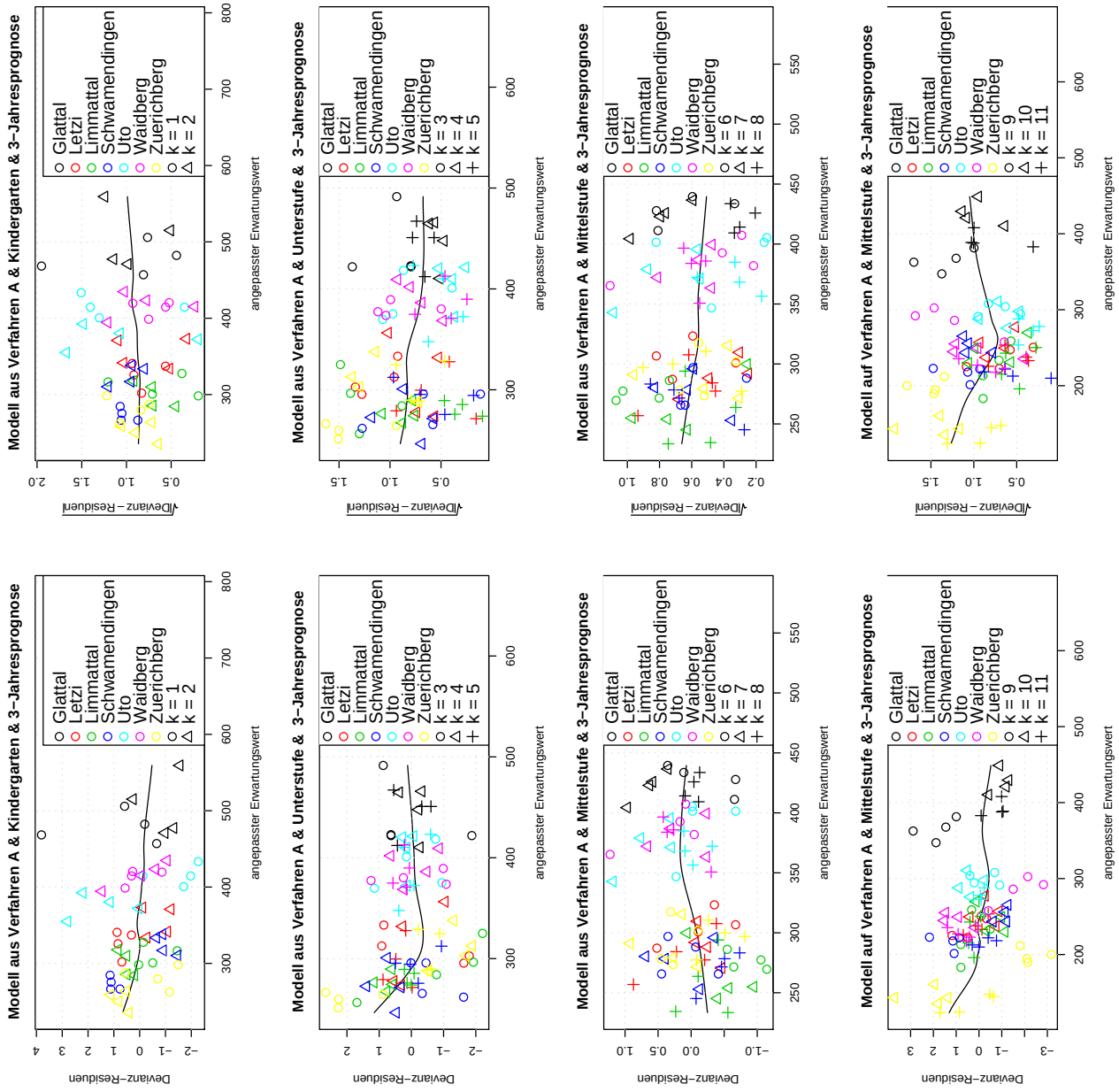

Abbildung B.9.: Residuenanalyse von Modell aus Verfahren A (Formel B.1 bis B.4 mit $\Delta_t = 1$, nach Variablenselektion)

