

- LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN -

INSTITUT FÜR STATISTIK

Saison-Trend-Zerlegung von Zeitreihen

BACHELORARBEIT

Autor: Franz Xaver Stelz

Matrikelnummer:xxxxxxxx

Betreuer: Prof. Dr. Stefan Mitnik, Andreas Fuest

München, den 09.01.2014

Inhaltsverzeichnis

Abbildungsverzeichnis	iii
Tabellenverzeichnis	v
1. Einführung	1
2. Grundlagen	4
2.1. Gleitender Durchschnitt	4
2.2. Loess	4
3. Saison-Trend Zerlegung mithilfe von Loess	8
3.1. Vorgehen	8
3.1.1. Innere Schleife	8
3.1.2. Äußere Schleife	9
3.2. Parametereinstellungen	10
3.2.1. Frequenz der Saison (np)	10
3.2.2. Saisonparameter	11
3.2.3. Trendparameter	15
3.2.4. Low-pass Parameter	19
3.2.5. Anzahl Iterationen der äußere Schleife	19
3.2.6. Anzahl Iterationen der innere Schleife	21
3.2.7. Robustheitsschätzung	22
3.2.8. Computationale Parameter	25
3.2.9. Multiplikative Umformung	26
3.2.10. Vergleich mit dem alten STL-Programm	28
4. Prognose	29
4.1. Methoden	30
4.1.1. Loess-Prognose	30
4.1.2. Random Walk	31
4.1.3. ARIMA-Modell	32
4.2. Prognosebeispiel	34
4.2.1. Punktprognose	34
4.2.2. Prognoseintervall	37
4.3. Prognosegüte	40
4.3.1. Mittlerer quadratischer Fehler	41
4.3.2. Vergleich verschiedener Prognosemethoden	41

Inhaltsverzeichnis

5. empirisches Beispiel	45
6. Schluss	49
6.1. Zusammenfassung	49
6.2. Ausblick	49
A. Beiliegende CD-Rom	51
Literaturverzeichnis	52

Abbildungsverzeichnis

1.1. Zerlegung der Produktion im Produzierenden Gewerbe in Deutschland	2
2.1. Darstellung der Gewichte bei einer Spannweite von 9	6
3.1. Saisonuntergruppengrafiken mit Saisonspannweite (ns) von 7 für die Produktion im Produzierenden Gewerbe in Deutschland	12
3.2. Saisonuntergruppengrafiken mit Saisonspannweite (ns) von 11 für die Produktion Produzierenden Gewerbe in Deutschland	13
3.3. Saisonuntergruppengrafiken bei linearer Schätzung (isdeg=1, ns=11) für die Produktion im Produzierenden Gewerbe in Deutschland	14
3.4. Trenddiagnose und Restkomponente mit ns=9 und nt=23 für die Produktion im Produzierenden Gewerbe in Deutschland	16
3.5. Trenddiagnosegrafik mit ns=9 und nt=13 für die Produktion im Produzierenden Gewerbe in Deutschland	17
3.6. Trenddiagnosegrafik mit ns=9 nt=13 und itdeg=2 für die Produktion im Produzierenden Gewerbe in Deutschland	18
3.7. Gewichtegrafik mit ns=9 und nt=13 (itdeg=1) für die Produktion im Produzierenden Gewerbe in Deutschland	20
3.8. Gewichtegrafik und Histogramm der Gewichte mit ns=9 und nt=19 (itdeg=2) für die Produktion im Produzierenden Gewerbe in Deutschland	21
3.9. Saisonuntergruppengrafiken mit eigenen übergebenen Robustheitsgewichten für die Produktion im Produzierenden Gewerbe in Deutschland	24
3.10. Gewichtegrafik mit Parameter wf=8 für die Produktion im Produzierenden Gewerbe in Deutschland	25
3.11. Multiplikative Zerlegung der Produktion im Produzierenden Gewerbe in Deutschland	27
4.1. Prognose und Vergleich der beiden Modelle für den Trend für die Produktion im Produzierenden Gewerbe in Deutschland	36
4.2. Prognoseintervall für die Prognose des Trendes mit Random-Walk mit Drift für die Produktion im Produzierenden Gewerbe in Deutschland	39
4.3. Prognoseintervall für die Prognose des Trendes mit Arima-Modell für die Produktion im Produzierenden Gewerbe in Deutschland	40
4.4. Vergleich der verschiedenen Prognosemethoden für die Produktion im Produzierenden Gewerbe in Deutschland (langfristige Prognose; rekursives Prognosefenster)	42
4.5. Vergleich der verschiedenen Prognosemethoden für die Produktion im Produzierenden Gewerbe in Deutschland (langfristige Prognose; rollierendes Prognosefenster)	44

ABBILDUNGSVERZEICHNIS

5.1. Arbeitslosigkeit in Deutschland von 1991 bis Oktober 2013	45
5.2. Zerlegung der Arbeitslosigkeit in Deutschland mit ns=13 und nt=7	46
5.3. Zerlegung der Arbeitslosigkeit in Deutschland mit nt=15 und ns=13	47
5.4. Prognosewerte und -intervalle der Arbeitslosigkeit in Deutschland	48

Tabellenverzeichnis

3.1. Box-Pierce Teststatistik verschiedener nt's mit ns=9 und itdeg=1 für die Produktion im Produzierenden Gewerbe in Deutschland	17
3.2. Box-Pierce Teststatistik verschiedener nt's mit ns=9 und itdeg=2 für die Produktion im Produzierenden Gewerbe in Deutschland	19
3.3. Abweichungen, die bei mehrmaliger Iteration der inneren Schleife entstehen	22
3.4. Bei Zerlegung der Produktion im Produzierenden Gewerbe in Deutschland erhaltenen Robustheitsgewichte für den Monat März (ns=9, nt=13, no=10)	22
3.5. Vom Analysten übergebene Robustheitsgewichte für den Monat März	23
3.6. Abweichung, die durch die computationalen Parameter bei der Zerlegung der Produktion im Produzierenden Gewerbe in Deutschland entsteht	26
3.7. Werte für das Jahr 2000 der multiplikativen Zerlegung für die Produktion im Produzierenden Gewerbe in Deutschland	28
4.1. Vergleich und Werte der verschiedenen Prognosemethode der Saisonkomponente für die Produktion im Produzierenden Gewerbe in Deutschland	34
4.2. Modellparameter eines Random-Walks mit Drift in der Trendkomponente für die Produktion im Produzierenden Gewerbe in Deutschland	35
4.3. Modellparameter eines ARIMA-Modells in der Trendkomponente für die Produktion im Produzierenden Gewerbe in Deutschland	35
4.4. Vergleich und Werte der verschiedenen Prognosemethoden der Trendkomponente für die Produktion im Produzierenden Gewerbe in Deutschland	36
4.5. Prognoseintervalle der beiden Modelle der Trendkomponente für die Produktion im Produzierenden Gewerbe in Deutschland	38
4.6. Kovarianzen und Korrelationen zwischen den einzelnen Residuen bei der Produktion im Produzierenden Gewerbe in Deutschland	38
4.7. Vergleich der verschiedenen Prognosemethoden für die Produktion im Produzierenden Gewerbe in Deutschland (kurzfristige Prognose; rekursives Prognosefenster)	42
4.8. Vergleich der verschiedenen Prognosemethoden für die Produktion im Produzierenden Gewerbe in Deutschland (kurzfristige Prognose, rollierendes Prognosefenster)	43

1. Einführung

In der Statistik, sowie in der Finanz- und Wirtschaftsökonomie nimmt die Zeitreihenanalyse eine sehr wichtige Rolle ein. Sie dient vor allem dazu, den Verlauf einer Zeitreihe zu modellieren und anhand dieser die zukünftigen Werte zu prognostizieren. Um die Prognose so einfach wie möglich zu gestalten, versucht man bereits in der Modellierung der Werte möglichst viele Informationen aus den Daten herauszulesen.

Eine Vorgehensweise in der Modellierung besteht darin, die Zeitreihe in mehrere Komponenten (Trend (T), Saison (S) und Rest (R)) zu zerlegen, da in den meisten Fällen Zeitreihen eine saisonale Struktur aufweisen und für den Analysten vor allem der Trend für den zukünftigen langfristigen Verlauf von großer Bedeutung ist.

Grundsätzlich gibt es bei der Komponentenzerlegung zwei unterschiedliche Ansätze: Das additive oder das multiplikative Verfahren.

Beim additiven Verfahren ($Y = T + S + R$) wird von einer konstanten Saison ausgegangen, deren Effekt absolut von Saison zu Saison immer ungefähr gleich stark ausgeprägt ist. Ein Beispiel stellt die Arbeitslosigkeit in Deutschland (in %) dar.

Der multiplikative Ansatz hingegen ($Y = T * S * R$), ist bei einer sich abschwächenden oder verstärkenden saisonalen Struktur deutlich besser geeignet, da diese nun multiplikativ von der Trendkomponente abhängig ist und sich somit äquivalent zum Trend entwickelt. Das heißt, bei zunehmendem Trend wird die Saison stärker, bei abnehmendem Trend schwächer. Ein Beispiel hierfür stellt die Stromgewinnung durch Solarenergie in Deutschland dar. So führt eine Erhöhung der installierten Leistung zu einer stärkeren Erhöhung des produzierten Stromes durch Sonnenenergie in den Sommermonaten, als in den Wintermonaten.

Bevor eine Zerlegungsmethode, die ich in R implementiert habe, vorgestellt wird, folgt eine kurze Beschreibung der Eigenschaften der verschiedenen Komponenten.

Die Saisonkomponente hat eine, mit festem Abstand (z.B. täglich, monatlich, jährlich, ...) wiederkehrende Struktur in den Daten, die in sich möglichst konstant sein soll.

Die Trendkomponente hingegen, soll eine geglättet Linie darstellen, die keinerlei saisonale Struktur im Zeitraum der Saisonkomponente aufweist und aus der eine langfristige Entwicklung ersichtlich sein soll. In dieser besteht trotzdem die Möglichkeit, dass mittel- bis langfristigen Schwankungen auftreten. Vor allem bei makroökonomischen Zeitreihen sind die circa 3-5 Jahren langen Konjunkturzyklen in der Trendkomponente enthalten.

Die dritte Komponente, der Rest, weist die Werte aus, die nicht über die Saison und den Trend erklärt werden können. Es sollte keine regelmäßige Struktur in diesen Daten vorhanden sein. Das bedeutet, alle Werte sollten stochastisch zufällig entstanden sein.

In der folgenden Arbeit soll nun ein Verfahren vorgestellt werden, mit dem die Zerlegung einer Zeitreihe durchgeführt werden kann. Es ist das sogenannte Saison-Trend-Zerlegungsverfahren anhand von Loess (Abkürzung: STL), entwickelt vor allem von Cleveland. Es beruht auf einem iterativen Algorithmus, dem additiven Ansatz der Komponentenzerlegung und verwendet vor allem das Loess-Verfahren zur Schätzung der einzelnen Werte der Komponenten. Durch Manipulation der Input-Zeitreihe besteht die Möglichkeit das

1. Einführung

Verfahren auch multiplikativ zu berechnen (Cleveland et al., 1990).

Das STL-Verfahren wurde von Cleveland bereits in R mithilfe der Programmiersprache Fortran77 implementiert. Ich habe nun den Fortran77-Code in R neu geschrieben und um einige Einstellungsmöglichkeiten erweitert. Außerdem habe ich einige Diagnoseplots zur besseren Bestimmung der Parameter in R implementiert.

