Statistical Consulting

Fallstudie: Measuring Quality Time

Evgeniya Ashmarina, Stefan Englert

17. Juni 2008

Inhaltsübersicht

Überblick

Methoden	Zeitreihenanalyse	ARIMA Modelle
Daten	Vier Jahre monatlicher Daten	Zwei Zeitreihen

Vorwort

Unsere Fallstudie beschreibt eine Form der Qualitätskontrolle.

Bei dieser Fallstudie hat eine Firma einen Index für die Qualität ihrer Produkte entwickelt, den IQ-Wert. Die genaue Entstehung dieses Wertes ist uns unbekannt, wir besitzen lediglich die entstandenen Daten und wissen, dass der IQ-Wert zwischen 0 und 100 liegt, wobei 100 den höchst möglichen Wert repräsentiert.

Die Firma hat diesen IQ-Wert für die Dauer von vier Jahren bestimmt und diese liegen uns als Datenmaterial vor.

Daten Teil 1

Monat	Stückzahl	IQ Wert	Monat	Stückzahl	IQ Wert
jan94	2339	86.63	jan96	2971	89.25
feb94	2275	84.60	feb96	3083	90.54
mar94	2881	87.04	mar96	3504	89.89
apr94	2780	87.19	apr96	3580	90.28
may94	3227	87.91	may96	3855	89.46
jun94	3291	87.99	jun96	3894	89.42
jul94	2944	88.09	jul96	3772	89.28
aug94	3163	88.25	aug96	3705	89.17
sep94	2770	87.62	sep96	3364	90.42
oct94	2827	87.43	oct96	3341	90.46
nov94	2392	86.74	nov96	2680	88.63
dec94	1973	84.86	dec96	2418	89.74

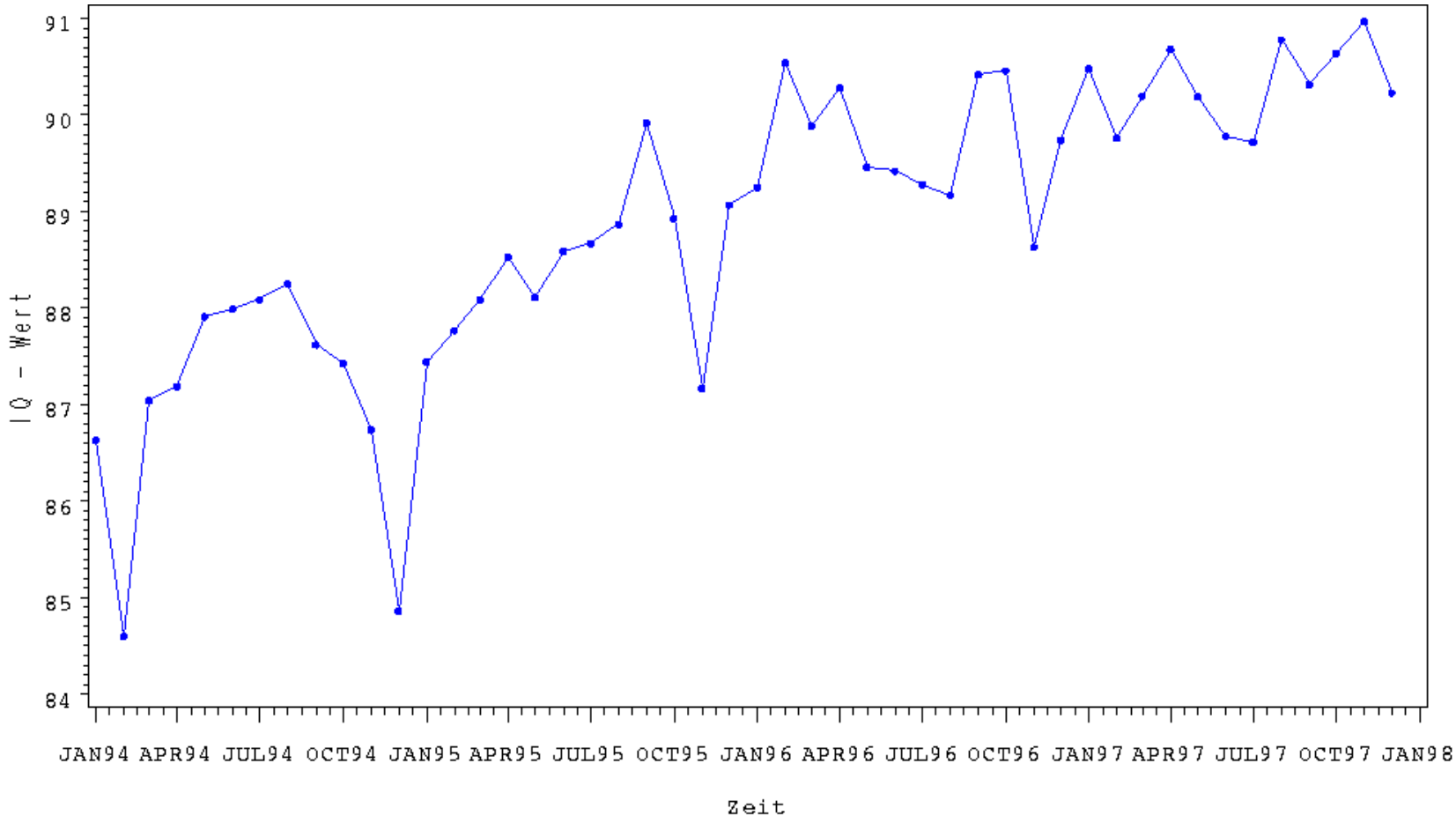
Daten Teil 2

Monat	Stückzahl	IQ Wert	Monat	Stückzahl	IQ Wert
jan95	3006	87.44	jan97	2963	90.48
feb95	2924	87.77	feb97	2890	89.76
mar95	3592	88.09	mar97	3455	90.20
apr95	3460	88.53	apr97	3747	90.68
may95	3807	88.11	may97	3685	90.19
jun95	3753	88.59	jun97	3672	89.78
jul95	3648	88.67	jul97	3865	89.72
aug95	3698	88.87	aug97	3729	90.78
sep95	3166	89.92	sep97	3205	90.32
oct95	3159	88.93	oct97	3158	90.64
nov95	2545	87.17	nov97	2552	90.97
dec95	2208	89.07	dec97	2135	90.23