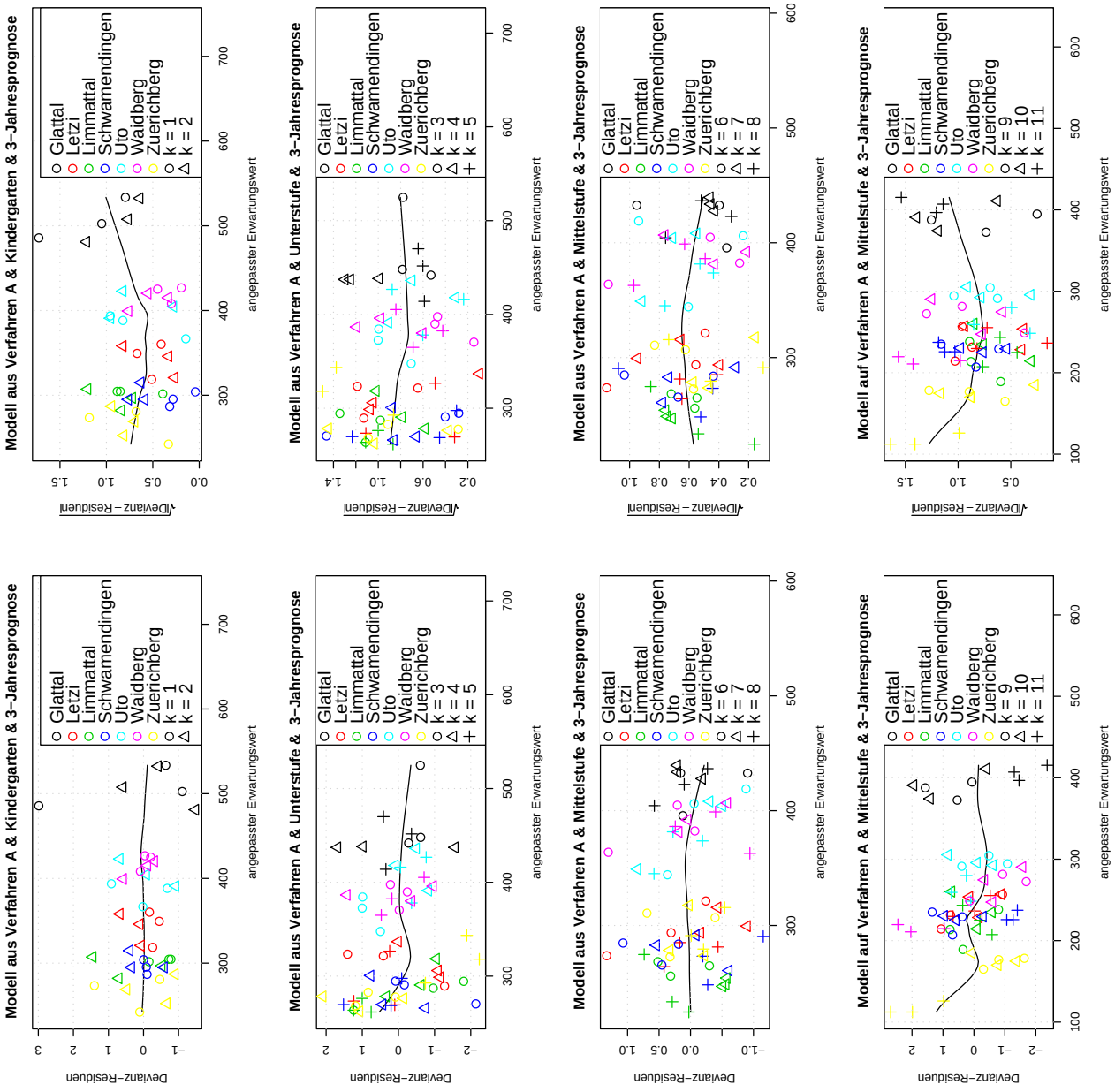

Abbildung B.10.: Residuenanalyse von Modell aus Verfahren A (Formel B.1 bis B.4 mit $\Delta_t = 2$, nach Variablenselektion)

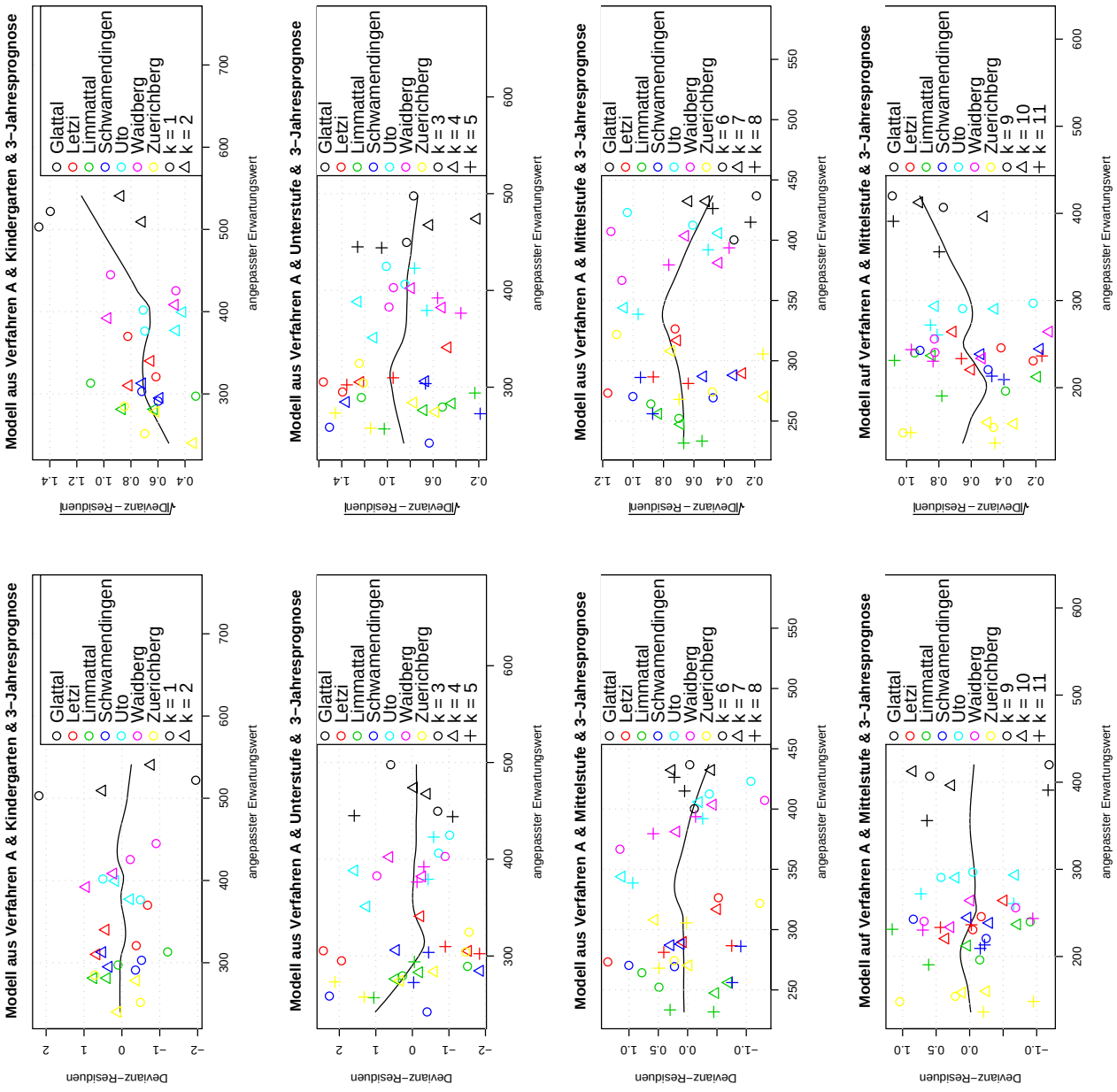

Abbildung B.11.: Residuenanalyse von Modell aus Verfahren A (Formel B.1 bis B.4 mit $\Delta_t = 3$, nach Variablenselektion)