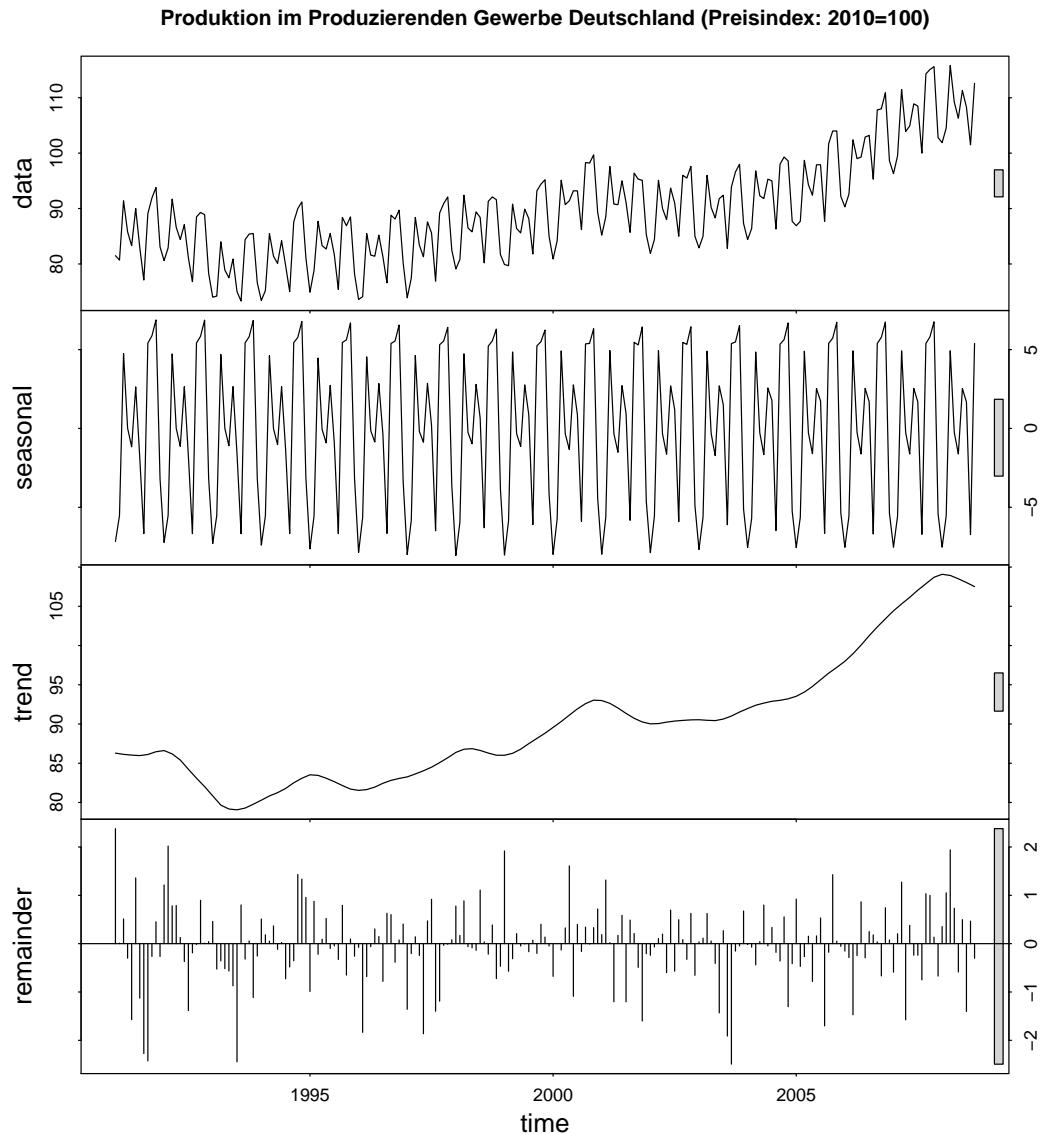


Abbildung 1.1.: Zerlegung der Produktion im Produzierenden Gewerbe in Deutschland

In der Abbildung 1.1 sind die Ergebnisse der Zerlegung einer makroökonomischen Zeitreihe dargestellt. Die Zerlegung wurde mit den STL-Verfahren durchgeführt. Die Ausgangszeitreihe stellt die Produktion im Produzierenden Gewerbe in Deutschland von Januar 1991 bis September 2008 dar. Die Werte wurden kalenderbereinigt und liegen in konstanten Preisen vor (inflationsbereinigt). Das Jahr 2010 stellt die Referenz dar und besitzt somit den Wert 100 (Deutsche Bundesbank, 2013). Um die Finanzkrise im Jahre 2008 und 2009 zu umgehen, wurde die Zeitreihe zu einem früheren Zeitpunkt abgeschnitten.

1. Einführung

Deswegen stellt das Jahr 2010 die Referenz dar, obwohl diese in der verwendeten Inputzeitreihe noch keine Realisation besitzt. Mithilfe der Produktion im Produzierenden Gewerbe (01.1991-09.2008) werden im späteren Verlauf die verschiedenen Einstellungsmöglichkeiten des STL-Verfahrens erläutert. Zuerst beachte man die Balken am rechten Rand. Sie geben an, in welcher Achsen-skalierung sich die jeweiligen Grafiken befinden. So entspricht die Höhe der jeweiligen Balken der gleichen y-Achsenabschnittsgröße.

In der ersten Grafik ist die Ausgangszeitreihe dargestellt. Sie weist sehr viel Schwankung auf, wobei eine gleichbleibende und wiederkehrende Struktur in der Zeitreihe erkennbar ist. Dies spricht für einen saisonalen Verlauf. Über den kompletten Zeitverlauf ist ein positiver Trend festzustellen. In den folgenden Grafiken sind die einzelnen Komponenten der zerlegten Zeitreihe dargestellt.

In der zweiten Grafik erkennt man die konstante Struktur der Saisonkomponente (jährlich), so ist der Verlauf innerhalb eines Jahres konstant über den Zeithorizont.

Die dritte Grafik weist den Verlauf der Trendkomponente auf, sie besitzt einen positiven Trend. Die Konjunkturzyklen von 3-5 Jahren sind deutlich erkennbar.

Die Restkomponente ist in der vierten Grafik dargestellt. Sie weist scheinbar sehr wenig Struktur auf. Verfahren um dies festzustellen werden später näher erläutert. Am Balken am rechten Rand erkennt man, dass die Restkomponente die geringsten Werte aufweist, so ist die Saison deutlich stärker ausgeprägt, als der Rest.

Im nächsten Kapitel werden zunächst die im STL vorkommenden statistischen Glättungsverfahren erläutert, der gleitenden Durchschnitt und das Loess-Verfahren.

2. Grundlagen

2.1. Gleitender Durchschnitt

Der gleitende Durchschnitt stellt ein einfaches statistisches Verfahren dar, um eine Zeitreihe zu glätten. Bei diesem wird der Mittelwert einer bestimmten Anzahl umliegender Punkte (τ), des zu schätzenden Zeitpunktes, berechnet. Betrachtet man eine Zeitreihe (y_1, y_2, \dots, y_t) , so wird der geschätzte Mittelwert \hat{y}_i zum Zeitpunkt i folgendermaßen berechnet (Treiber, 2010, vgl. 17(a)).

$$\text{Für } \tau \text{ gerade: } m = \frac{\tau - 1}{2} ; \hat{y}_i = \frac{1}{\tau} \sum_{j=i-m}^{i+m} y_j \quad (2.1)$$

$$\text{Für } \tau \text{ ungerade: } m = \frac{\tau}{2} ; \hat{y}_i = \frac{1}{\tau} \left(\frac{y_{j-m} + y_{j+m}}{2} + \sum_{j=i-m+1}^{i+m-1} y_j \right) \quad (2.2)$$

Dieses Verfahren stellt eine gute Grundlage dar, um Trendverläufe zu erlangen, da man über einen bestimmten Zeitraum die Werte mittelt. Würde man z.B. den Wert von τ in Höhe der Anzahl der Zeitpunkte je Saison einstellen, so könnte man diese aus der Zeitreihe wegmitteln (Treiber, 2010, vgl. 17(b)). Jedoch treten bei diesem Verfahren immer Probleme an den links- und rechtsseitigen Grenzen der Zeitreihe auf. Wie man anhand der Formel erkennen kann, geht der untere Laufindex der Summe von $i = i - m$. Das bedeutet der erste zu schätzende Zeitpunkt muss größer als m sein. Somit können die ersten m - und die letzten m -Zeitpunkte der Zeitreihe nicht geschätzt werden. Dies ist nicht zufriedenstellend, da vor allem die rechtsseitige Grenze von großer Bedeutung für die Prognose ist.

Dass dieses Verfahren trotzdem Eingang in das STL gefunden hat, beruht darauf, dass innerhalb des Algorithmus mit dem Loess-Verfahren Werte prognostiziert werden und aufgrund dessen die Abschneidung der m -letzten Werte keine Probleme darstellt (vgl. Kap.3.1.1). Außerdem können mithilfe des gleitenden Durchschnitts langfristige Schwankungen, die bei der Berechnung der Saison auftreten und in die Trendkomponente einfließen sollen, einfach und unkompliziert berechnet werden.

2.2. Loess

Das Loess-Verfahren kann dazu benutzt werden eine Zeitreihen zu glätten. So muss bei diesem für die Berechnung immer eine zeitliche Komponente x_i und die Werte zu diesen Zeitpunkten $y(x_i)$ übergeben werden ($i = 1, \dots, n$), um ein Ergebnis zu erhalten. Es besteht auch die Möglichkeit, dass die Ausgangszeitreihe an einzelnen Zeitpunkten fehlende Werte enthält.

Loess beruht auf dem Verfahren der **Lokal** gewichteten Regression. So wird für jeden Zeitpunkt x_i eine individuelle Regressionsschätzung durchgeführt, bei der nur der geschätzte Wert $\hat{y}(x_i)$ von Bedeutung ist, nicht aber die einzelnen Regressionsparameter. Bevor die Schätzung der Regression mithilfe der Kleinsten-Quadrat-Schätzung vorgenommen

2. Grundlagen

werden kann, müssen zuerst die Gewichte berechnet werden, die bestimmen mit welchem Anteil die umliegenden Werte in die Schätzung mit einfließen. So soll erreicht werden, dass Werte, die in der Nähe des zu schätzenden Zeitpunktes liegen, höhere Gewichte erhalten, als Werte, die weiter vom Zeitpunkt entfernt liegen. Im folgenden wird zur Vereinfachung der Zeitpunkt, zu dem ein Wert geschätzt wird, mit x_0 benannt.

Die Gewichte sind für jede individuelle Regressionschätzung unterschiedlich und werden für jeden Zeitpunkt anhand des zeitlichen Abstandes zum schätzenden Punkt, individuell bestimmt.

Zuerst übergibt der Statistiker die Spannweite (q). Sie gibt die Anzahl der Punkte (umliegenden) an, die in die Schätzung mit einfließen sollen. Der Wert von dieser muss ungerade sein, da der Wert des zu schätzenden Zeitpunktes in der Spannweite mit eingerechnet wird. So impliziert eine Spannweite von 13, dass die vorherigen und nachfolgenden sechs Werte, sowie der Wert zum Zeitpunkt x_0 in die Schätzung aufgenommen werden. Sollten nicht ausreichend vorherige oder nachfolgende Werte zur Verfügung stehen (links- oder rechtsseitige Grenze der Zeitreihe), wird ein zusätzlicher nachfolgender oder vorheriger Wert mit einbezogen. Nur für diese Werte wird ein Gewicht $w_{x_0}(x_j)$ zur Schätzung des Zeitpunktes x_0 ermittelt.

Für die Gleichung der Gewichtsberechnung müssen zwei unterschiedliche Fälle unterschieden werden.

Für den ersten Fall nimmt man an, dass $q \leq n$ ist (n =Anzahl der Zeitpunkte). Danach wird der Wert des maximalen zeitlichen Abstandes der einfließenden Punkte zum schätzenden Wert bestimmt ($\lambda(x_0) = \max|x_0 - x_j|$) und mithilfe einer Kerndichtefunktion die einzelnen Gewichte ermittelt. Beim Loess-Verfahren geschieht dies mithilfe einer trikubischen Funktion (Cleveland et al., 1990, vgl. S. 5-6).

$$K(u) = \begin{cases} (1 - u^3)^3 & \text{falls } 0 < u < 1 \\ 0 & \text{sonst} \end{cases} \quad (2.3)$$

Somit erhält man folgende Gleichung für die Berechnung der einzelnen Gewichte:

$$w_{x_0}(x_j) = K\left(\frac{|x_j - x_0|}{\lambda(x_0)}\right) \quad (2.4)$$

Wie man an der Formel 2.4 erkennen kann, nehmen die Gewichte nur Werte zwischen null und eins an, da die Kerndichtefunktion nur diesen Wertebereich annimmt. Außerdem erkennt man, dass Werte, die nahe des zu schätzenden Zeitpunktes liegen, höhere Gewichte bekommen, als Werte, die weiter entfernt liegen (vgl. Abbildung 2.1). Der Punkt mit maximalem Abstand fließt nicht mehr in die Schätzung mit ein, da dessen Gewicht null beträgt. Somit fließen nur $q - 2$ Werte in die Schätzung mit ein, wenn die Ausgangswerte symmetrisch um den zu schätzenden Zeitpunkt liegen. Deswegen stellt eine Spannweite die kleiner als drei ist, keinen sinnvollen Wert dar, da sonst der Schätzwert der Realisation entspricht.

Für zweiten Fall nimmt man an, dass $q > n$ ist. Wenn dies eintritt, wird $\lambda(x_0)$ noch mit dem Faktor q/n multipliziert und man erhält für die Gewichtsberechnung folgende Gleichung.

$$w_{x_0}(x_j) = K\left(\frac{|x_j - x_0|}{\lambda(x_0) \frac{q}{n}}\right) \quad (2.5)$$

2. Grundlagen

Der Bruch q/n ist größer als eins ($q > n$). Dies führt dazu, dass der Nenner von u größer wird und daraus folgt, dass der Wert von u verringert wird. Da $K(u) = (1 - u^3)^3$ gilt und u kleiner wird, nimmt der Wert der Kerndichtefunktion und damit der Wert für das jeweilige Gewicht zu. Somit führt dies dazu, dass weiter entfernt liegende Werte, wie vom Benutzer erwünscht, ein höheres Gewicht bekommen.