Erster Überblick

- Es handelt sich um vier Jahre monatlicher Daten
- Insgesamt liegen uns 48 Beobachtungen vor
- Der Datensatz enthält keine fehlenden Werte
- Für jeden Messpunkt ist jeweils der Monat mitsamt Jahr, die gefertigte Stückzahl des Produktes und der durch die Firma bestimmte IQ-Wert vorhanden
- Der IQ-Wert wurde immer mit der gleichen Methode und gleicher Skala bestimmt

Plot


91
90
89
88
87
86
85
84
IQ - Wert
JAN94 APR94 JUL94 OCT94 JAN95 APR95 JUL95 OCT95 JAN96 APR96 JUL96 OCT96 JAN97 APR97 JUL97 OCT97 JAN98
Zeit

Die Daten erfüllen die Eigenschaften einer Zeitreihe:

1. Es handelt sich um Beobachtungen einer Zufallsvariablen an aufeinander folgenden Zeitpunkten.
2. Die Zeitpunkte haben den gleichen Abstand.
3. Die Werte wurden durch die gleiche Methode bestimmt.

Wir können also die Methoden der Zeitreihenanalyse verwenden, um unsere Daten zu analysieren.

Zielsetzung

Ziel dieser Analyse ist es ein gutes Zeitreihenmodell zur Modelierung der IQ-Werte zu finden und weitere Werte vorherzusagen.

Methoden

Sei y_t der Zeitreihenwerte zum Zeitpunkt t. In unserem Fall ist das Zeitintervall jeweils ein Monat und wir möchten den IQ-Wert $= y_t$ vorhersagen.

In einem Regressionsmodell könnten wir X als Zeitindex und Y als die Antwortvariable betrachten. Wir könnten dann ein geeignetes Regressionsmodell wählen um $\hat{y}_{t+1}$ zu erhalten.

Hierfür würden wir aber benötigen, dass die Fehler des Modells unabhängig sind. Diese Annahme ist aber nicht sinnvoll für einen Prozess, der sich über die Zeit entwickelt. Deshalb verwenden wir ARIMA-Modelle.

ARIMA Modelle

Die ARIMA Modelle wurden durch Box und Jenkins (1970) populär. Ihre Herangehensweise wird auch als Box–Jenkins Methode beschrieben und schließt die folgenden drei Schritte ein:

Identification Die Reihe y_t wird durch Differenzbildung stationär gemacht.

Estimation Ein $ARMA(p, q)$ Modell wird an die so entstandene Reihe angepasst.

Forecasting Das ARIMA Modell wird angewendet um zukünftige Werte y_t vorherzusagen.

Identification

Zuerst müssen wir die Saisonalität und den Trend aus der Reihe entfernen, um diese stationär zu machen.

In der Praxis geschieht dies durch Differenzbildung.

$$\Delta\, y_t = y_t - y_{t-1} \quad \Delta_s\, y_t = y_t - y_{t-s}$$

Dieses Verfahren wenden wir solange an, bis die entstandene Serie w_t stationär ist, d.h. dass der Erwartungswert und die Varianz der w_t ungefähr konstant ist.

Zuerst sollte eine einfache Differenz auf die Daten angewendet werden, außerdem sollte man nicht mehr als zwei einfache Differenzen ($\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$) anwenden, da sonst das $ARMA(p, q)$ Modell instabil wird.

Falls notwendig kann man auch eine saisonale Differenz anwenden. So könnte man mit $s = 12$ eine jährliche Saisonalität aus monatlichen Daten entfernen.

$$\Delta_{12}\, y_t = y_t - y_{t-12}$$

Estimation

Die w_t werden dann durch ein $ARMA(p, q)$ Modell der Form

$$w_t + \sum_{j=1}^{p} \alpha_j w_{t-j} = \epsilon_t + \sum_{j=1}^{q} \beta_j \epsilon_{t-j},$$

modelliert, wobei die ϵ_t unbekannte zufällige Komponenten sind, die als unabhängig und identisch verteilt angenommen werden.

Zuerst müssen die Ordnungen p, q des *ARMA* Modells festgelegt werden. Diese können durch Untersuchung der Korellogramme für die Autocorrelation (ACF) und der partiellen Autocorrelation (PACF) bestimmt werden, wie wir gleich sehen werden.

Durch die zwei Informationskriterien *AIC* und *BIC*, die wir später noch genauer vorstellen werden, kann dabei das beste Modell gewählt werden.

Nachdem man die Ordnungen p, q festgelegt hat, kann man die Parameter des *ARMA*–Modells schätzen.

ACF

Das Autocorrelationskorrelogramm (ACF) ist eine graphische Darstellung von $\rho_k = Corr(y_{t-k}, y_t)$ gegen k.

Für einen $AR(p)$ Prozess nimmt die Korrelation zwischen y_t und y_{t-k} exponentiell mit wachsendem Zeitabstand k ab.

Im Gegensatz dazu ist $\rho_k = 0$ für alle $k > q$ ein reiner $MA(q)$ Prozess.

ACF eines AR(1) Prozesses

Lag	Covariance	Correlation	Autocorrelations
0	9.328577	1.00000	|**********|
1	4.404094	0.47211	|*********. |
2	1.719905	0.18437	|****. |
3	1.067388	0.11442	|**. |
4	0.885341	0.09491	|**. |
5	0.682238	0.07313	|*. |
6	0.375455	0.04025	|*. |
7	-0.082802	-.00888	| . |
8	-0.136101	-.01459	| . |
9	-0.336445	-.03607	| *. |
10	-0.068452	-.00734	| . |
11	0.557274	0.05974	|*. |
12	0.159944	0.01715	| . |
13	-0.168594	-.01807	| . |
14	-0.624676	-.06696	| *. |
15	-0.300576	-.03222	| *. |
16	0.015192	0.00163	| . |
17	-0.305454	-.03274	| *. |
18	-0.233878	-.02507	| *. |
19	-0.443853	-.04758	| *. |
20	-0.856874	-.09185	| **. |

"." marks two standard errors

PACF

Die partielle Autocorrelation zwischen y_t und y_{t-k} ist die Korrelation nachdem die Effekte von $y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}$ durch eine Regression entfernt wurden. Dabei sind ϕ_k die Regressionskoeffizienten.