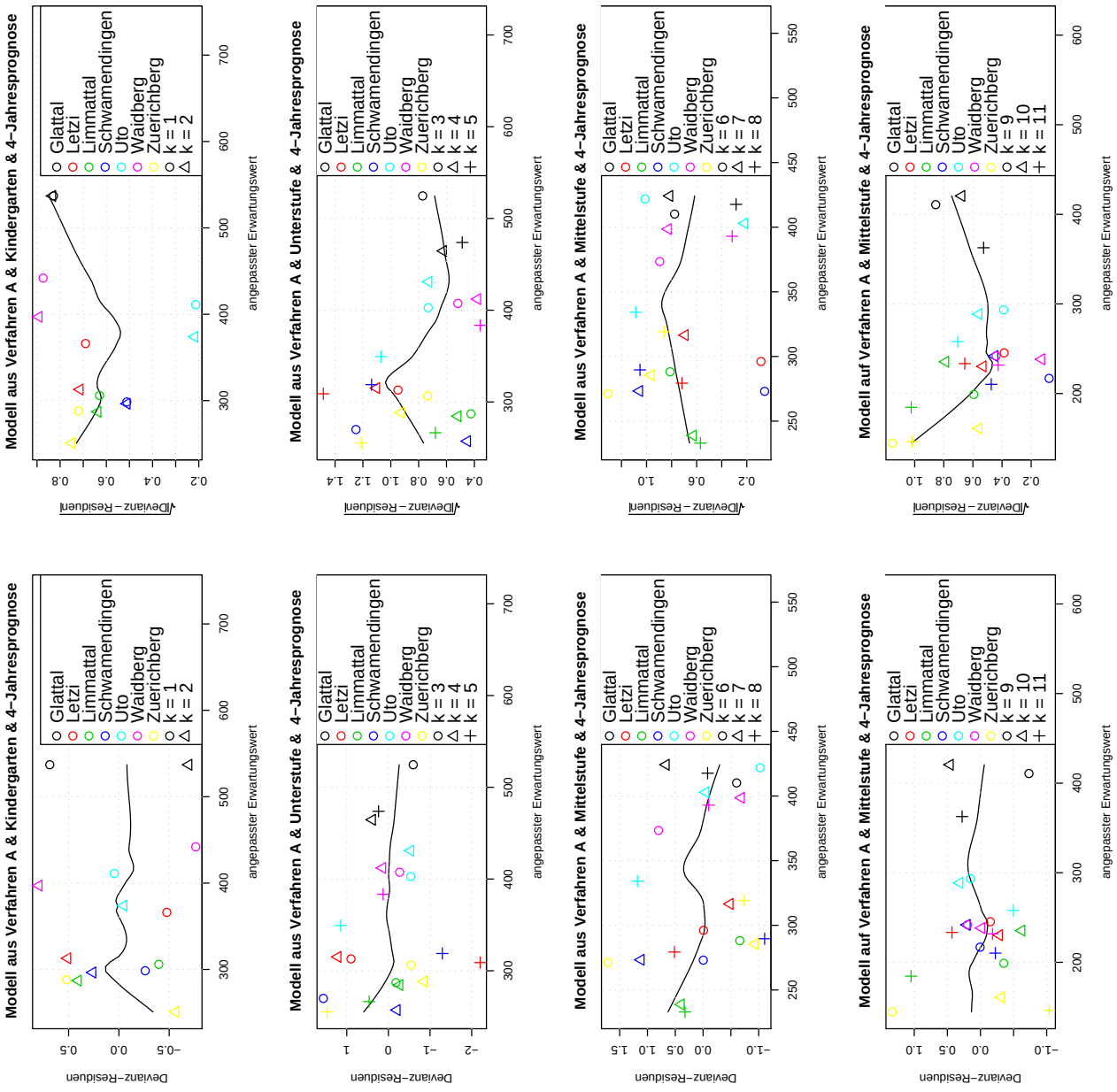

Abbildung B.12.: Residuenanalyse von Modell aus Verfahren A (Formel B.1 bis B.4 mit $\Delta_t = 4$, nach Variablenselektion)

B.3.4. Out-of-sample-Prognose

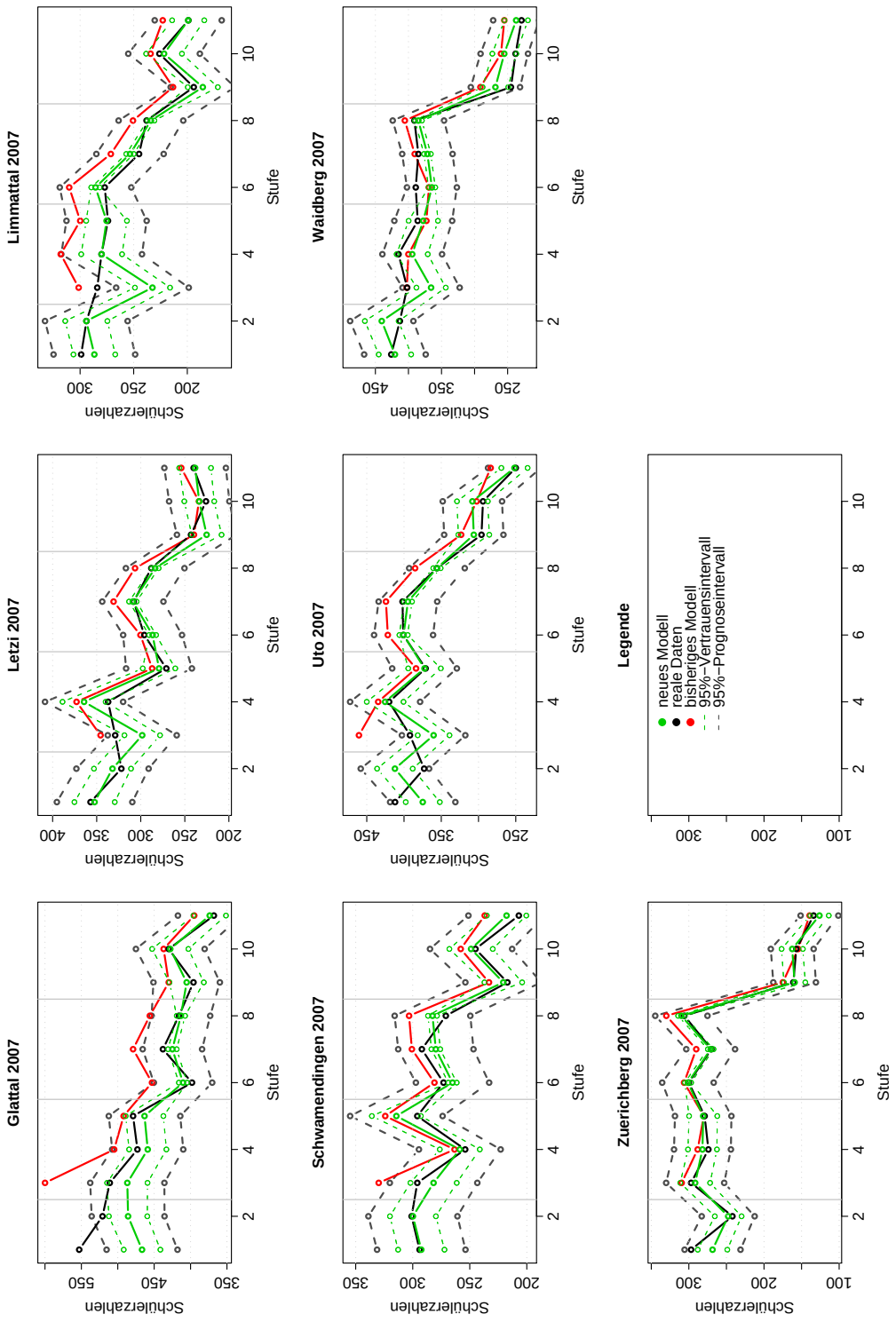

Abbildung B.13.: Ausgehend vom Jahr 2006 eine 1-Jahresprognose der Out-of-sample-Modelle für das Jahr 2007