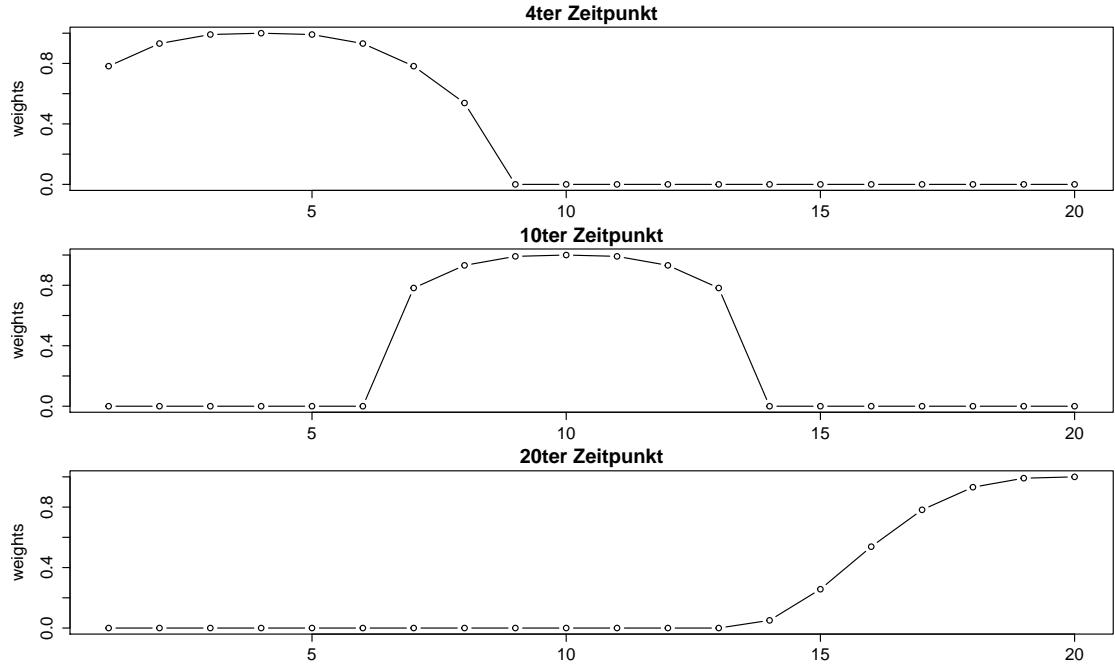


Abbildung 2.1.: Darstellung der Gewichte bei einer Spannweite von 9

In Abbildung 2.1 sieht man beispielhaft, welche Werte die Gewichte mit zunehmendem Abstand annehmen. Die Spannweite wurde mit dem Wert neun festgelegt ($q = 9$). In der ersten Grafik sieht man die Gewichte für die Schätzung des vierten Zeitpunktes, in der zweiten Graphik für die Schätzung des zehnten Zeitpunktes und in der dritten Graphik die Schätzung für den zwanzigsten Zeitpunkt. In der letzten Grafik erkennt man, dass nur Zeitpunkte, die links vom zu schätzenden Punkt liegen, für die Berechnung benutzt werden. Ebenso sieht man, dass immer neun Werte in die Berechnung mit einfließen. So dienen die ersten neun Punkte für die Schätzung der ersten fünf Zeitpunkte, jedoch jeweils mit unterschiedlicher Gewichtung.

Nachdem man nun die einzelnen Gewichte berechnet hat, wird nun die lokal gewichtete Regression anhand der Kleinsten-Quadrat-Schätzung vorgenommen.

Zuerst wird die Regressionsgleichung einer lokalen Regression ohne Gewichte gezeigt.

$$y(x_j) = \beta_0 + \beta_1(x_0 - x_j) + \beta_2(x_0 - x_j)^2 + \dots + \beta_p(x_0 - x_j)^p + \epsilon(x_j) \quad (2.6)$$

Man erkennt, dass für die Schätzung des Wertes $\hat{y}(x_0)$ an der Stelle x_0 nur die Konstante (Parameter β_0) von Bedeutung ist, da die Regression ausgehend vom Punkt (x_0) ermittelt wird und der Faktor $(x_0 - x_j)^p$ den Wert null annimmt.

Doch wie schätzt man den Parameter β_0 ? Dies geschieht über die Minimierung der Fehlerquadratsumme, wobei bei der gewichteten Regression die einzelnen Fehlerquadrate noch mit dem ihnen zugewiesenen Gewicht multipliziert werden. Somit erhält man folgendes Minimierungsproblem für die Bestimmung der Parameter $(\beta_0, \dots, \beta_p)$ (Foster, b, vgl. S. 1).

2. Grundlagen

$$Q(\beta_0, \dots, \beta_p) = \min \sum_{j=1}^n \epsilon(x_j)^2 * w_{x_0}(x_j) \quad (2.7)$$

$$= \min \sum_{j=1}^n [y(x_j) - \beta_0 - \beta_1(x_j - x_0) - \dots - \beta_p(x_j - x_0)^p]^2 * w_{x_0}(x_j) \quad (2.8)$$

Zur Vereinfachung betrachtet man die Bestimmung des Wertes über die Matrizen- und Vektordarstellung. Man benötigt folgende Matrizen und Vektoren (Foster, b, vgl. S. 1-3),

$$X_{x_0} = \begin{Bmatrix} 1 & (x_1 - x_0) & \dots & (x_1 - x_0)^p \\ 1 & (x_2 - x_0) & \dots & (x_2 - x_0)^p \\ \dots & \dots & \dots & \dots \\ 1 & (x_n - x_0) & \dots & (x_n - x_0)^p \end{Bmatrix} \quad (2.9)$$

$$Y = (y(x_1), \dots, y(x_n))^T \quad (2.10)$$

$$B = (\beta_0, \beta_1, \dots, \beta_p)^T \quad (2.11)$$

$$W_{x_0} = \begin{Bmatrix} w_{x_0}(x_1) & 0 & \dots & 0 \\ 0 & w_{x_0}(x_2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & w_{x_0}(x_n) \end{Bmatrix} \quad (2.12)$$

und damit muss nun folgende Fehlerquadratsumme minimiert werden.

$$\min[(Y - X_{x_0}B)^T W_{x_0} (Y - X_{x_0}B)] \quad (2.13)$$

Nachdem man den vorherigen Term aus Formel 2.13 abgeleitet und null gesetzt hat, erhält man folgende Lösung für die Gleichung.

$$\hat{B} = (X_{x_0}^T W_{x_0} X_{x_0})^{-1} (X_{x_0}^T W_{x_0} Y) \quad (2.14)$$

Da bei der lokalen Regression die Parameter in Abhängigkeit des zu schätzenden Zeitpunktes bestimmt werden, gibt, wie oben gezeigt, der Wert von β_0 den zu schätzenden Wert an. Um den Wert somit zu erhalten, ist es ausreichend, die geschätzten Parameter \hat{B} mit einem Vektor zu multiplizieren, dessen Länge $p+1$ ist und der eine führende eins und sonst nur Nullen besitzt, er wird mit e_1 definiert.

$$\hat{y}(x_0) = e_1 (X_{x_0}^T W_{x_0} X_{x_0})^{-1} (X_{x_0}^T W_{x_0} Y) \text{ mit } e_1 = (1, 0, \dots, 0) \quad (2.15)$$

Mithilfe der Formel aus 2.15 erhält man den geschätzten Wert zum Zeitpunkt x_0 . Um die komplette Loess-Schätzung zu erhalten, wird nun für jeden Punkt der Zeitreihe der Wert mithilfe der lokalen gewichteten Regression ermittelt. Im später vorgestellten Verfahren besteht die Möglichkeit einer konstanten, linearen oder quadratischen Regressionsschätzung ($p = 0, 1, 2$).

3. Saison-Trend Zerlegung mithilfe von Loess

3.1. Vorgehen

Das STL-Programm beruht auf einem Algorithmus und ist ein iteratives Saison-Trend Zerlegungsverfahren, dessen Hauptbestandteile eine innere und äußere Schleife sind. In der inneren Schleife werden die einzelnen Komponenten (Saison und Trend) geschätzt, wobei die innere Schleife jeweils ni -mal durchlaufen wird. Die äußeren Schleife besteht aus der inneren Schleife und einer Robustheitsschätzung. Bei dieser bekommen Werte, die nur schlecht über die beiden Komponenten Saison und Trend erklärt werden können, geringere Gewichte als Werte, die gut erklärbar sind. Bei einem erneuten Durchlauf der inneren Schleife fließen diese Gewichte in die Komponentenschätzung mit ein. Insgesamt wird die äußere Schleife no -mal (ni und no sind Parameter des STL-Programms und müssen und müssen vom Analysten angegeben werden. Sie geben die Anzahl der Schleifendurchläufe an) durchlaufen.

Als Output erhält man:

- Saisonkomponente ($S(x_i)$)
- Trendkomponente ($T(x_i)$)
- Restkomponente ($R(x_i)$)
- Robustheitsgewichte ($w_{ro}(x_i)$)

3.1.1. Innere Schleife

Die Abfolge der Vorgänge in der inneren Schleife werden im folgenden dargestellt, wobei noch nicht auf die genauen Parametereinstellungen eingegangen wird. Die Namen in Klammern stellen die Variablennamen der Parameter in den von mir geschriebenen R-Code dar (Cleveland et al., 1990, vgl. S.6-8).

1. Schritt: Zuerst wird der im vorherigen Umlauf geschätzte Trend, von der Originalzeitreihe abgezogen. In der ersten Iteration ist noch keine Trendkomponente vorhanden und die Originalzeitreihe dient als Input für die ersten Berechnungen. ($DT(x_i) = Y(x_i) - T(x_i)$)

2. Schritt: Die enttrendete Zeitreihe ($DT(x_i)$) wird in ihre saisonalen Untergruppen aufgeteilt. Dies geschieht anhand des Parameters np , der die Anzahl der Zeitpunkte je Saison angibt. Für monatliche Daten z.B. ist $np=12$ und es wird für jeden Monat eine eigene Zeitreihe erstellt. Auf diese Saisonuntergruppen wird nun das Loess-Verfahren angewendet, wobei auch ein Wert für den vorherigen und nachfolgenden Wert der Zeitreihe geschätzt wird. Die Spannweite (ns) und den Grad ($isdeg$) der Loess-Schätzung müssen vom Benutzer angegeben werden. Danach werden die einzelnen Saisonuntergruppenschätzungen wieder zu einer Zeitreihe zusammengefügt ($C(x_i)$). Diese Zeitreihe besteht nun aus $n+2*np$ Beobachtungen.

3. Saison-Trend Zerlegung mithilfe von Loess

3.Schritt: Um die endgültige Saisonkomponente zu erhalten, werden noch einige weitere Berechnungen vorgenommen. So wird versucht langfristige Schwankungen, die durch die Schätzung in den Saisonuntergruppen entstanden sind, zu ermitteln und aus dieser zu entfernen. Dies geschieht, indem man zweimal einen gleitenden Durchschnitt mit der Spannweite von np auf die zusammengefügte Saisonuntergruppenschätzung anwendet. Danach wird noch ein gleitender Durchschnitt mit Spannweite drei berechnet. Wie im Kapitel 2.1 bereits erwähnt, treten bei der gleitenden Durchschnitteberechnung Probleme an den Rändern der Zeitreihe auf. Deswegen werden im vorangegangenen Schritt eine Loess-Schätzung für den vorherigen und folgenden Zeitpunkt jeder Saisonuntergruppe vorgenommen, damit die einzelnen Punkte problemlos abgeschnitten werden können. Die nach den verschiedenen gleitenden Durchschnitten erhaltene Zeitreihe weist nun wieder n -Beobachtungen auf. Im letzten Schritt wird nun noch einmal ein Loess-Verfahren auf die Zeitreihe angewendet. Man erhält nun eine Zeitreihe, die eine Korrektur der Saisonkomponente darstellt ($SC(x_i)$). Die Spannweite (nl) und der Grad (ildeg) des Loess-Verfahrens können wie vorher vom Analysten bestimmt werden, wobei in diesem Fall Defaultwerte vorliegen.

4.Schritt: Jetzt wird die endgültige Saisonkomponente berechnet. Dazu wird die in Schritt 3 berechnete Zeitreihe von der Saisonuntergruppenschätzung abgezogen ($S(x_i) = C(x_{i+np}) - SC(x_i)$).

5.Schritt: Nachdem nun die Saison ausgewiesen worden ist, wird nun die Trendkomponente berechnet. Den Input hierfür liefert eine desaisonalisierte Zeitreihe. Es wird die im vorherigen Schritt erhaltene Saison von der Originalzeitreihe abgezogen ($DS(x_i) = Y(x_i) - S(x_i)$).