Für einen $AR(p)$ Prozess ist $\phi_k = 0$ für $k > p$, d.h. es gibt keine „Verbindung“ mehr zwischen y_t und y_{t-k} für $k > p$.

Ein $MA(q)$ Prozess kann dabei repräsentiert werden durch einen $AR(\infty)$ Prozess; die ϕ_k nehmen dann exponentiell ab.

PACF ist ein Plot von ϕ_k gegen k.

PACF eines AR(1) Prozesses

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.47211	| . |********** |
2	-0.04956	| . * | . |
3	0.05994	| . |* . |
4	0.03131	| . |* . |
5	0.01533	| . | . |
6	-0.00713	| . | . |
7	-0.03929	| . * | . |
8	0.00203	| . | . |
9	-0.03851	| . * | . |
10	0.03311	| . |* . |
11	0.07273	| . |* . |
12	-0.04956	| . * | . |
13	-0.01246	| . | . |
14	-0.06943	| . * | . |
15	0.03182	| . |* . |
16	0.00836	| . | . |
17	-0.04655	| . * | . |
18	0.02404	| . | . |
19	-0.04947	| . * | . |
20	-0.05700	| . * | . |

AIC/BIC

Durch ACF und PACF Korrelogramme kann die Ordnung reiner *AR* oder *MA* Prozesse sehr gut bestimmt werden.

Durch Kombination können die Ordnungen eines *ARMA* Modells angegeben werden. Bei dieser Entscheidung helfen die Informationskriterien AIC und BIC.

Diese sind definiert als

$$AIC(k) = \log \hat{\sigma}_k^2 + \frac{2k}{n} \quad BIC(k) = \log \hat{\sigma}_k^2 + \frac{k \ \log n}{n},$$

wobei σ_k^2 die Varianz der Residuen bezeichnet.

Das beste Modell besitzt dabei die minimalen Werte für AIC und BIC.

Forecasting

Zur Vorhersage weiterer Werte muss das „integrierte“ Modell aus dem ARMA Modell und den im Identificationsschritt angewendeten Differenzen rekonstruiert werden, um ein lineares Modell der ursprünglichen Zeitreihe y_t zu erhalten.

Ein $ARIMA(1,1,1)$ Modell, das auf einem $ARMA(1,1)$ Modell zusammen mit einfachen Differenzen beruht, hat beispielsweise die Form:

$$y_t - y_{t-1} + \alpha(y_{t-1} - y_{t-2}) = \epsilon_t + \beta_1 \epsilon_{t-1}$$

Saisonale ARIMA Modelle

ARIMA Modelle können auch saisonale Effekte berücksichtigen.

Zuerst passt man an die Daten ein $ARMA(p, q)$-Modell an und wendet auf dieses geschätzte Modell erneut ein saisonales $ARMA(P, Q)$-Modell mit saisonalem Lag s an.

Die saisonalen ARIMA Modelle werden oft als $(P, D, Q)_s \times (p, d, q)$ geschrieben. Dabei ist

D = Ordnung der saisonalen Differenzen

P,Q = Ordnungen des saisonalen ARMA-Modells

s = Saisonale Lag

Beispiel

Ein saisonales ARIMA Modell $(1,0,0)_{12} \times (1,1,0)$ hat die Form:

$$w_t - \alpha\, w_{t-1} - \beta\,(w_{t-12} - \alpha\, w_{t-13}) = \epsilon_t$$

wobei

$$w_t = y_t - y_{t-1}$$

eine einfache Differenz ist.

ARIMA Modelle können oft das geeignete Mittel zur Untersuchung einer Zeitreihe sein.

Sie benötigen jedoch, dass die Originalreihe (y_t) auf eine stationäre Reihe (w_t) zurückgeführt werden kann, in der Mittelwert und Varianz konstant sind.

Mittelwertsstationarität kann im Allgemeinen durch Differenzbildung erreicht werden. Für eine konstante Varianz wird manchmal eine nichtlineare Transformation benötigt.

Statistical Consulting

Fallstudie: Measuring Quality Time

Evgeniya Ashmarina, Stefan Englert

17. Juni 2008

Wiederholung

Bei dieser Fallstudie hat eine Firma einen Index für die Qualität ihrer Produkte entwickelt, den IQ-Wert.