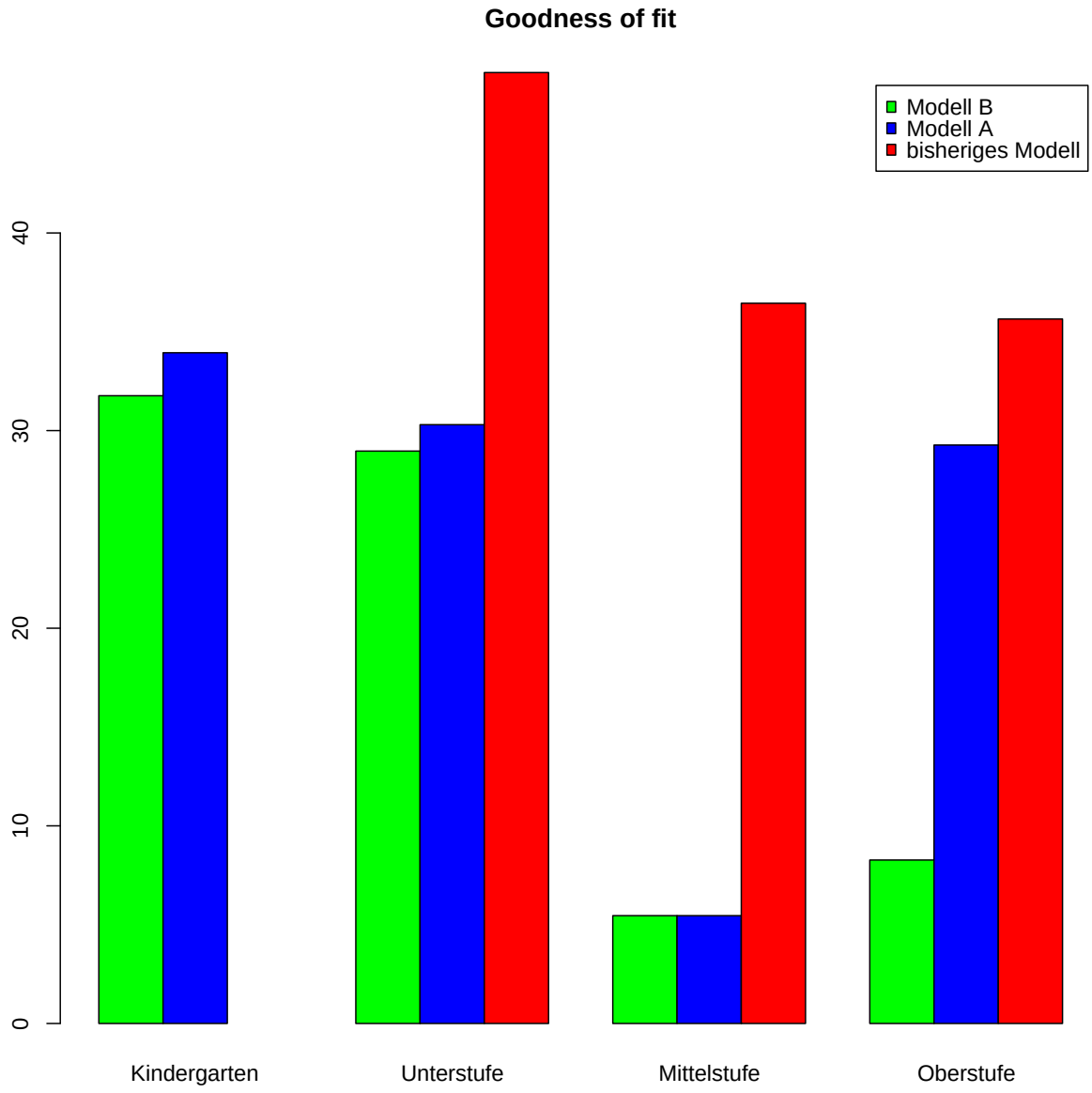

Abbildung B.14.: Goodness-of-fit der 1-Jahresprognosen für 2007: Prognose mit Modell aus Verfahren A (grün), Prognose mit Modell aus Verfahren B (blau), Prognose des Schul- und Sportdepartements (rot)

Die entsprechenden Daten zur Abbildung B.13

```
Schulkreis      Stufe       Anzahl Schüler
```

% Glattal		1		466.6747
33	Glattal	1		466.6747
23	Glattal	2		485.8585
24	Glattal	3		486.9177
25	Glattal	4		458.8741
26	Glattal	5		463.2665
27	Glattal	6		410.3658
28	Glattal	7		424.9869
29	Glattal	8		413.2900
30	Glattal	9		405.3894
31	Glattal	10		427.8997
32	Glattal	11		373.7318
77	Letzi	1		352.3960
67	Letzi	2		332.0605
68	Letzi	3		298.3337
69	Letzi	4		364.3318
70	Letzi	5		279.2994
71	Letzi	6		286.5737
72	Letzi	7		308.9928
73	Letzi	8		283.6495
74	Letzi	9		225.2611
75	Letzi	10		233.6316
76	Letzi	11		238.3103
121	Limmattal	1		286.6968
111	Limmattal	2		294.1908
112	Limmattal	3		232.4956
113	Limmattal	4		279.9036
114	Limmattal	5		275.3061
115	Limmattal	6		285.5990
116	Limmattal	7		253.4326
117	Limmattal	8		233.9377
118	Limmattal	9		185.5536
119	Limmattal	10		221.6835
120	Limmattal	11		199.1688
165	Schwamendingen	1		292.5198
155	Schwamendingen	2		299.8471
156	Schwamendingen	3		281.7500
157	Schwamendingen	4		258.7540
158	Schwamendingen	5		314.3609
159	Schwamendingen	6		265.1294
160	Schwamendingen	7		279.7506
161	Schwamendingen	8		282.6748
162	Schwamendingen	9		220.6027
163	Schwamendingen	10		248.8591
164	Schwamendingen	11		217.7453
209	Uto	1		374.8488
199	Uto	2		412.2149