6.Schritt: Auf die desaisonalisierte Zeitreihe ($DS(x_i)$) wird ein Loess-Verfahren angewendet. Der Benutzer muss in diesem Fall die Spannweite (nt) und den Grad der Schätzung (itdeg) vorgeben. Die erhaltene Zeitreihe stellt die neue Trendkomponente dar und dient als Input für die erneute Berechnung der Saisonkomponente, falls ein erneuter Schleifendurchlauf stattfindet.

7.Schritt: Im letzten Schritt der inneren Schleife wird die Restkomponente berechnet. Es werden einfach die beiden in den vorherigen Schritten berechneten Komponenten (Saisonkomponente und Trendkomponente) von der Originalzeitreihe abgezogen ($R(x_i) = Y(x_i) - S(x_i) - T(x_i)$). Nachdem der letzte Schritt geschehen ist, werden die einzelnen Teilschritte der inneren Schleife erneut durchgeführt.

3.1.2. Äußere Schleife

In der äußeren Schleife findet eine Robustheitsschätzung statt. Das heißt, man versucht Werte, die nur schlecht über den Trend und die Saison erklärt werden können, zu gewichten, damit diese im erneuten Durchlauf der inneren Schleife weniger Einfluss auf die Komponentenschätzung bekommen. Dies stellt sich als sehr sinnvoll dar, da sich vor allem bei makroökonomischen Zeitreihen oft unplausible Werte, aufgrund von externen Effekten ergeben, z.B. extrem kalte Winter führen zu geringer Produktion.

Die Gewichte der Robustheitsschätzung werden über die Werte der Restkomponente berechnet. So muss zuerst der Median der absoluten Werte der Restkomponente berechnet werden. Anschließend werden die Gewichte äquivalent der Gewichtsberechnung des Loess-Verfahrens mithilfe einer Kerndichtefunktion erstellt, wobei eine quadratische Funktion benutzt wird und der Nenner des Bruchs defaultmäßig den sechsfachen Median der Rest-

3. Saison-Trend Zerlegung mithilfe von Loess

komponente aufweist (Cleveland et al., 1990, vgl. S. 8).

$$K(ro) = \begin{cases} (1 - (ro)^2)^2 & \text{falls } 0 < ro < 1 \\ 0 & \text{sonst} \end{cases} \quad (3.1)$$

Somit erhält man folgende Berechnung für die einzelnen Robustheitsgewichte:

$$w_{ro}(x_j) = K\left(\frac{|R(x_j)|}{6 * med(|R(x)|)}\right) \quad (3.2)$$

Die erhaltenen Gewichte werden im erneuten Durchlauf der inneren Schleife mit den Abstandsgewichten multipliziert ($w(x_j) = w_{x0}(x_j) * w_{ro}(x_j)$) und mit diesen wird die Kleinsten-Quadrat-Schätzung einer lokal gewichteten Regression berechnet.

Für den Benutzer besteht die Möglichkeit den Faktor vor dem Median zu bestimmen (Parameter: wf). So führt ein größerer Wert von diesem allgemein dazu, dass die Ausreißer nicht so stark aus der Schätzung herausgewichtet werden. Eine andere Möglichkeit besteht darin, dem Algorithmus von Anfang an eigene Robustheitsgewichte zu übergeben. In diesem Fall wird nur die Komponentenschätzung (innere Schleife) mit den vorgegebenen Gewichten durchgeführt.

3.2. Parametereinstellungen

In diesem Teilabschnitt wird die optimale Wahl und die Default-Einstellungen der Parameter erklärt. Dies geschieht mithilfe der im ersten Kapitel (Abbildung 1.1) gezeigten Zeitreihe (Produktion im Produzierenden Gewerbe in Deutschland).

Die wichtigsten Parameter sind:

- Frequenz der Saison (np)
- Spannweite bei der Loessschätzung der Saison (ns)
- Grad der Loessschätzung der Saison (isdeg)
- Spannweite bei der Loessschätzung des Trendes (nt)
- Grad der Loessschätzung des Trendes (itdeg)
- Anzahl Iterationen der äußeren Schleife (no)
- Anzahl Iterationen der inneren Schleife (ni)

Außerdem wird sich noch mit einigen anderen Parametereinstellungen und Möglichkeiten kritisch auseinandergesetzt, damit die bestmögliche Zerlegung der Zeitreihe gefunden werden kann.

3.2.1. Frequenz der Saison (np)

Der Parameter np gibt die Anzahl der Zeitpunkte innerhalb einer Saison an. Sie wird in R über den Input bestimmt. Die Input Zeitreihe muss vom Typ ts (time-series) sein und weist somit bereits einen Wert für np auf. Sie ist in den meisten Fällen ziemlich einfach, zum Beispiel für vierteljährliche Daten ist np=4, für monatliche Daten ist np=12 und für tägliche Daten ist np=7.

3. Saison-Trend Zerlegung mithilfe von Loess

3.2.2. Saisonparameter

Für die Loess-Schätzung der Saisonuntergruppen muss vom Benutzer die Spannweite (ns) und der Grad der Schätzung (isdeg) übergeben werden (vgl. Kap. 2.1, Schritt 3).

Die Bestimmung der Spannweite ist in erster Linie von den Wünschen und Zielen des Analysten abhängig. Jedoch ist das Ziel in den meisten Fällen, eine in sich homogenen Saison mit möglichst wenig Variation innerhalb der einzelnen Saisonuntergruppen über den Zeitverlauf zu erhalten. Als Saisonuntergruppe bezeichnet man eine Zeitreihen, die nur aus Werte eines Saisonabschnitts bestehen (z.B. bei monatlichen Daten, Zeitreihe nur mit Januar-Werten).

Der Wert von ns muss ungerade sein und sollte größer oder gleich sieben sein. Dies liegt daran, da sonst zu wenige Werte mit in die Schätzung einfließen würden und somit mehrere zu schätzende Zeitpunkte nur von einem beliebigen Zeitpunkt abhängig sein könnten und den gleichen Schätzwert erhalten.

Im Ausnahmefall kann ns auch den Wert fünf annehmen. Dies sollte aber nur geschehen, wenn es eine konstante Regressionsschätzung gibt und nur wenige Zeitpunkte je Saisonuntergruppe vorhanden sind. In diesem Fall ist besonders vom Analysten darauf zu achten, dass aufgrund der Robustheitsschätzung keine unplausiblen Ergebnisse innerhalb einzelner Saisonuntergruppen entstehen, da zwei heraus gewichtete Saisonwerte bereits dazuführen können, dass mehrere Werte den gleichen Schätzwert erhalten. Wenn nur die innere Schleife durchlaufen wird, stellt eine Spannweite von fünf kein Problem dar, da somit nicht die Möglichkeit besteht, dass Werte herausgewichtet werden und somit jeder Wert aus mindestens drei Zeitpunkten geschätzt wird, wenn die Ausgangs-Zeitreihe keine fehlenden Werte besitzt.

Doch wie bestimmt der Benutzer den Parameter ns? Dies geschieht visuell mithilfe eines Diagnoseplots der Saisonuntergruppengrafiken. Dort ist der Input (Punkte) und Output (rote Linie) der Loess-Schätzung jeder Saisonuntergruppe im Zeitverlauf dargestellt. Da die Saison in sich möglichst konstant sein soll, soll möglichst wenig Schwankung in der Linie der geschätzten Werte innerhalb einer Saisonuntergruppe sichtbar sein.

Beispielhaft betrachtet man die Saisonuntergruppengrafiken für die Produktion im Produzierenden Gewerbe in Deutschland, einmal mit einer geringen Spannweite ($q=7$) für die Saison und einmal mit einer großen Spannweite ($q=11$). Dadurch kann man aufzeigen, wie sich die Höhe der Spannweite auf den Verlauf der Saisonuntergruppen auswirkt und wie man die optimale Wahl der Saisonspannweite mithilfe dieses Diagnoseplots finden kann.

In den Abbildungen 3.1 und 3.2 sind die Saisonuntergruppen für die Produktion im Produzierenden Gewerbe dargestellt. Jede Untergrafik stellt die Werte der Loess-Schätzung für einen anderen Monat dar. Die Grafik beginnt mit dem Januar (Nr.1). Die gestrichelte Linie stellt die Nulllinie dar, anhand der Verschiebung zu dieser kann man erkennen, ob es sich um einen Monat mit hoher oder schwacher Produktion, im Vergleich zum von der Trendkomponente hervorgesagten Verlauf, handelt.

Entscheidend für die Wahl des Parameters ns ist die rote Linie. Sie bildet die Ergebnisse der Loess-Schätzung ab. In Abbildung 3.1 erkennt man, dass es noch mehr Schwankungen innerhalb der einzelnen Monate (Saisonuntergruppengrafiken) gibt, als in Abbildung 3.2. So weist der Februar und Mai (Nr. 2,5) noch sehr viel Variation auf. Im Vergleich dazu sind die Variationen innerhalb der Saisonuntergruppengrafiken in Abbildung 3.2 kaum mehr ersichtlich. Jede Grafik weist in dieser eine konstante Linie aus. Allgemein kann man sagen, dass es sich um eine sehr gleichbleibenden Saisonstruktur handelt und deswegen

3. Saison-Trend Zerlegung mithilfe von Loess

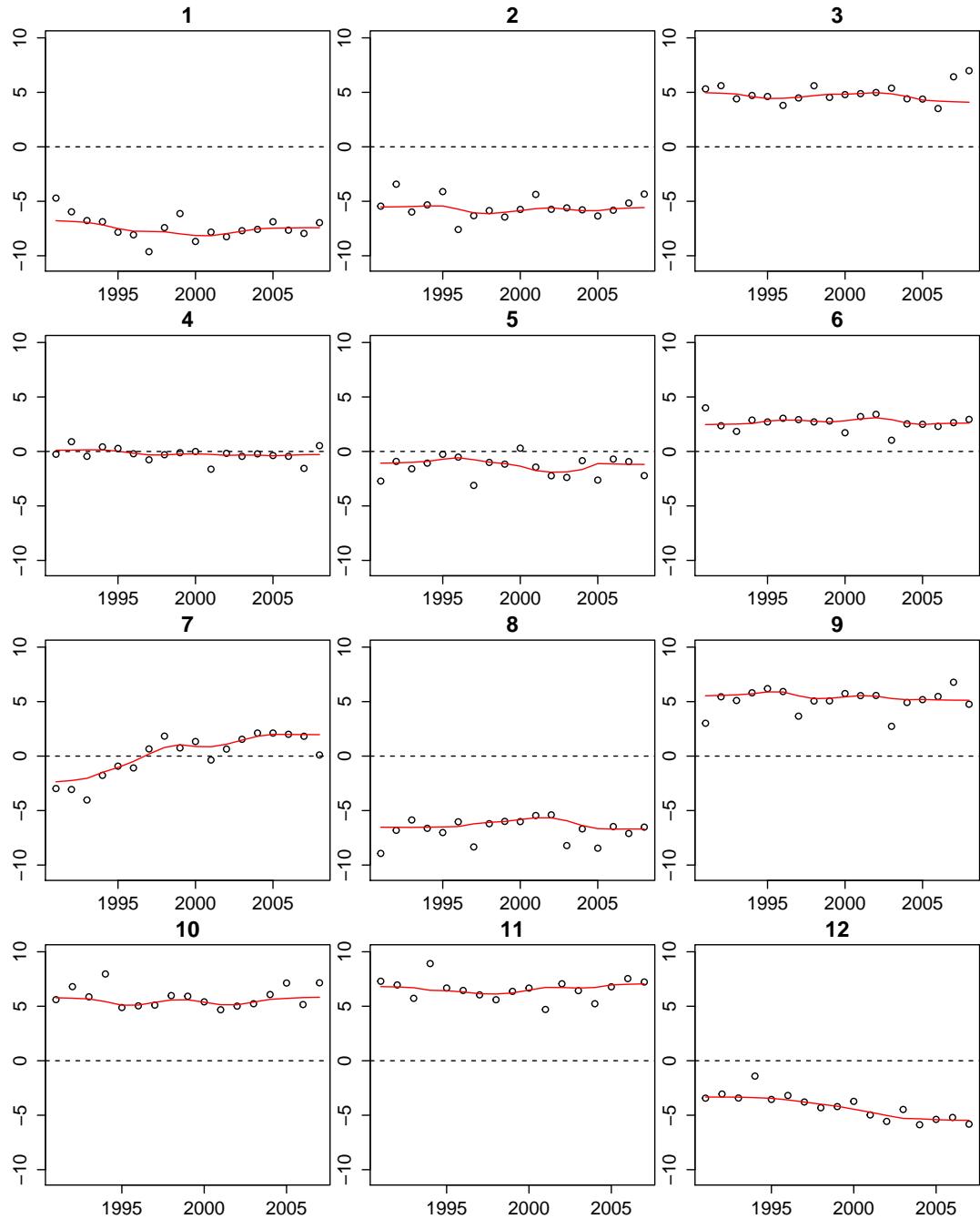


Abbildung 3.1.: Saisonuntergruppengrafiken mit Saisonspannweite (ns) von 7 für die Produktion im Produzierenden Gewerbe in Deutschland

bereits niedrige Werte von ns zu guten Ergebnissen führen.