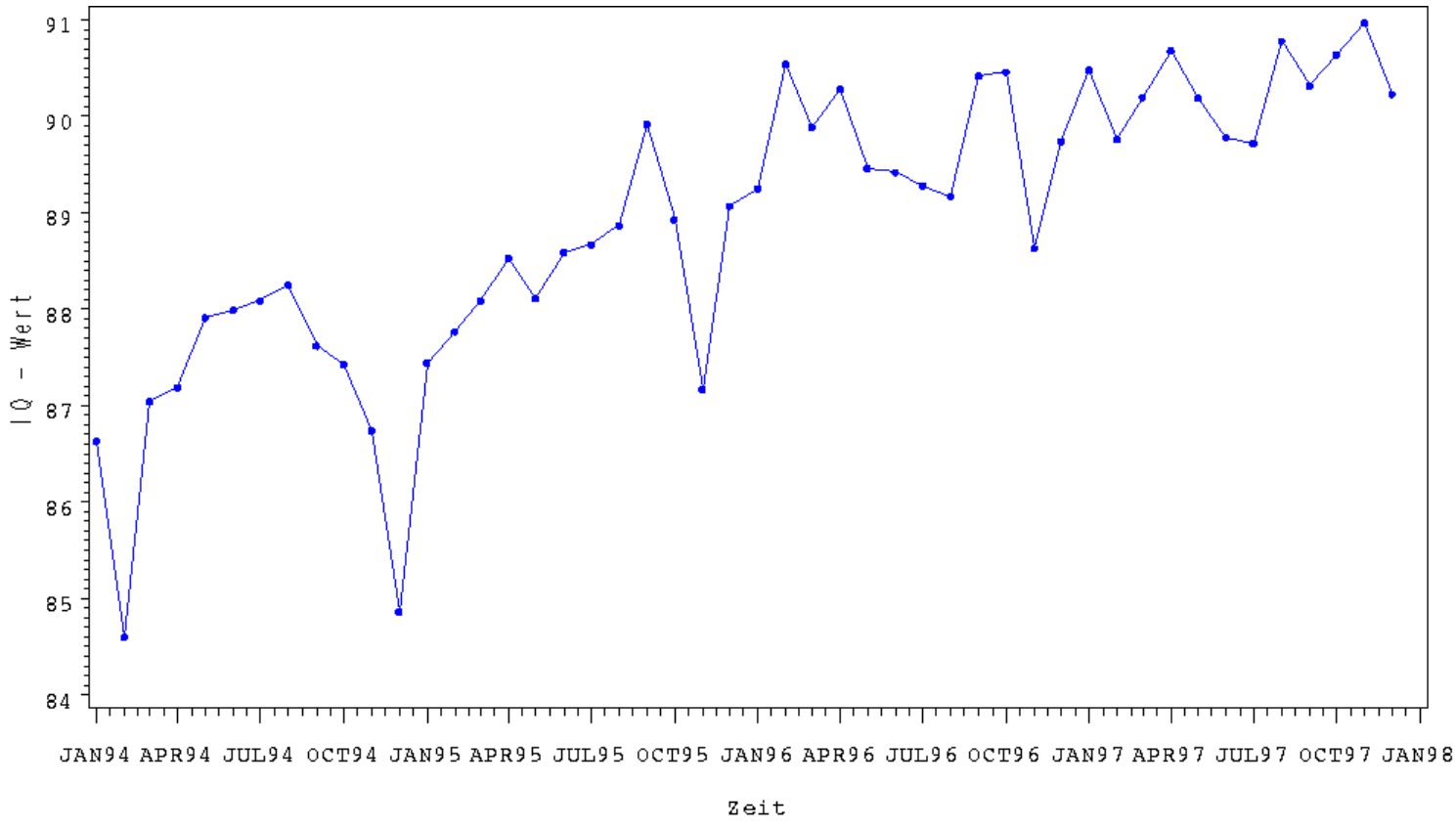
Die Firma hat diesen IQ-Wert für die Dauer von vier Jahren bestimmt und diese liegen uns als Datenmaterial vor.

Ziel dieser Analyse ist es ein gutes Zeitreihenmodell zur Modelierung der IQ-Werte zu finden und weitere Werte vorherzusagen.

Wiederholung

- Wir besitzen vier Jahre monatlicher Daten ohne fehlende Werte.
- Die Daten erfüllen die Eigenschaften einer Zeitreihe.
- Wir können die Methoden der Zeitreihenanalyse verwenden, um unsere Daten zu analysieren.

Analyse der Daten



Die Daten zeigen eine jährliche Saisonalität und einen Trend nach oben, der in unserem Fallbeispiel sicher gewünscht ist.

Trendmodell

Beim *klassischen Komponentenmodell* der Zeitreihenanalyse zieht man vier Komponenten in Betracht:

Den Trend T_t, die (mittelfristige) Konjunktur K_t, die (regelmäßige) Saison S_t und die zufällige Streuung ϵ_t.

Da wir nur vier Jahre Datenmaterial zur Verfügung haben, verzichten wir auf eine Konjunktur.

Für ein *additives Modell* lautet dann die Zusammensetzung:

$$X_t = T_t + S_t + \epsilon_t$$

Im Falle eines additiven Modells wird die saisonale Schwankung einfach aufaddiert, d.h. die Amplitude der Schwankung bleibt konstant.
Dies ist bei unseren Daten nicht der Fall, weshalb wir zu einem multiplikativen Modell übergehen, bei dem sich die Amplitude mit dem Trend verändert.

$$X_t = T_t \cdot S_t \cdot \epsilon_t$$

Dieses Modell lässt sich durch eine Logarithmustransformation $X_t \mapsto \ln X_t$ in ein additives Modell transformieren.

$$\ln X_t = \ln T_t + \ln S_t + \ln \epsilon_t$$

Mit diesem Modell werden wir im folgenden Arbeiten, d.h. wir werden ein additives Modell an die logarithmierten Ausgangsdaten anpassen.

Wir führen nun die zuvor genannten Schritte an unseren (logarithmierten) Daten durch. Diese waren im einzelnen:

Identification Die Reihe y_t wird durch Differenzbildung stationär gemacht.

Estimation Ein $ARMA(p, q)$ Modell wird an die so entstandene Reihe angepasst.

Forecasting Das ARIMA Modell wird angewendet um zukünftige Werte y_t vorherzusagen.