200	Uto	3	360.2857
201	Uto	4	425.3738
202	Uto	5	372.4769
203	Uto	6	400.6184
204	Uto	7	394.7699
205	Uto	8	355.7803
206	Uto	9	306.3462
207	Uto	10	308.2171
208	Uto	11	251.7425
253	Waidberg	1	420.4739
243	Waidberg	2	440.3928
244	Waidberg	3	365.7735
245	Waidberg	4	394.3229
246	Waidberg	5	377.6390
247	Waidberg	6	364.5530
248	Waidberg	7	371.3762
249	Waidberg	8	385.0225
250	Waidberg	9	268.4890
251	Waidberg	10	255.0839
252	Waidberg	11	237.1414
297	Zuerichberg	1	268.4758
287	Zuerichberg	2	247.6600
288	Zuerichberg	3	291.5946
289	Zuerichberg	4	281.9063
290	Zuerichberg	5	280.9221
291	Zuerichberg	6	301.1948
292	Zuerichberg	7	270.9779
293	Zuerichberg	8	309.9675
294	Zuerichberg	9	159.2789
295	Zuerichberg	10	162.4556
296	Zuerichberg	11	126.2056

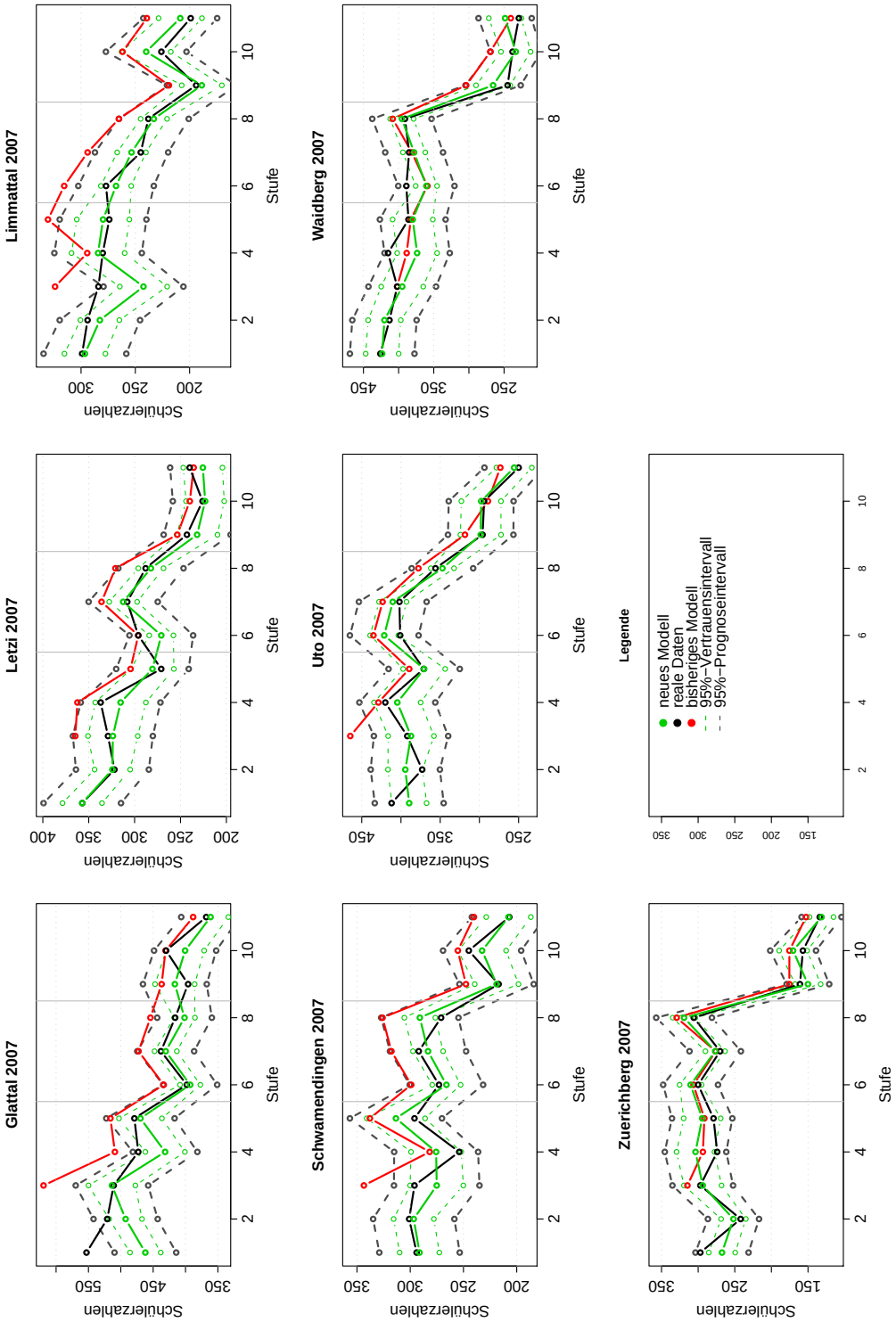

Abbildung B.15.: Ausgehend vom Jahr 2005 eine 2-Jahresprognose der Out-of-sample-Modelle für das Jahr 2007