Da es kein Kriterium für die Saisonspannweite (ns) gibt, muss diese also durch ausprobieren und betrachten des Diagnoseplots vom Analysten bestimmt werden. In unserem Fall, würde man sich für die große Spannweite entscheiden, da diese eine bessere Anpassung an die gewünschten Eigenschaften der Saisonkomponente liefert. Versuche mit einer Spannweite von neun haben zum Ergebnis geführt, dass für die Produktion im Produzierenden

3. Saison-Trend Zerlegung mithilfe von Loess

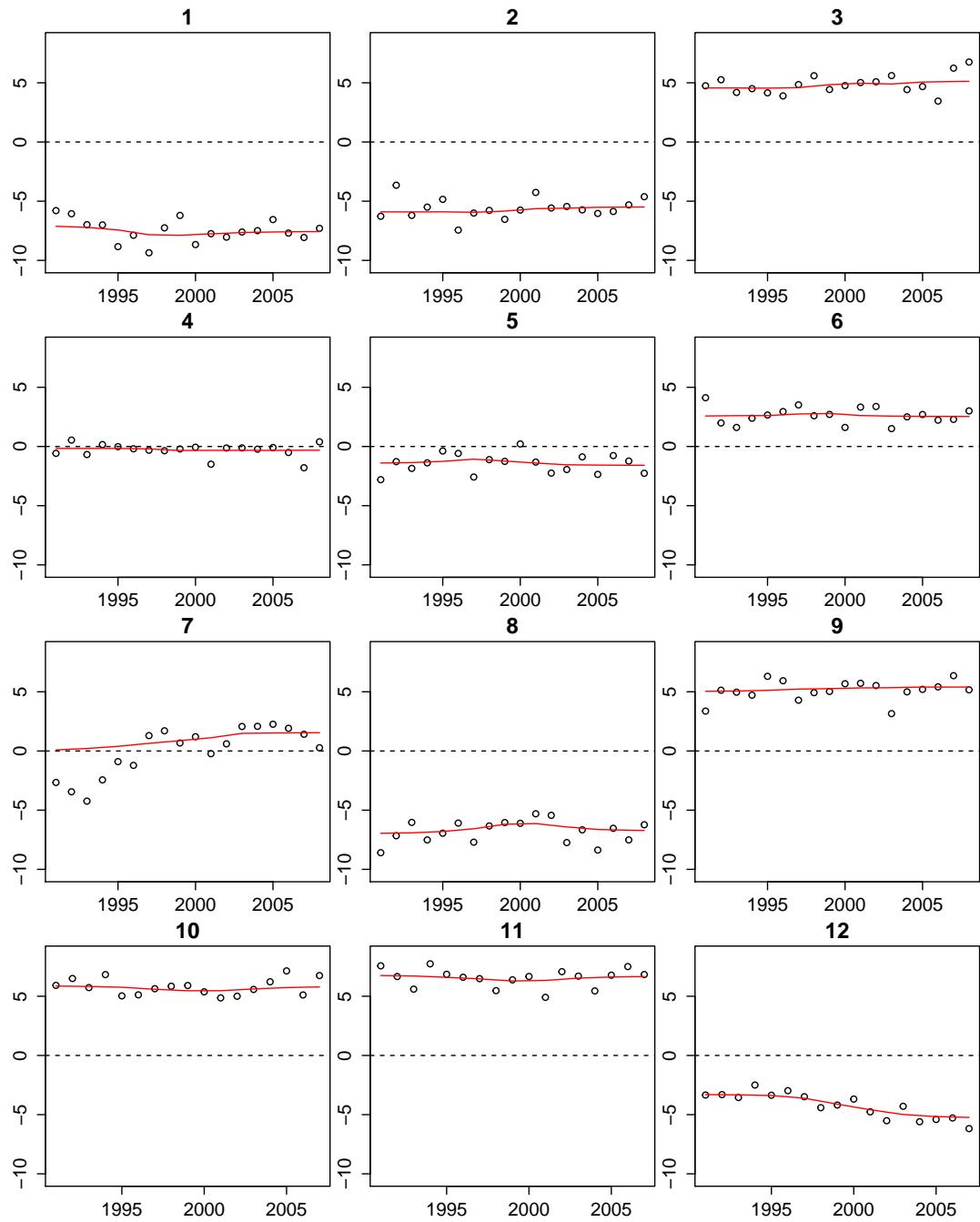


Abbildung 3.2.: Saisonuntergruppengrafiken mit Saisonspannweite (ns) von 11 für die Produktion Produzierenden Gewerbe in Deutschland

Gewerbe in Deutschland eine Spannweite von neun am besten für die Saisonspannweite geeignet ist. Für die weitere Findung der Parametereinstellungen wird nun immer diese Größe verwendet (ns=9).

Ein weiterer Parameter, der für die Berechnung der Saisonkomponente eingestellt werden kann, ist der Grad der Loess-Schätzung (isdeg). Da die Struktur der Saisonkomponente über den Zeitverlauf konstant sein soll, ist die Einstellung von diesem defaultmäßig eine

3. Saison-Trend Zerlegung mithilfe von Loess

konstante Regressions schätzung (isdeg=0).

Ein positiver oder negativer Trend in allen Saisonuntergruppengrafiken, der eine lineare Regressionsschätzung rechtfertigen würde, tritt äußerst selten auf. Trotzdem besteht die Möglichkeit eine lineare Regressionsschätzung durchzuführen. Davon ist in den meisten Fällen jedoch abzuraten, da es bei dieser ebenso oftmals zu einer Anpassung an die Inputwerte der rechts- und linksseitigen Grenzen der Saisonuntergruppenzeitreihen kommt.

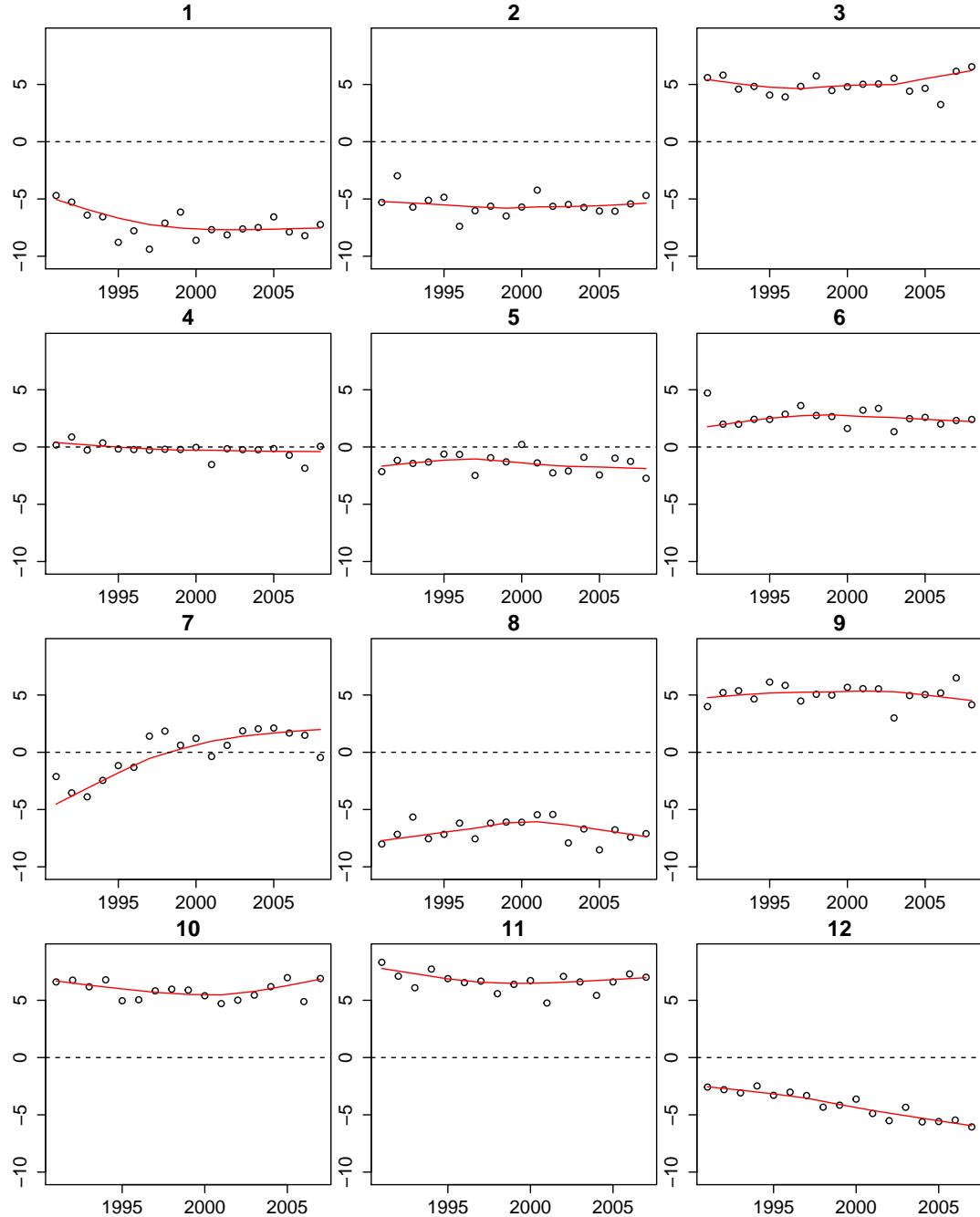


Abbildung 3.3.: Saisonuntergruppengrafiken bei linearer Schätzung (isdeg=1, ns=11) für die Produktion im Produzierenden Gewerbe in Deutschland

3. Saison-Trend Zerlegung mithilfe von Loess

In Abbildung 3.3 sind die Saisonuntergruppen einer linearen Regressionssaisonschätzung dargestellt. So sind insbesondere in den Untergrafiken 1 und 3 (Januar und März) die Unterschiede zu Abbildung 3.2 deutlich sichtbar. Im Falle vom Januar führt die lineare Schätzung zu einer Anpassung der rechtsseitigen Werte. Die Grafik weist zu Beginn einen linear abfallenden Trend auf. Dies wird bei einer linearen Schätzung als plausibel eingestuft, wogegen bei der konstanten Schätzung eindeutig sichtbar ist, dass dies im Gegensatz zu den nachfolgenden Werten unplausibel erscheint und auf z.B. besondere Ereignisse zurückzuführen sein könnte. In Untergrafik 3 (März) tritt dieses Problem an der linksseitigen Grenze auf. Bei konstanter Schätzung findet noch keine Anpassung an die Werte der letzten beiden Jahre statt, bei linearer Schätzung hingegen schon und die drei vorigen Werte werden als zu schwach eingeschätzt. Es findet also eine lineare positive Anpassung statt.

Wenn den Analysten bei konstanter Schätzung Strukturbreüche wie in Abbildung 3.2 Untergrafik 3 auffallen, sollte geklärt werden, welche möglichen Gründe es für diese gibt. Falls der Bruch erklärt werden kann, gibt es eine sinnvolle Vorgehensweise diesen in den Daten anzupassen, ohne den Grad der Saisonschätzung auf linear zu setzen. Diese weitere Einstellungsmöglichkeit wird in Kapitel 3.2.7 erläutert. Sollte es keinen triftigen Grund geben, sollten die Werte der konstanten Regressionsschätzung verwendet werden.

3.2.3. Trendparameter

Für die Trendschätzung müssen, wie bei der Saison, die Spannweite nt und der Grad der Loess-Schätzung $itdeg$ angegeben werden (vgl. Kap 2.1, Schritt 6). Zuerst geht man erneut auf die vom Analysten gewünschten Eigenschaften des Trendes ein. Er soll eine in sich geglättete Linie sein, in der nur noch mittel- bis langfristige Schwankungen auftreten. Da das bei alleiniger Betrachtung der Trendkomponente schwierig festzustellen ist, muss oder kann die Parametereinstellung über zwei unterschiedliche Arten bestimmt werden, nämlich mithilfe des Saisonparameters ns und oder der Betrachtung der Restkomponente.

Die Spannweite für die Loessschätzung des Trends (nt) muss auf jeden Fall ungerade sein. Wird vom Analysten eine gerade Zahl übergeben, so wird zu dieser der Wert eins addiert. Eine andere Bedingung ist, dass nt größer gleich sieben sein muss. Außerdem sollte nt auf jeden Fall größer als die Frequenz der Saison np sein. Es hat sich gezeigt, dass die Trendkomponente sonst noch zu viel saisonale Struktur aufweist.

Kommt man nun zuerst zur Bestimmung der Trendspannweite nt über den Saisonparameter ns .