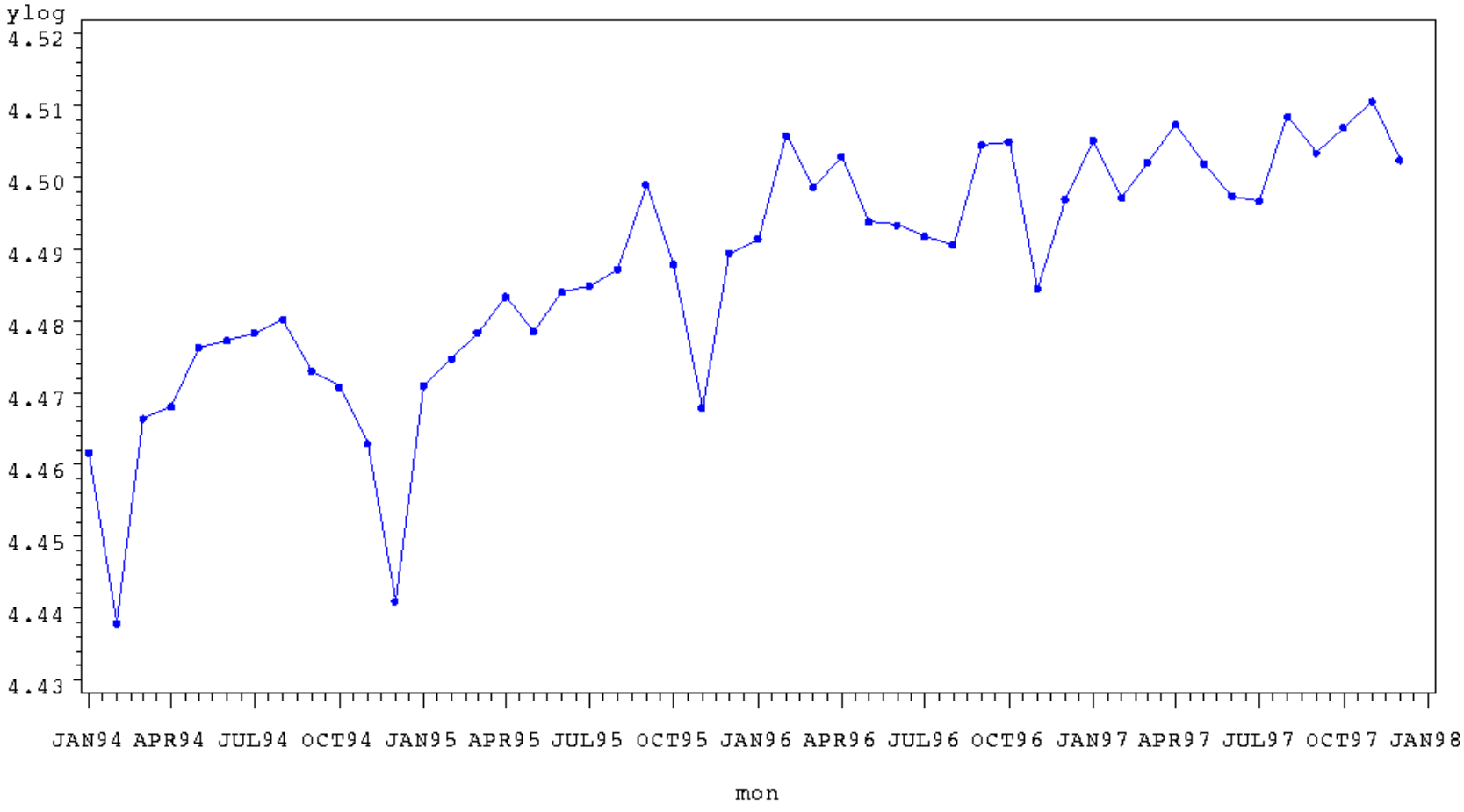
Identification

Zuerst müssen wir aus der Reihe y_t eine stationäre Reihe machen. In unserem Beispiel bieten sich die folgenden Differenzbildungen an:

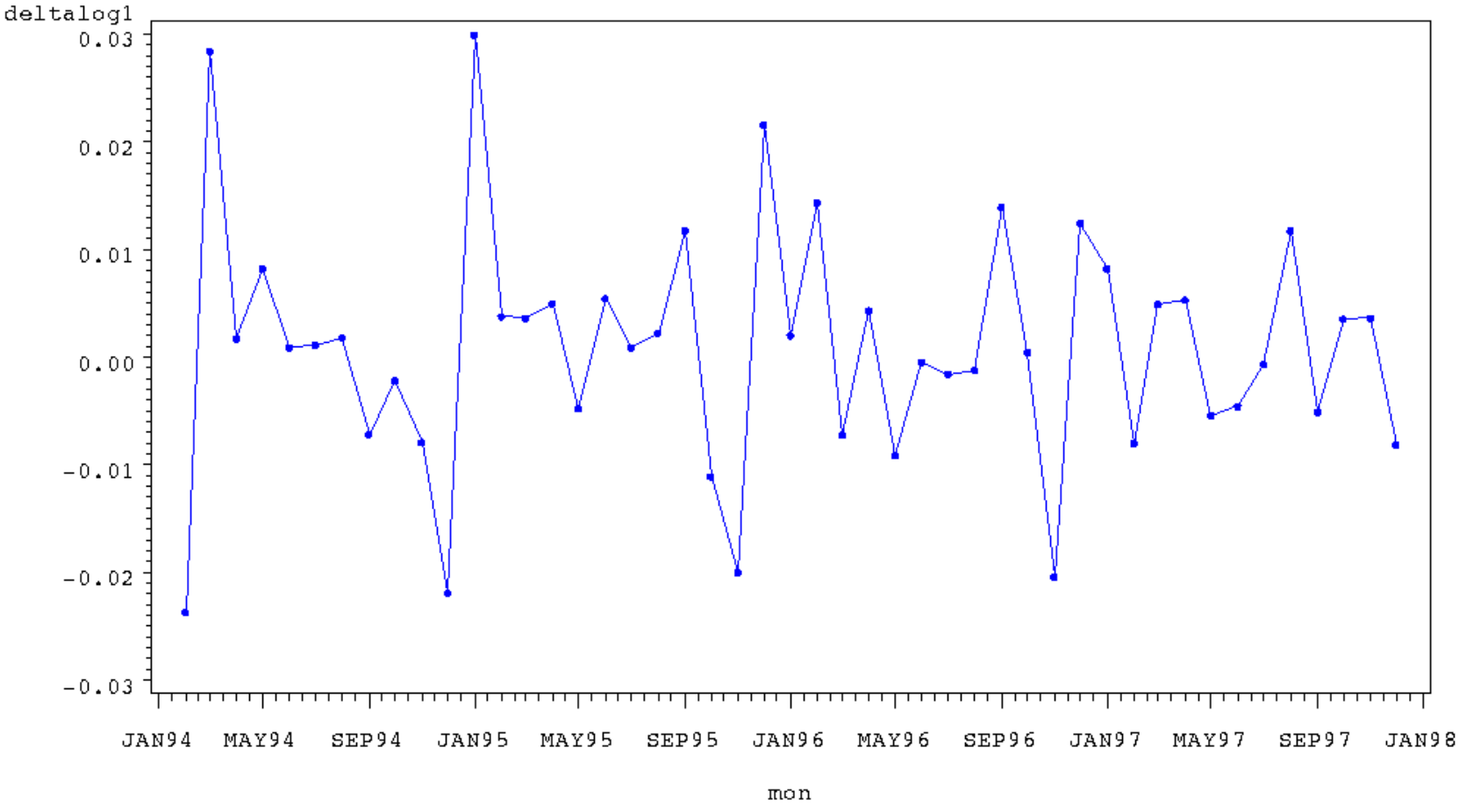
- y
- $y(1) := y_t - y_{t-1}$
- $y(12) := y_t - y_{t-12}$
- $y(1,12) := (y_t - y_{t-1}) - (y_{t-12} - y_{t-13})$

Wir plotten im folgenden die dazugehörigen Diagramme.