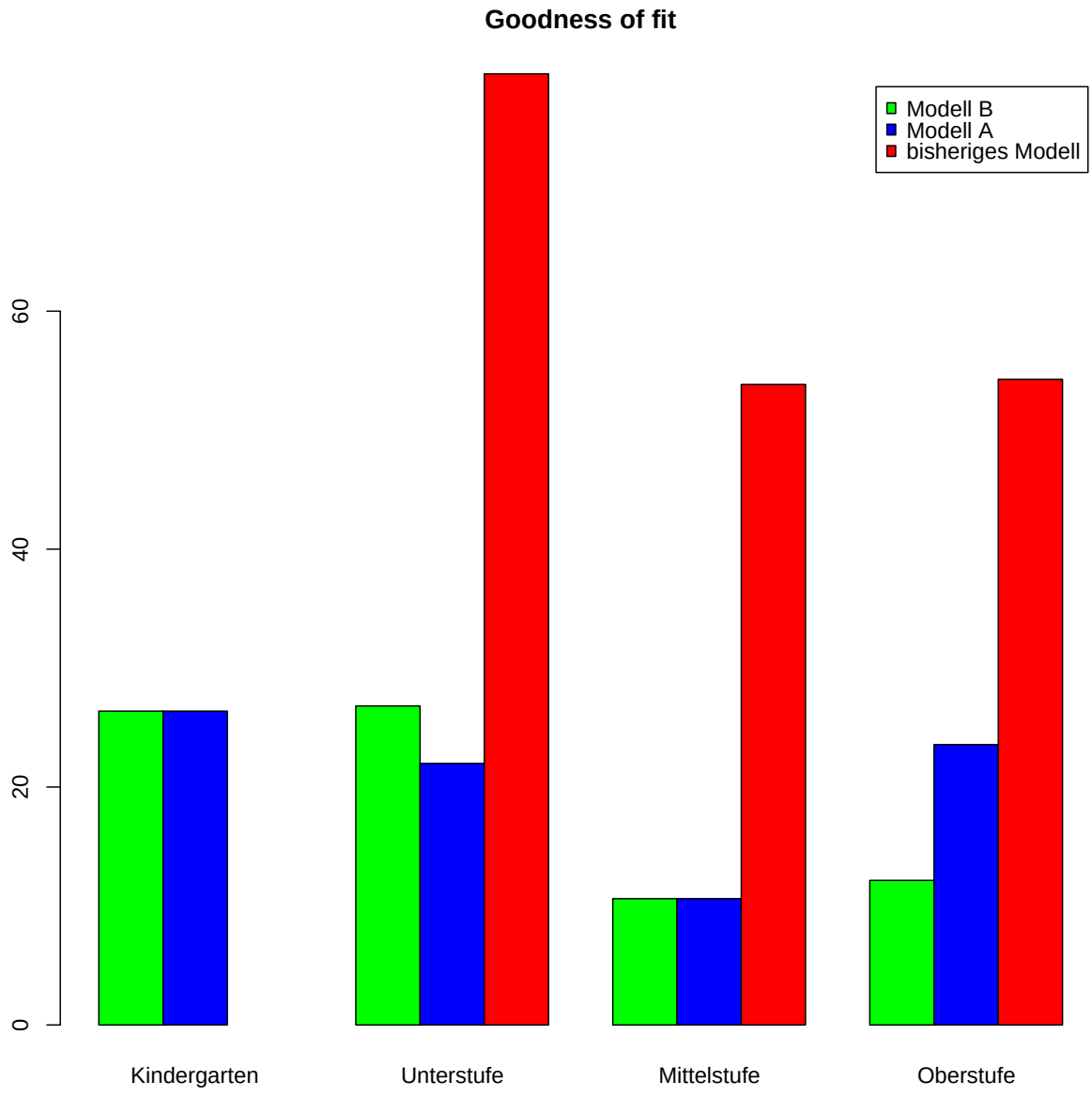

Abbildung B.16.: Goodness-of-fit der 2-Jahresprognosen für 2007: Prognose mit Modell aus Verfahren A (grün), Prognose mit Modell aus Verfahren B (blau), Prognose des Schul- und Sportdepartements (rot)

Die entsprechenden Daten zur Abbildung B.15

```
10    Glattal          1           462.0524
11    Glattal          2           492.5629
```

1	Glattal	3	513.6920
2	Glattal	4	431.5195
3	Glattal	5	469.6638
4	Glattal	6	392.7920
5	Glattal	7	430.6164
6	Glattal	8	401.5207
7	Glattal	9	416.4772
8	Glattal	10	400.6148
9	Glattal	11	361.2935
43	Letzi	1	357.0424
44	Letzi	2	324.2403
34	Letzi	3	323.6877
35	Letzi	4	315.2943
36	Letzi	5	280.5695
37	Letzi	6	270.9255
38	Letzi	7	312.4611
39	Letzi	8	282.2534
40	Letzi	9	231.9170
41	Letzi	10	223.1502
42	Letzi	11	225.6486
76	Limmattal	1	296.4399
77	Limmattal	2	282.6598
67	Limmattal	3	242.5000
68	Limmattal	4	284.3158
69	Limmattal	5	279.7263
70	Limmattal	6	267.7686
71	Limmattal	7	253.4876
72	Limmattal	8	232.9587
73	Limmattal	9	188.7723
74	Limmattal	10	239.9914
75	Limmattal	11	208.6122
109	Schwamendingen	1	291.2877
110	Schwamendingen	2	296.6890
100	Schwamendingen	3	275.1226
101	Schwamendingen	4	275.6710
102	Schwamendingen	5	313.4280
103	Schwamendingen	6	266.0041
104	Schwamendingen	7	283.2896
105	Schwamendingen	8	290.9720
106	Schwamendingen	9	218.9005
107	Schwamendingen	10	232.5141
108	Schwamendingen	11	207.6971
142	Uto	1	389.5204
143	Uto	2	394.5101
133	Uto	3	387.2117
134	Uto	4	404.6671
135	Uto	5	370.4548

136	Uto	6	421.3635
137	Uto	7	410.4832
138	Uto	8	347.1798
139	Uto	9	298.3588
140	Uto	10	297.7807
141	Uto	11	255.7262
175	Waidberg	1	423.3798
176	Waidberg	2	420.1494
166	Waidberg	3	394.8342
167	Waidberg	4	373.6462
168	Waidberg	5	379.9329
169	Waidberg	6	360.5355
170	Waidberg	7	377.9340
171	Waidberg	8	395.3325
172	Waidberg	9	265.5961
173	Waidberg	10	233.3566
174	Waidberg	11	248.5328
208	Zuerichberg	1	267.4343
209	Zuerichberg	2	251.7915
199	Zuerichberg	3	293.7211
200	Zuerichberg	4	303.5756
201	Zuerichberg	5	294.5000
202	Zuerichberg	6	310.3183
203	Zuerichberg	7	276.7169
204	Zuerichberg	8	319.2127
205	Zuerichberg	9	150.1342
206	Zuerichberg	10	170.3655
207	Zuerichberg	11	131.5739