Eine Möglichkeit den Trendparameter zu bekommen liefert Cleveland in seinen Skript. Er schlägt vor, dass folgende Formel eine gute Grundlage zur Bestimmung des Parameters nt liefert (Cleveland et al., 1990, vgl. S. 20).

$$nt \geq \frac{1.5 * np}{1 - 1.5 * ns^{-1}} \quad (3.3)$$

Da nt ungerade sein muss, nimmt sie in der Formel die nächst größte ungerade Zahl an. Anhand der Formel erkennt man, dass nt immer größer als np sein muss, da der Nenner immer Werte zwischen 0.7 und 1 annimmt (da $ns \in 5,7,\dots$) und daraus folgt, dass der Faktor mit dem np multipliziert wird immer größer als eins ist.

Für unser Beispiel, der Produktion im Produzierenden Gewerbe in Deutschland mit $ns=9$, würde man sich somit für $nt \geq 21.6 \sim 23$ entscheiden.

Betrachtet man für diesen Fall einen Diagnostikplot für die Trendschätzung, so ist dort der Loess-Input (Kreise) und Output (rote Linie) für die Schätzung der Trendkomponente

3. Saison-Trend Zerlegung mithilfe von Loess

dargestellt. Eine weitere Einschätzung über die Bestimmung des Trendparameters kann über die Betrachtung der Restkomponente (Remainder) geschehen.

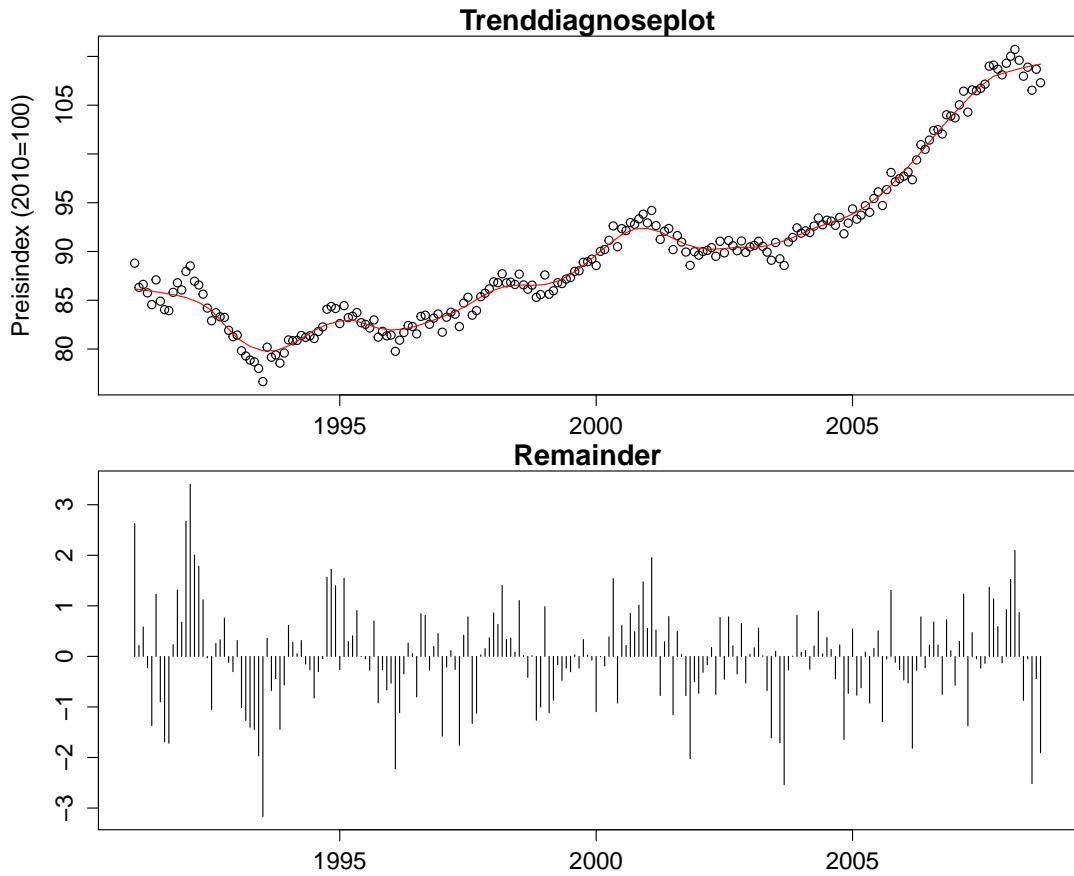


Abbildung 3.4.: Trenddiagnose und Restkomponente mit $ns=9$ und $nt=23$ für die Produktion im Produzierenden Gewerbe in Deutschland

In Abbildung 3.4 ist ein Diagnoseplot für die Trendschätzung und der Verlauf der Restkomponente für eine Spannweite von $q = 23$ dargestellt. Bei Betrachtung der Grafik ist ersichtlich, dass vor allem zu Zeitpunkten, in denen der Trend ein lokales Maxima oder lokales Minima aufweist, vermehrt Punkte auftreten, die über oder unterhalb der roten Linie liegen. In der Restkomponente ist dies ebenso zu sehen, da zu diesen Zeitpunkten vermehrt Zeitpunkte mit großer positiver oder negativer Abweichung auftreten. Dies ist oftmals ein Zeichen eines zu großen Trendparameters nt , da die konjunkturelle Schwankung des Trendes nur unzureichend abgebildet wird und diese somit in die Restkomponente mit einfließt.

Darum wird oftmals versucht den Parameter nt über die Restkomponente zu bestimmen. So wird dieser ausgehend von den von Cleveland vorgeschlagenen Parameterwert sukzessive reduziert, bis keine Autokorrelationen mehr in der Restkomponente vorhanden ist. Es ist jedoch darauf zu achten, dass dies nicht immer möglich ist. Deswegen ist vom Analysten eine alleinige Bestimmung über die Betrachtung des Trenddiagnoseplot oftmals auch sinnvoller.

Zur Bestimmung, ob Unkorreliertheit vorliegt, wird der Box-Pierce Test verwendet. Er testet die Zeitreihen auf vorhandene Autokorrelationen. Die Ablehnung der Nullhypothese

3. Saison-Trend Zerlegung mithilfe von Loess

besagt, dass die Daten der Zeitpunkte, bis zu einem bestimmten Zeitabstand, miteinander korreliert sind. Die Nullhypothese lautet somit, dass keine Korrelation zwischen den Werten der Zeitreihe eines bestimmten Zeitintervalls bestehen. Bei einem p-wert von kleiner 0.1 wird die Nullhypothese abgelehnt und die Werte werden als voneinander abhängig angenommen. Bei diesem Test muss auch immer eine Anzahl, für die Höhe des Zeitabstandes, bis zu dem getestet werden soll, angegeben werden. Da die vermehrten positiven beziehungsweise negativen Abweichungen in der Restkomponente hintereinander auftreten, hat man sich dazu entschieden, dass es vollkommen ausreichend ist, auf einem Zeitabstand (Lag) von eins zu testen.

Für die Produktion im Produzierenden Gewerbe in Deutschland erhält man ausgehend von nt=23 folgende Werte für die Teststatistik und p-Werte.

nt	Teststatistik	p-Wert	Ablehnung der Nullhypothese
23	68.86	$1.72 * 10^{-7}$	ja
21	20.96	$4.68 * 10^{-6}$	ja
19	11.72	$6.20 * 10^{-4}$	ja
17	6.53	0.01059	ja
15	2.91	0.08796	ja
13	1.01	0.3161	nein

Tabelle 3.1.: Box-Pierce Teststatistik verschiedener nt's mit ns=9 und itdeg=1 für die Produktion im Produzierenden Gewerbe in Deutschland

Aus der Tabelle 3.1 ist ersichtlich, dass man sich für einen Trendparameter von nt=13 entscheiden würde, dabei betrachtet man für diesen Fall auch den Trenddiagnoseplot, um einen Vergleich mit der größeren Trendspannweite zu erhalten.

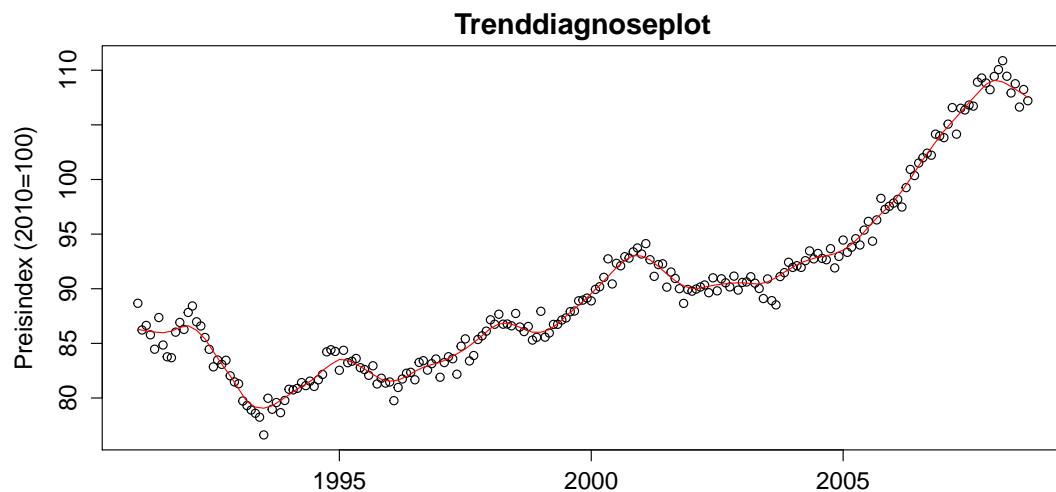


Abbildung 3.5.: Trenddiagnosegrafik mit ns=9 und nt=13 für die Produktion im Produzierenden Gewerbe in Deutschland

In Abbildung 3.5 ist dargestellt, wie im Falle der unkorrelierten Restkomponente der Trendverlauf ist. So sind hier die Konjunkturzyklen stärker ausgeprägt. Das bedeutet, die lokalen Maxima und Minima nähern sich stärker den Realisationen der Zeitreihe an.

3. Saison-Trend Zerlegung mithilfe von Loess

Außerdem wird der Umschwung, der an den Grenzen der Zeitreihe stattfindet, deutlich besser abgebildet. Während in Abbildung 3.4 die schwachen Werte an der rechtsseitigen Grenze noch keinen Einfluss auf die Trendschätzung haben, findet in Abbildung 3.5 eine Anpassung an diese statt.

Es sollte trotzdem darauf geachtet werden, den Trendparameter nicht zu gering zu halten, auch wenn dadurch noch keine Unkorreliertheit in der Restkomponente vorliegt, damit sich nicht zu viel Schwankung in der Trendkomponente wiederfindet.

Ein weiterer Parameter, der für den Trend eingestellt werden kann, ist der Grad der Regressionsschätzung im Loess-Verfahren der Trendkomponente. Defaultmäßig ist dieser linear (eins), er kann aber auch konstant oder quadratisch sein. Eine konstante Schätzung kommt in den meisten Fällen nicht in Frage, da bei steigendem oder fallendem Trend die Werte an den Grenzen damit nur unzureichend angepasst werden, da die Schätzung von den vorherigen Werten abhängig ist und somit der zu schätzende Zeitpunkt, bei steigendem oder fallendem Trend immer einen geringeren oder höheren Wert erhält, als er eigentlich bekommen müsste. Deshalb ist eine konstante Schätzung nur bei einem konstanten Trendverlauf sinnvoll.

Doch wie schaut es mit einer quadratischen Regressionsschätzung der Trendkomponente aus? Dieser sollte vor allem bei Zeitreihen mit vielen langfristigen Schwankungen bessere Ergebnisse liefern, da durch diesen logischerweise die Buckel (Umschwünge) besser abgebildet werden können. Jedoch muss an den Grenzen der Zeitreihe aufgepasst werden, dass der quadratische Trend nicht zu stark abgebildet wird und somit die Trenderwartung zu stark oder zu schwach ausfällt. Damit würden Probleme bei der Prognose entstehen, da der Trend über- oder unterschätzt wird.