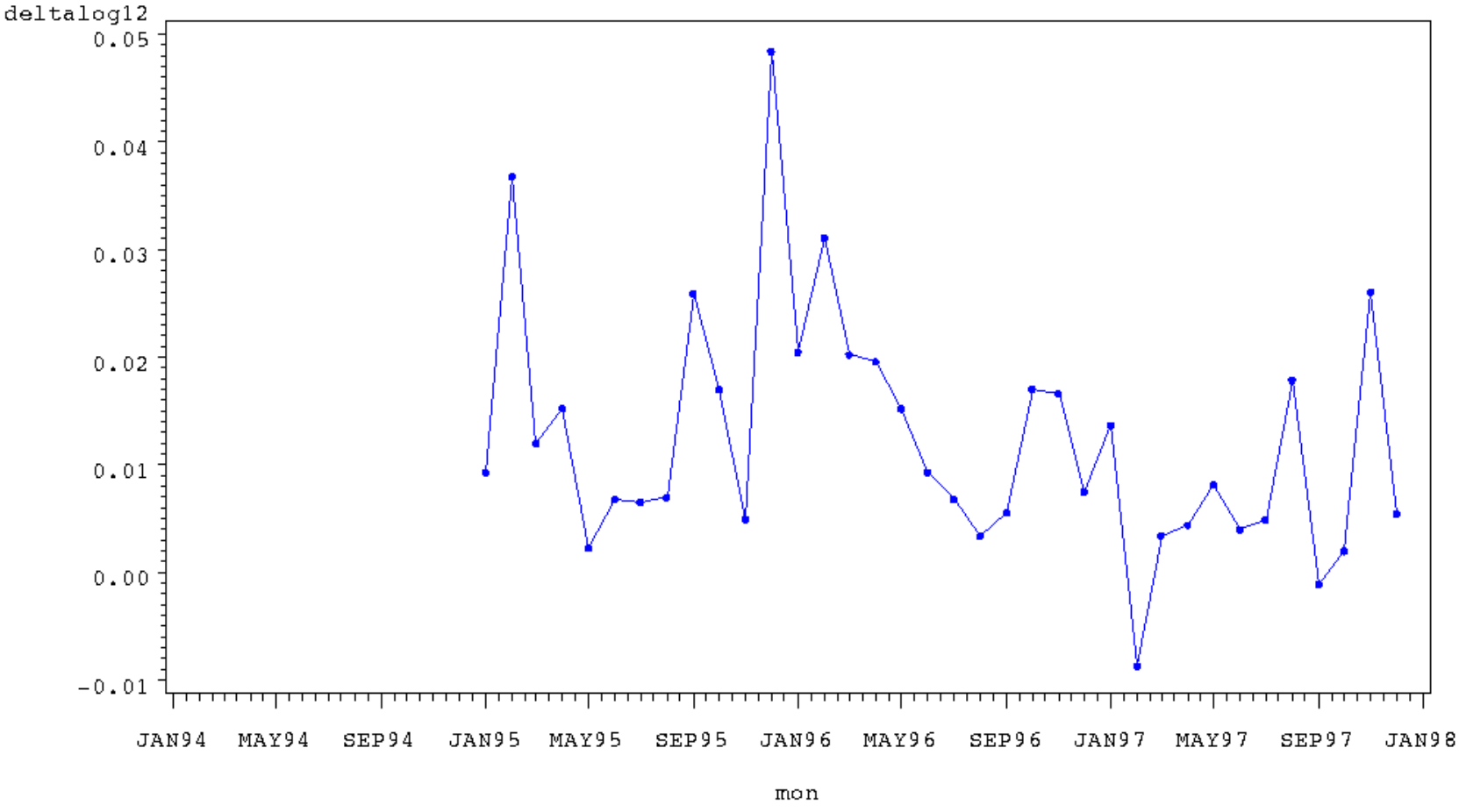
y


ylog
4.52
4.51
4.50
4.49
4.48
4.47
4.46
4.45
4.44
4.43
JAN94 APR94 JUL94 OCT94 JAN95 APR95 JUL95 OCT95 JAN96 APR96 JUL96 OCT96 JAN97 APR97 JUL97 OCT97 JAN98
mon