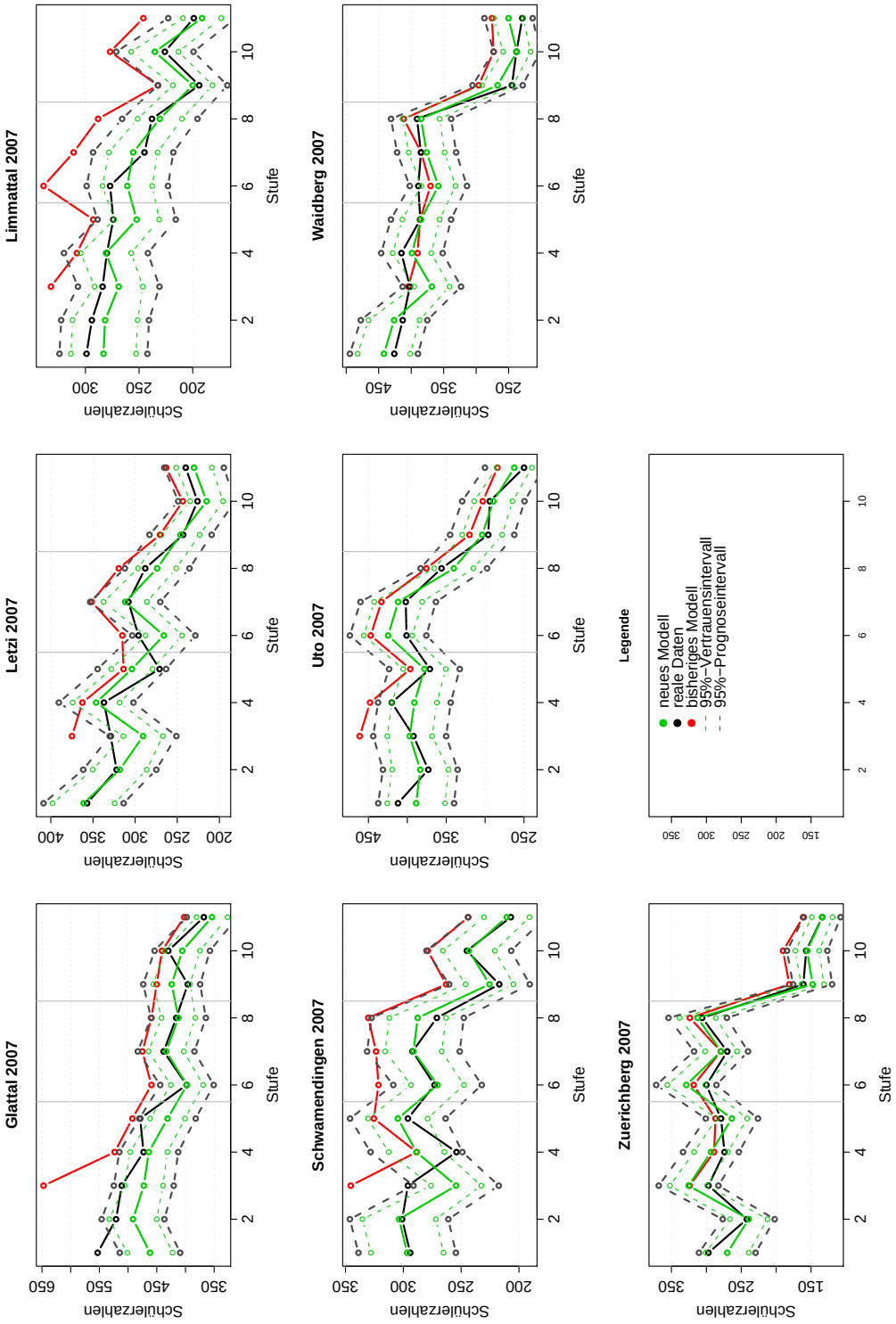

Abbildung B.17.: Ausgehend vom Jahr 2004 eine 3-Jahresprognose der Out-of-sample-Modelle für das Jahr 2007

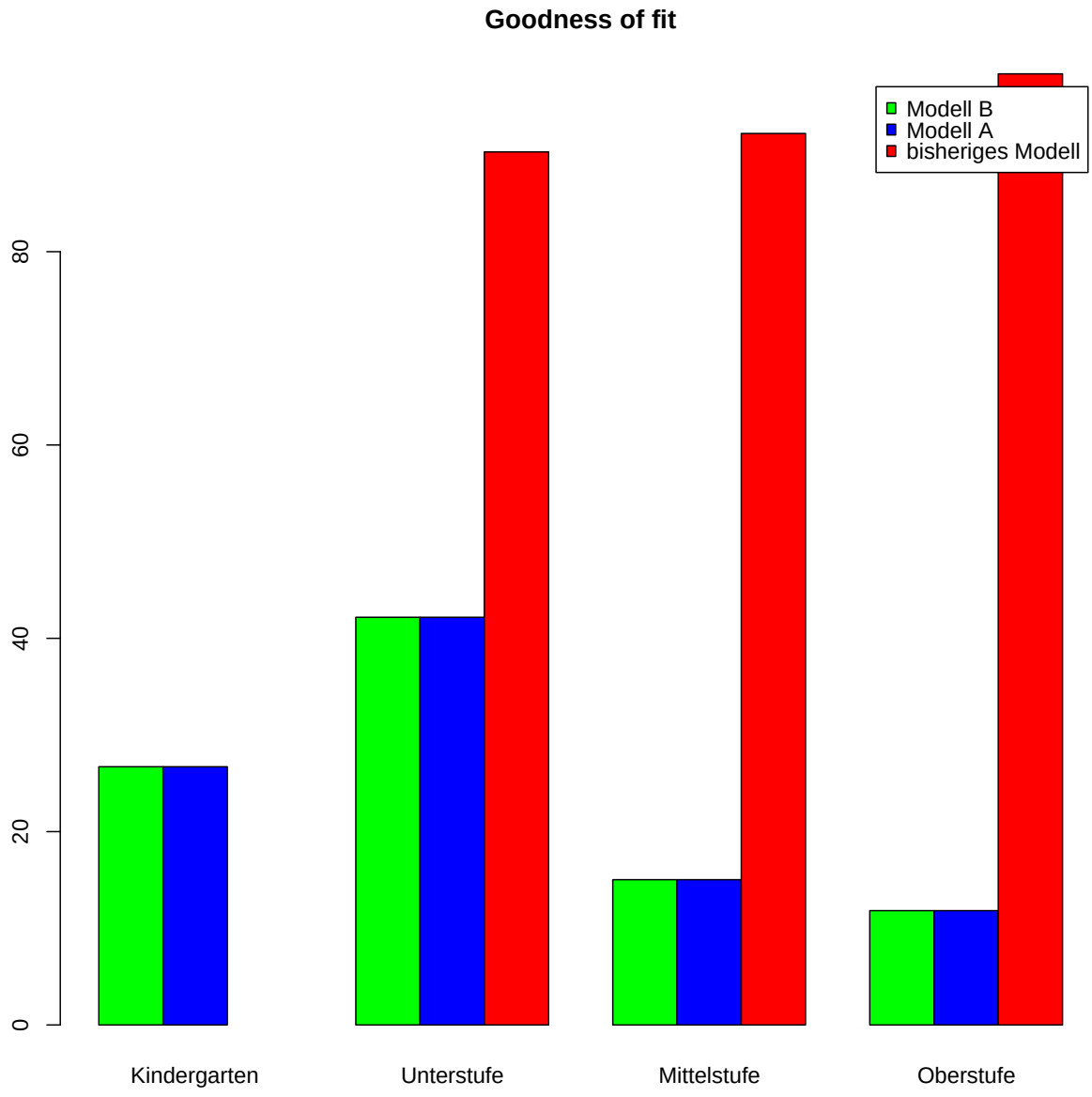

Abbildung B.18.: Goodness-of-fit der 3-Jahresprognosen für 2007: Prognose mit Modell aus Verfahren A (grün), Prognose mit Modell aus Verfahren B (blau), Prognose des Schul- und Sportdepartements (rot)