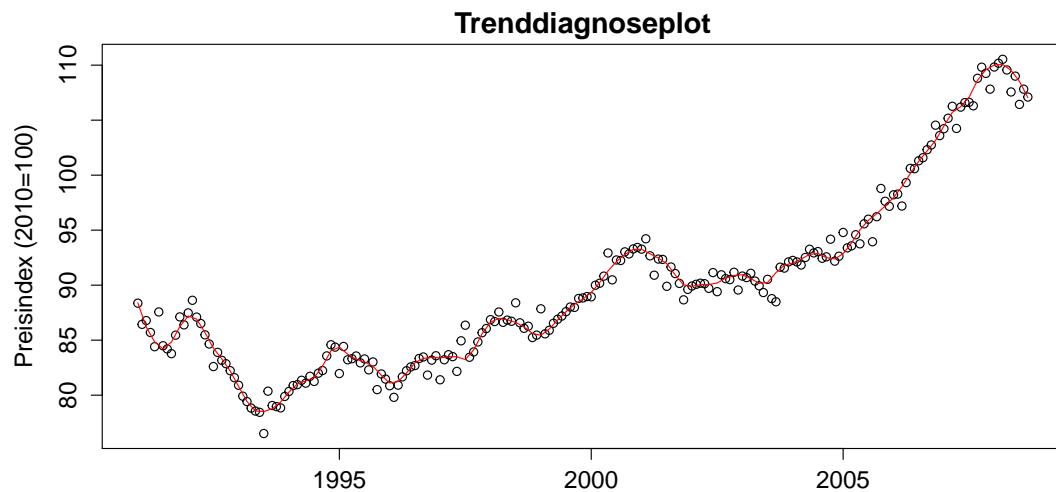


Abbildung 3.6.: Trenddiagnosegrafik mit $ns=9$ $nt=13$ und $itdeg=2$ für die Produktion im Produzierenden Gewerbe in Deutschland

Aus Abbildung 3.6 würde man schließen, dass die quadratische Schätzung eine deutlich bessere und plausiblere Anpassung an den links- und rechtsseitigen Grenzen liefern würde, als in Abbildung 3.5 bei der linearen Regressionsschätzung gezeigt wurde. Jedoch ist jetzt auch deutlich mehr Schwankung in der Trendkomponente vorhanden. Dies will man eigentlich vermeiden, da diese keine kurzfristigen Schwankungen darstellen soll. Um trotzdem die quadratische Schätzung weiter in Erwägung zu ziehen, sollte eine Erhöhung

3. Saison-Trend Zerlegung mithilfe von Loess

des Trendparameters nt vorgenommen werden, um eine glattere Linie zu bekommen. Dafür führt man erneut eine Prüfung der Höhe des Trendparameters nt durch. Man betrachtet, ausgehend von dem von Cleveland vorgeschlagenen Wert der Trendspannweite (nt=23) die Ergebnisse des Box-Pierce Tests.

nt	Teststatistik	p-Wert	Ablehnung der Nullhypothese
23	3.207	0.07365	ja
21	2.461	0.1167	nein
19	1.075	0.2999	nein

Tabelle 3.2.: Box-Pierce Teststatistik verschiedener nt's mit ns=9 und itdeg=2 für die Produktion im Produzierenden Gewerbe in Deutschland

Nach der Betrachtung der Tabelle 3.2 würde man sich jetzt für nt=21 entscheiden. Da der p-wert zu diesem Zeitpunkt aber noch sehr nahe dem Wert 0.1 liegt, habe ich mich dazu entschlossen nt=19 zu nehmen. Dieser Wert für die Trendspannweite ist größer, als der Wert bei linearer Schätzung, somit fließen bei quadratischer Regressionsschätzung der Trendkomponente deutlich mehr Punkte mit in die Schätzung ein.

Es bestehen somit zwei Möglichkeiten für die Trendparameter, entweder nt=13 und itdeg=1 oder nt=19 und itdeg=2.

3.2.4. Low-pass Parameter

Bei der Low-Pass Loess-Schätzung muss die Spannweite und der Grad der Regressions schätzung angegeben werden (vgl. Kap. 3.1.1, Schritt 4). Für beide Parameter liegen Defaultwerte vor und diese sollten nur von Analysten, die sich sehr gut mit den STL-Verfahren auskennen, verändert werden. Da saisonale Schwankungen durch die Low-Pass Schätzung entfernt werden sollen, entspricht der Defaultwert für die Spannweite der Höhe des Parameters np. Sollte dieser Wert gerade sein, erhält nl die nächste größere ungerade Zahl. Der Grad der Schätzung weist den Wert eins auf. Es handelt sich somit um eine lineare Kleinste-Quadrat Schätzung.

3.2.5. Anzahl Iterationen der äußere Schleife

Die Anzahl der Durchläufe der Robustheitsschätzung no kann durch den Analysten mithilfe eines Diagnoseplots bestimmt werden. So wird in diesem die Entwicklung der Gewichte, die bei jeder Robustheitsschätzung am Ende des Durchlaufs der äußeren Schleife entstehen, für jeden Zeitpunkt dargestellt.

Als Defaultwert ist no=15 vorgegeben. Ein größerer Wert sollte nicht gewählt werden, da sich gezeigt hat, dass dies die Laufzeit zu stark erhöhen würde. Bei 15 Umläufen sind in den allermeisten Fällen schon die optimalen Ergebnisse erreicht worden und eine Verbesserung der Ergebnisse ist nicht mehr möglich.

Doch warum sollte die Anzahl der Schleifendurchläufe der äußeren Schleife überhaupt angepasst werden? Dies dient in erster Linie zur Erkennung von Problemen, die entstehen können, wenn zu viele Werte als Ausreißer eingeschätzt werden. Außerdem kann auch die Laufzeit verkürzt werden, wenn bereits zu einem früheren Zeitpunkt eine Konvergenz in den Gewichten erreicht worden ist. Bei erneuter Zerlegung einer bereits vom Analysten analysierten Zeitreihe kann die Laufzeit somit verbessert werden, wenn man die Umläufe

3. Saison-Trend Zerlegung mithilfe von Loess

der äußeren Schleife verringert.

Als Beispiel betrachtet man hierfür den Diagnoseplot der Robustheitsgewichte (Abbildung 3.7) für die Produktion im Produzierenden Gewerbe in Deutschland mit $ns=9$ und $nt=13$ ($itdeg=1$).

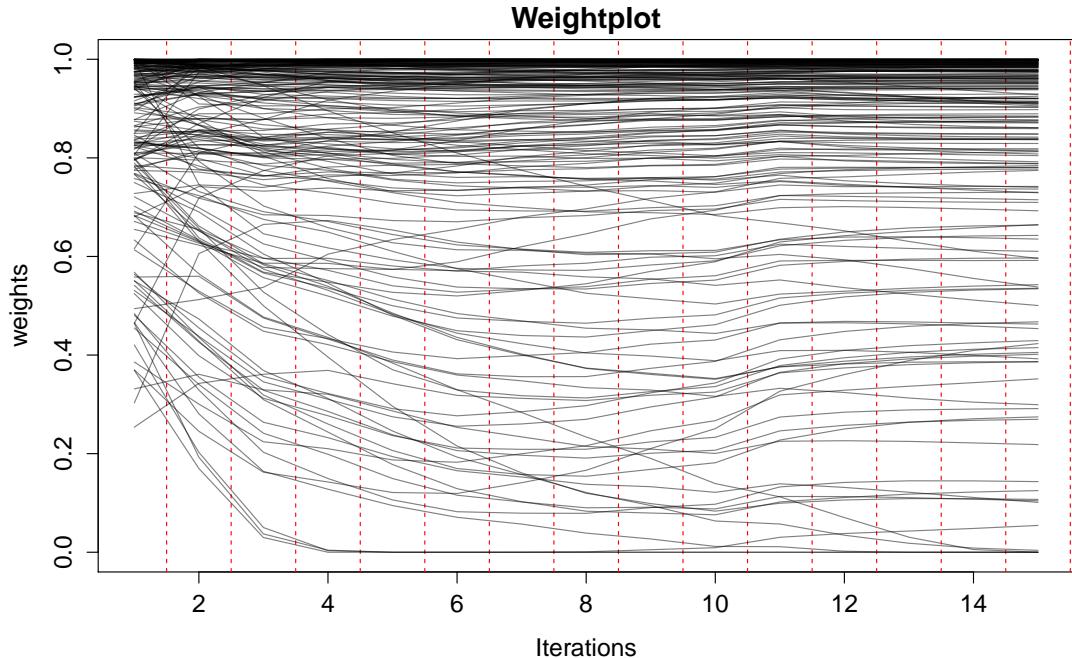


Abbildung 3.7.: Gewichtegrafik mit $ns=9$ und $nt=13$ ($itdeg=1$) für die Produktion im Produzierenden Gewerbe in Deutschland

Man erkennt in Abbildung 3.7, dass sich bereits nach circa sieben Durchläufen der äußeren Schleife eine Konvergenz in den Gewichten einstellt. So führt eine erneuter Durchlauf der äußeren Schleife zu keinen eklatanten anderen Gewichten für die verschiedenen Zeitpunkte. Wenn man nun einen Wert für die Produktion im Produzierenden Gewerbe in Deutschland für Oktober 2008 erhält und die Zeitreihe erneut, mit den gleichen Parametereinstellungen, in ihre Komponenten zerlegen will, ist es von Vorteil, die äußere Schleife nur noch zehn-mal durchlaufen zu lassen, um die Laufzeit zu verringern, da sich bei höheren Iterationen keine neuen Gewichte ergeben und somit die Komponentenschätzung keine signifikanten besseren Ergebnisse liefert.

Ein anderes Problem, das auftreten kann, ist, dass zu viele Werte aus der Schätzung herausgenommen werden. In diesem Fall müssen entweder die kompletten Parametereinstellungen überdacht werden, oder man bricht zu einem früheren Zeitpunkt ab, zu dem noch genügend Werte in die Rechnung mit einfließen. Eine andere Möglichkeit, dieses Problem zu handhaben, findet man in Kapitel 3.2.7.

In Abbildung 3.8 ist der Diagnoseplot und ein Histogramm für die Werte der Gewichte, mit den in Kapitel 3.2.4 vorgeschlagenen Parameterwerte ($ns=9$, $nt=19$, $itdeg=2$), dargestellt. Anhand der Grafik, die die Entwicklung der Gewichte zeigt, ist erkennbar, dass mit zunehmender Schleifendurchlaufzahl immer mehr Werte, für die Schätzung der Komponenten, herausgewichtet werden. Dies kann vor allem zu Problemen bei der Saisonberechnung

3. Saison-Trend Zerlegung mithilfe von Loess

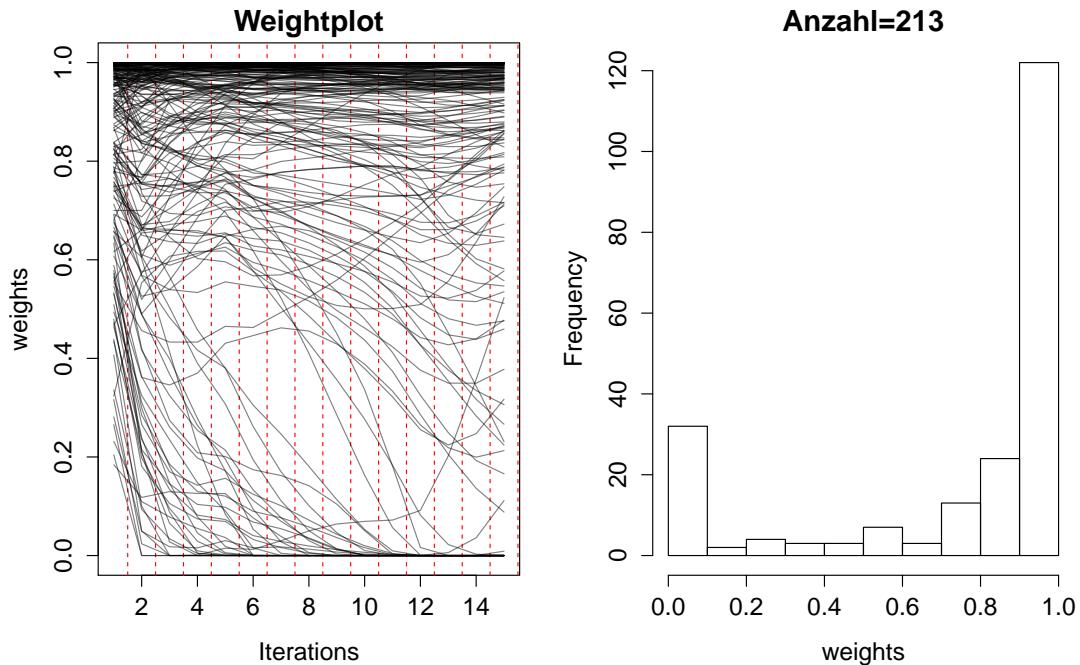


Abbildung 3.8.: Gewichtetgrafik und Histogramm der Gewichte mit $ns=9$ und $nt=19$ ($itdeg=2$) für die Produktion im Produzierenden Gewerbe in Deutschland

führen, da die Möglichkeit besteht, dass zu viele Robustheitsgewichte einer Saisonuntergruppe den Wert null annehmen und somit die Schätzung einzelner Zeitpunkte nur noch von einem Zeitpunkt abhängig ist.

In der rechten Grafik sieht man die Anzahl der herausgewichteten Werte. Zum Schluss weisen bereits über 30 Werte Gewichte von null oder nahe null auf. Dies sind circa 20% der kompletten Ausgangswerte. Das ist oftmals nicht hinnehmbar, da die Komponenten nicht nur anhand einzelner Werte, sondern mithilfe möglichst vieler Zeitpunkte bestimmt werden sollen. Nun kann der Analyst entscheiden, ob er die Berechnung entweder zu einen früheren Zeitpunkt beendet (z.B. nach sechs Umläufen) oder sich für komplett andere Parametereinstellungen entscheidet.