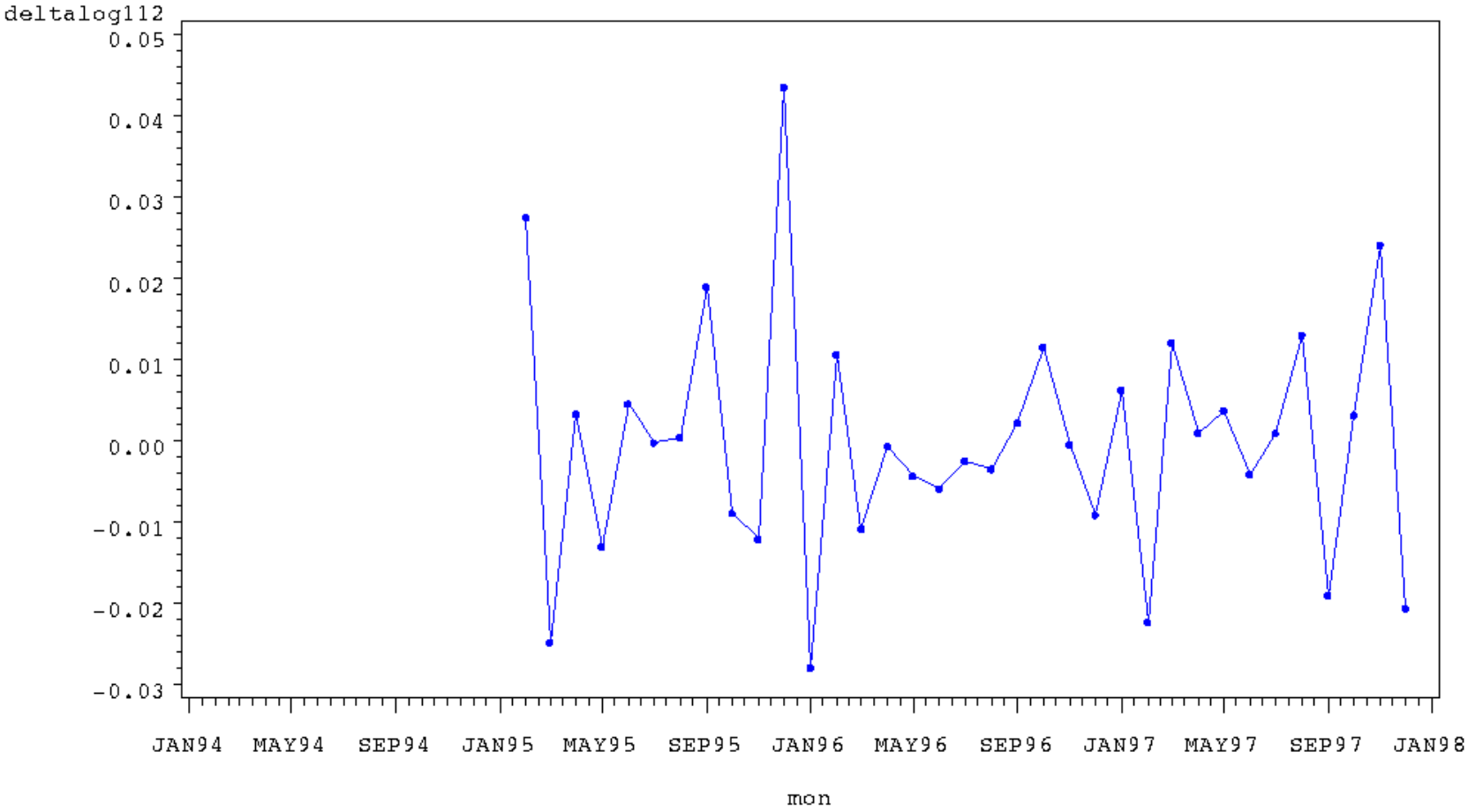
y(1)


deltalog1
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
JAN94
MAY94
SEP94
JAN95
MAY95
SEP95
JAN96
MAY96
SEP96
JAN97
MAY97
SEP97
JAN98
mon

y(12)


deltalog12
0.05
0.04
0.03
0.02
0.01
0.00
-0.01
JAN94
MAY94
SEP94
JAN95
MAY95
SEP95
JAN96
MAY96
SEP96
JAN97
MAY97
SEP97
JAN98
mon

y(1,12)


deltalog112
0.05
0.04
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
JAN94
MAY94
SEP94
JAN95
MAY95
SEP95
JAN96
MAY96
SEP96
JAN97
MAY97
SEP97
JAN98
mon

SAS bietet uns außerdem die Möglichkeit eines Stationaritätstest (augmented Dickey-Fuller).

Zusammen mit diesen Informationen entscheiden wir uns für:

$$y(1) := y_t - y_{t-1}$$

Die immer noch enthaltene saisonale Komponente werden wir durch ein geeignetes saisonales ARIMA-Modell herausmodellieren.

Estimation

Als nächstes müssen wir die Ordnungen p und q des $ARMA(p, q)$ Modells bestimmen.

Dazu verfahren wir wie zuvor erläutert und betrachten die ACF und PACF–Korrelogramme.

ACF–Korrelogramm

Lag	Covariance	Correlation	Autocorrelations
0	0.00012271	1.00000	|**********|**********|
1	-0.0000380	-.30991	|******| . |
2	-0.0000145	-.11783	| . **| . |
3	2.62728E-6	0.02141	| . | . |
4	-0.0000142	-.11558	| . **| . |
5	-2.2065E-6	-.01798	| . | . |
6	-6.1827E-7	-.00504	| . | . |
7	4.53585E-6	0.03696	| . |* . |
8	9.30111E-6	0.07580	| . |** . |
9	-0.0000317	-.25869	| . *****| . |
10	0.00002348	0.19134	| . |**** . |
11	2.04807E-6	0.01669	| . | . |
12	5.37765E-6	0.04382	| . |* . |
13	0.00001707	0.13912	| . |*** . |
14	-0.0000274	-.22313	| . ****| . |
15	5.98523E-6	0.04877	| . |* . |

"." marks two standard errors

PACF–Korrelogramm

Partial Autocorrelations

Lag	Correlation	−1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	−0.30991	| . . . . ****** | . . . . |
2	−0.23659	| . . . .***** | . . . . |
3	−0.11375	| . . . . ** | . . . . |
4	−0.20784	| . . . . **** | . . . . |
5	−0.18454	| . . . . **** | . . . . |
6	−0.18292	| . . . . **** | . . . . |
7	−0.11961	| . . . . ** | . . . . |
8	−0.02384	| . . . . | . . . . |
9	−0.35887	| . . . . ******* | . . . . |
10	−0.12094	| . . . . ** | . . . . |
11	−0.13610	| . . . . *** | . . . . |
12	−0.00663	| . . . . | . . . . |
13	0.16086	| . . . . |*** . . . . |
14	−0.09341	| . . . . ** | . . . . |
15	0.03780	| . . . . |* . . . . |

Wir können durch diese Korrelogramme unsere Auswahl auf $p = 1$ für ein $AR(1)$–Modell, $q = 1$ für ein $MA(1)$–Modell und $p = 1, q = 1$ für ein $ARMA(1, 1)$–Modell einschränken.

Durch die Werte der Informationskriterien AIC und BIC entscheiden wir uns für das $AR(1)$–Modell, das auch als $ARMA(1, 0)$–Modell bezeichnet werden kann.

Da das PACF-Korrelogramm bei Lag 9 einen erhöhten Wert zeigt, entscheiden wir uns für ein saisonales ARIMA-Modell mit Lag 9, i.Z. $(1, 0, 0)_9 \times (1, 1, 0)$.

Anmerkung: SAS bezeichnet das BIC Informationskriterium abweichend als SBC Informationskriterium.