Die entsprechenden Daten zur Abbildung B.17

```
22    Glattal          1           461.8790
12    Glattal          2           491.4633
```

13	Glattal	3	472.6705
14	Glattal	4	463.9865
15	Glattal	5	431.2226
16	Glattal	6	397.4378
17	Glattal	7	433.7509
18	Glattal	8	412.2908
19	Glattal	9	424.0047
20	Glattal	10	405.5863
21	Glattal	11	353.7375
55	Letzi	1	361.1513
45	Letzi	2	318.2362
46	Letzi	3	290.6227
47	Letzi	4	346.2527
48	Letzi	5	303.8661
49	Letzi	6	265.9840
50	Letzi	7	311.7564
51	Letzi	8	273.9233
52	Letzi	9	246.1228
53	Letzi	10	215.1886
54	Letzi	11	230.1017
88	Limmattal	1	283.1763
78	Limmattal	2	281.6873
79	Limmattal	3	268.9908
80	Limmattal	4	280.8872
81	Limmattal	5	252.2473
82	Limmattal	6	260.9567
83	Limmattal	7	255.5011
84	Limmattal	8	230.6953
85	Limmattal	9	200.0645
86	Limmattal	10	235.2863
87	Limmattal	11	191.3468
121	Schwamendingen	1	296.6701
111	Schwamendingen	2	303.6243
112	Schwamendingen	3	254.3132
113	Schwamendingen	4	288.5740
114	Schwamendingen	5	304.7167
115	Schwamendingen	6	270.4269
116	Schwamendingen	7	291.2507
117	Schwamendingen	8	287.5688
118	Schwamendingen	9	225.4667
119	Schwamendingen	10	243.3107
120	Schwamendingen	11	210.5767
154	Uto	1	388.6325
144	Uto	2	383.3486
145	Uto	3	396.9244
146	Uto	4	390.9296
147	Uto	5	378.0056

148	Uto	6	424.6932
149	Uto	7	411.7362
150	Uto	8	340.2438
151	Uto	9	303.7458
152	Uto	10	289.4367
153	Uto	11	262.5326
187	Waidberg	1	441.9491
177	Waidberg	2	426.3549
178	Waidberg	3	368.2384
179	Waidberg	4	398.8484
180	Waidberg	5	384.8886
181	Waidberg	6	358.1539
182	Waidberg	7	376.2529
183	Waidberg	8	384.7891
184	Waidberg	9	267.0393
185	Waidberg	10	237.4150
186	Waidberg	11	250.2411
220	Zuerichberg	1	270.0341
210	Zuerichberg	2	239.3524
211	Zuerichberg	3	325.4145
212	Zuerichberg	4	293.4131
213	Zuerichberg	5	263.4437
214	Zuerichberg	6	328.7040
215	Zuerichberg	7	278.9636
216	Zuerichberg	8	312.4061
217	Zuerichberg	9	147.8285
218	Zuerichberg	10	155.6195
219	Zuerichberg	11	134.0300

B.3.5. Zu- und Abnahme der Schülerzahlen

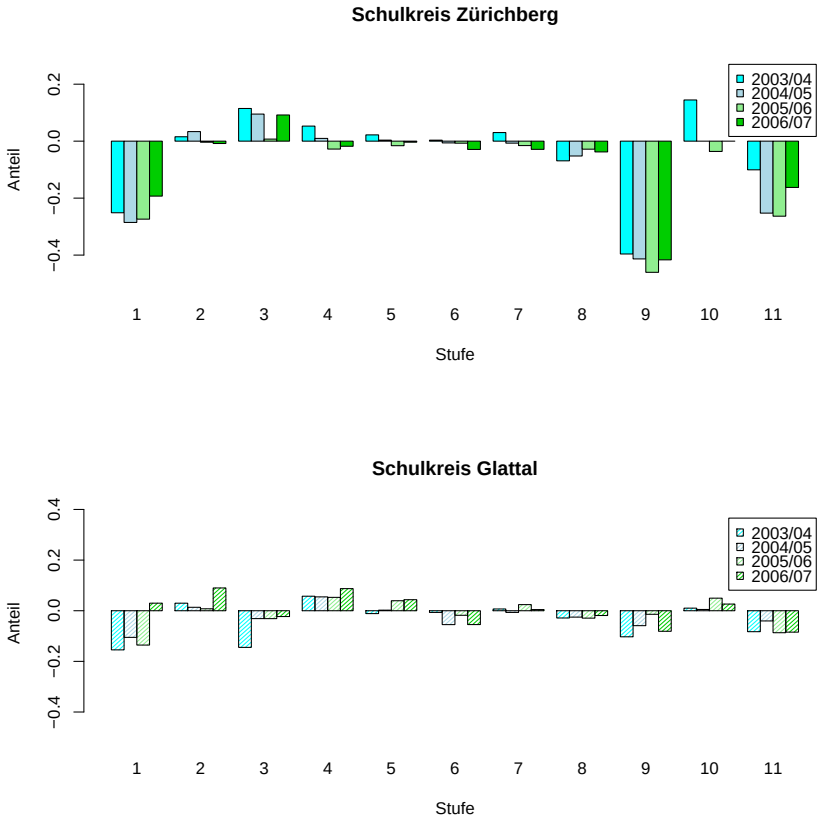

Abbildung B.19.: Zu- und Abnahme der Schülerzahlen pro Stufe im Schulkreis Zürichberg (oben) und Glattal (unten). Prozentualer Anteil der Differenz zwischen Anzahl Schüler in der Stufe (k-1) im Jahr (t-1) und der Anzahl Schüler in der Stufe k im Jahr t gegenüber der Stufe k und dem Jahr t, siehe auch FormelB.5.

$$Anteil = \frac{Y_{ikt,j} - Y_{i(k-1)(t-1),j)}}{Y_{i(k-1)(t-1),j)}} \tag{B.5}$$