Schätzung der Parameter

The ARIMA Procedure

Unconditional Least Squares Estimation

Parameter	Schätzwert	Standardfehler	t-Wert	Approx Pr > |t|	Lag
AR1,1	-0.29247	0.14389	-2.03	0.0480	1
AR2,1	-0.30628	0.14699	-2.08	0.0429	9

Variance Estimate	0.000109
Std Error Estimate	0.010426
AIC	-292.649
SBC	-288.949
Number of Residuals	47

Correlations of Parameter Estimates

Parameter	AR1,1	AR2,1
AR1,1	1.000	-0.015
AR2,1	-0.015	1.000

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	0.00010871	1.00000	|**********|
1	-2.4497E-6	-.02253	| . | . |
2	-0.0000215	-.19814	| . **** | . |
3	1.18881E-6	0.01094	| . | . |
4	-0.0000126	-.11596	| . ** | . |
5	-0.0000164	-.15090	| . *** | . |
6	-1.0043E-7	-.00092	| . | . |
7	7.83337E-7	0.00721	| . | . |
8	-1.8487E-6	-.01701	| . | . |
9	9.87277E-6	0.09082	| . |** . |
10	0.00001663	0.15294	| . |*** . |
11	0.00001213	0.11154	| . |** . |
12	0.00001719	0.15810	| . |*** . |
13	4.30402E-6	0.03959	| . |* . |
14	-0.0000195	-.17958	| . **** | . |
15	-4.7326E-6	-.04353	| . * | . |

"." marks two standard errors

Autocorrelation Check of Residuals (Portmanteau-test)

To Lag	Chi-Square	DF	Pr > ChiSq
6	4.01	4	0.4047
12	8.43	10	0.5873
18	11.66	16	0.7673
24	19.66	22	0.6041

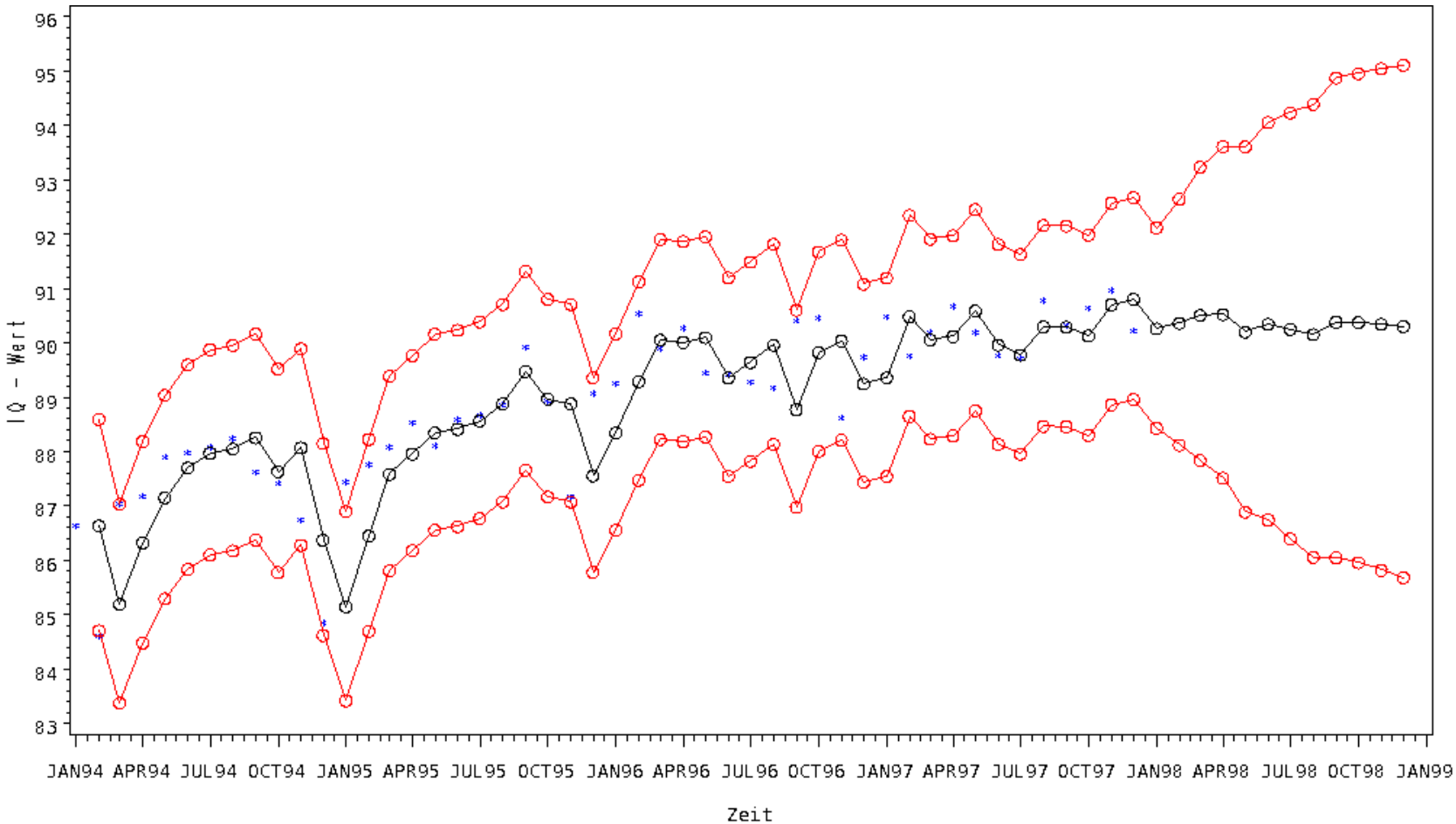
Forecasting

Wir wollen durch das so gewonnene Modell die Entwicklung unser Daten für das kommende Jahr vorhersagen.

SAS-Code für die Auswertung

```
PROC ARIMA
identify var=ylog(1) nlag=15;
estimate p=(1)(9) nonconstant method=cls;
forecast out=estimation lead=12 id=date interval=month;
END;
```

Vorhersage


96
95
94
93
92
91
90
89
88
87
86
85
84
83
IQ - Wert
JAN94 APR94 JUL94 OCT94 JAN95 APR95 JUL95 OCT95 JAN96 APR96 JUL96 OCT96 JAN97 APR97 JUL97 OCT97 JAN98 APR98 JUL98 OCT98 JAN99
Zeit