Für den Fall, der Produktion im Produzierenden Gewerbe in Deutschland, hat man nun für die wichtigsten Parametereinstellungen Werte gefunden. Man würde sich für folgende Parameter entscheiden: $ns=9$, $isdeg=0$, $nt=13$, $itdeg=1$ und $no=10$.

3.2.6. Anzahl Iterationen der innere Schleife

Der Parameter ni bestimmt die Anzahl, wie oft die innere Schleife durchlaufen wird. Für den Defaultwert von dieser, unterscheidet man zwei unterschiedliche Fälle.

Im ersten Fall ist der Wert des Parameters $no > 0$, dabei hat sich gezeigt, dass der einmalige Durchlauf der inneren Schleife vollkommen ausreichend ist, um für die Komponentenschätzung bereits konvergente Ergebnisse zu erzielen.

Falls der Parameter $no = 0$, muss die innere Schleife öfter durchlaufen werden. Der Defaultwert ist nun zwei. Dies reicht laut Cleveland bereits aus, um konvergente Ergebnisse zu erhalten. Versuche von diesem, mithilfe von Abbruchkriterien, haben keine Erkenntnisse geliefert, die für eine höhere Einstellung von ni sprechen würden (Cleveland et al., 1990,

3. Saison-Trend Zerlegung mithilfe von Loess

vgl. S. 14) .

Man betrachtet für die Produktion im Produzierenden Gewerbe die durchschnittlichen und maximalen absoluten Abweichungen, die zwischen einem zweimaligen und einem fünfmaligen Durchlaufen der inneren Schleife entstehen ($ni=2$ zu $ni=5$). In beiden Fällen wird die äußere Schleife nicht durchlaufen.

Art	Saison	Trend
Maximale absolute Abweichung	0.1980536	0.2424799
Durchschnittliche absolute Abweichung	0.0412737	0.0202778

Tabelle 3.3.: Abweichungen, die bei mehrmaliger Iteration der inneren Schleife entstehen

Man erkennt, dass die maximale Abweichung des Trendes zwischen den beiden Schätzungen nur 0.24 (circa 0.3%) beträgt. Für die Saison ergibt sich 0.19. Die durchschnittliche Abweichung in beiden Komponenten ist deutlich geringer, sie entspricht nur ungefähr 10% der maximalen Abweichung. Die Abweichung der Saison ist prozentual nicht darstellbar und bewegt sich auch noch im Rahmen. Somit ist ein zweimaliges Durchlaufen der inneren Schleife durchaus ausreichend, um aussagekräftige Ergebnisse zu bekommen.

3.2.7. Robustheitsschätzung

Für die Robustheitsschätzung können zwei Parameter verändert werden, wobei bei einem (bei rw) die Vorgehensweise der Schätzung an sich nicht verändert wird, sondern vom Benutzer direkt die Robustheitsgewichte übergeben werden.

Betrachtet man zuerst diese Möglichkeit, den Parameter rw . Für diesen muss man für jeden Zeitpunkt ein individuelles Gewicht angeben. Er sollte verwendet werden, wenn der Analyst aufgrund von Vorinformationen eigene Gewichte übergeben will oder ein Bruch in den Daten zu erkennen ist, der aufgrund der bisher wenig vorhandenen Werte an der rechtsseitigen Grenze heraus gewichtet wird. Die Plausibilität des Bruches muss natürlich immer vom Analysten überprüft werden. Es ist darauf zu achten, dass bei Übergabe von Gewichten keine erneute Robustheitsschätzung stattfindet und nur die innere Schleife durchlaufen wird.

Die Funktionsweise des Parameters rw soll mithilfe der Produktion im Produzierenden Gewerbe in Deutschland gezeigt werden. Ruft man sich Abbildung 3.2 zurück ins Gedächtnis, so sieht man in der Saisonuntergruppengrafik für den Monat März (Nr. 3), dass die letzten beiden Werte oberhalb der roten Linie liegen und somit stark heraus gewichtet werden. Man nimmt nun an, dass die nachfolgenden Märzwerke ebenfalls das Niveau der letzten beiden Werte besitzen und diese somit plausible Werte darstellen, an die die Kurve der Loess-Schätzung angepasst werden soll.

Dazu schaut man sich zuerst die Robustheitsgewichte der letzten Märzwerke an.

Zeitpunkt	03.2003	03.2004	03.2005	03.2006	03.2007	03.2008
Robustheitsgewicht	0.85628	0.91842	0.95909	0.26694	0.54139	0.12156

Tabelle 3.4.: Bei Zerlegung der Produktion im Produzierenden Gewerbe in Deutschland erhaltenen Robustheitsgewichte für den Monat März ($ns=9$, $nt=13$, $no=10$)

3. Saison-Trend Zerlegung mithilfe von Loess

Man erkennt aus der Tabelle, dass die letzten beiden Werte eher geringe Gewichte, im Gegensatz zu den vorherigen Gewichten besitzen. Der Wert für 2006 ist deswegen so gering, da dessen Niveau sehr stark unterhalb des erwarteten Saisonwertes liegt. Nun wird versucht eine Anpassung an die letzten beiden Werte zu erzielen, indem dem Programm für die Märzwerte andere Robustheitsgewichte übergeben werden.

Für die Anzahl der Gewichte, die verändert werden sollen, nimmt man folgende Faustformel:

Ab dem Zeitpunkt, zu dem der Strukturbruch sichtbar ist, sollen die X vor dem Strukturbruch liegende Robustheitsgewichte verändert werden. Der Wert für X sollte die Hälfte der Saisonspannweite bzw. Trendspannweite sein, je nachdem in welcher Komponente der Strukturbruch festgestellt worden ist. Es sollten auch alle Gewichte der darauffolgenden Zeitpunkte verändert werden.

In der Tabelle 3.5 sieht man die Gewichte, die für die vorhin durchgeführten STL-Zerlegung bei erneuter Komponentenschätzung benutzt werden.

Zeitpunkt	03.2003	03.2004	03.2005	03.2006	03.2007	03.2008
Robustheitsgewicht	0.70	0.50	0.30	0.00	0.95	0.95

Tabelle 3.5.: Vom Analysten übergebene Robustheitsgewichte für den Monat März

Die Robustheitsgewichte nehmen nur Werte zwischen null und eins an. Die Werte nach dem Strukturbruch erhalten hohe Gewichte, die Werte davor, niedrige Gewichte, wobei eine linksseitige bzw. rechtsseitige Abstufung stattfindet. Für die übrigen Monate werden die Gewichte der Ausgangszerlegung verwendet. Diese sind im Output des STL-Programms in R vorhanden.

In Abbildung 3.9 erkennt man nun für den Monat März (Nr. 3), dass es zu einer Anpassung an die Werte der letzten beiden Monate gekommen ist, wie man es sich vorgestellt hat. Mit dem Parameter *rw* kann somit eine Anpassung an die Werte der rechtsseitigen Grenze der Saisonuntergruppen vorgenommen werden. Es kann ebenso eine Anpassung der Trendkomponente erwirkt werden.

Ein andere Einstellungsmöglichkeit, die bei der Robustheitsschätzung vorgenommen werden kann, ist der Parameter *wf* (weightsfactor). Er gibt an, wie stark Ausreißer bei der Gewichtung heraus gewichtet werden.

Dazu schaut man sich noch einmal die Formel Nr. 2.1 und 3.2 an.

$$K(u) = \begin{cases} (1 - u^3)^3 & \text{falls } 0 < u < 1 \\ 0 & \text{sonst} \end{cases} \quad (3.4)$$

$$w_{ro}(x_j) = K\left(\frac{|R(x_j)|}{wf * \text{med}(|R(x)|)}\right) \quad (3.5)$$

In der Formel 3.2 steht im Nenner der Wert sechs. Dieser wird, wie in Formel Nr. 3.4 gezeigt, durch den Parameter *wf* ersetzt. Er gibt somit den Faktor, der vor dem Median eingesetzt wird an und ist, wie in Formel 3.2, defaultmäßig sechs. Wenn man einen großen Faktor angibt, sinkt der Wert innerhalb der Kerndichtefunktion und die Robustheitsgewichte nehmen größere Werte an, wenn man einen kleineren Faktor angibt, steigt der Wert innerhalb der Kerndichtefunktion und mehr Werte werden bei der Schätzung der Komponenten herausgewichtet.

3. Saison-Trend Zerlegung mithilfe von Loess

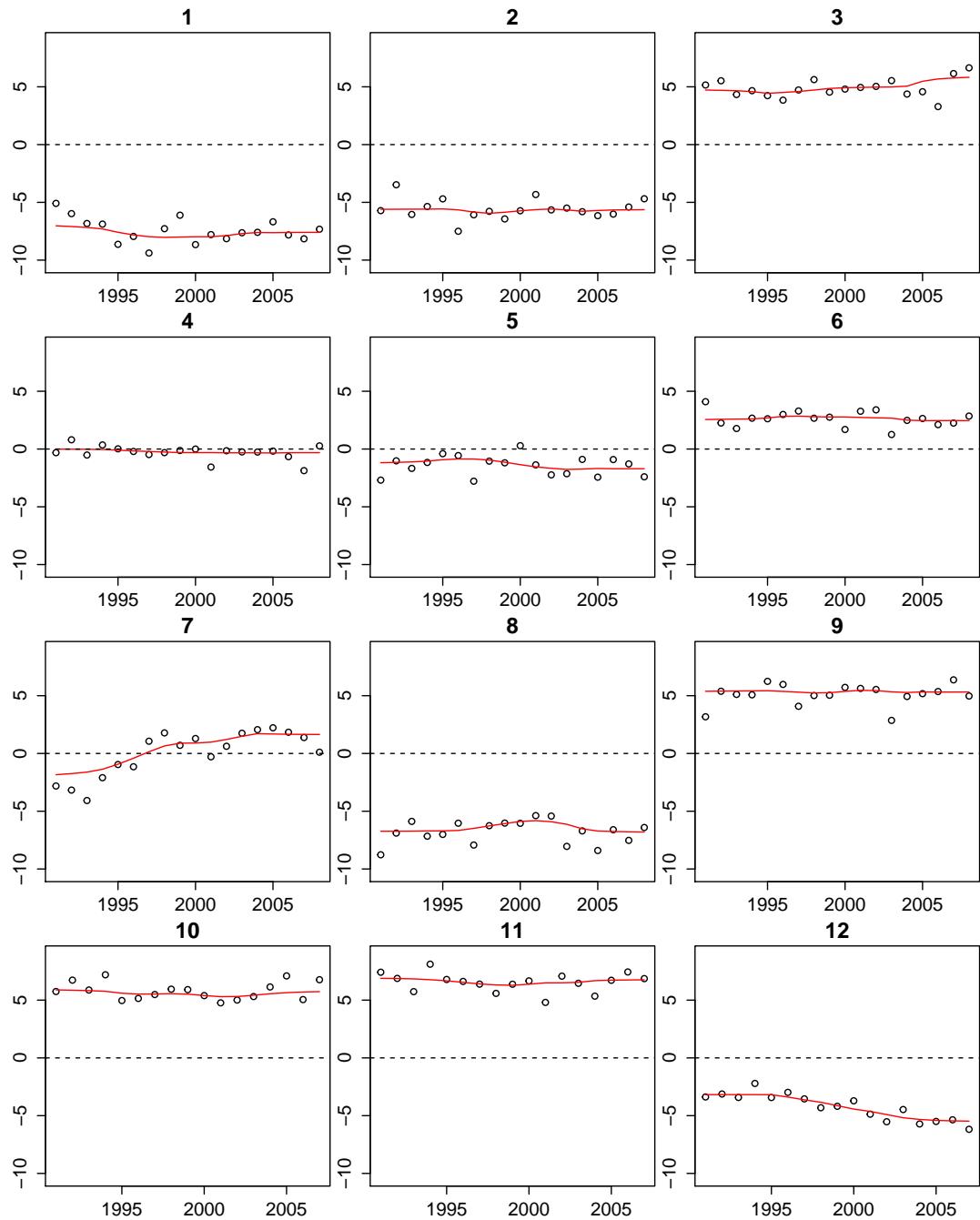


Abbildung 3.9.: Saisonuntergruppengrafiken mit eigenen übergebenen Robustheitsgewichten für die Produktion im Produzierenden Gewerbe in Deutschland

Versuche mit unterschiedlichen Zeitreihen haben gezeigt, dass es nur sinnvoll ist, größere Werte als sechs für den Faktor anzugeben, da bei $wf=6$ oftmals schon zu viele Werte geringe Robustheitsgewichte erhalten.

Betrachtet man für einen großen Faktor ($wf=8$) den Diagnoseplot für die Robustheitsgewichte. Dazu verwendet man, für die Zerlegung der Produktion im Produzierenden Gewerbe in Deutschland die gleichen Parametereinstellungen, wie in Abbildung 3.8 ($ns=9$,

3. Saison-Trend Zerlegung mithilfe von Loess

nt=19 und itdeg=2), in der man festgestellt hat, dass zu viele Zeitpunkte, herausgewichtet werden.

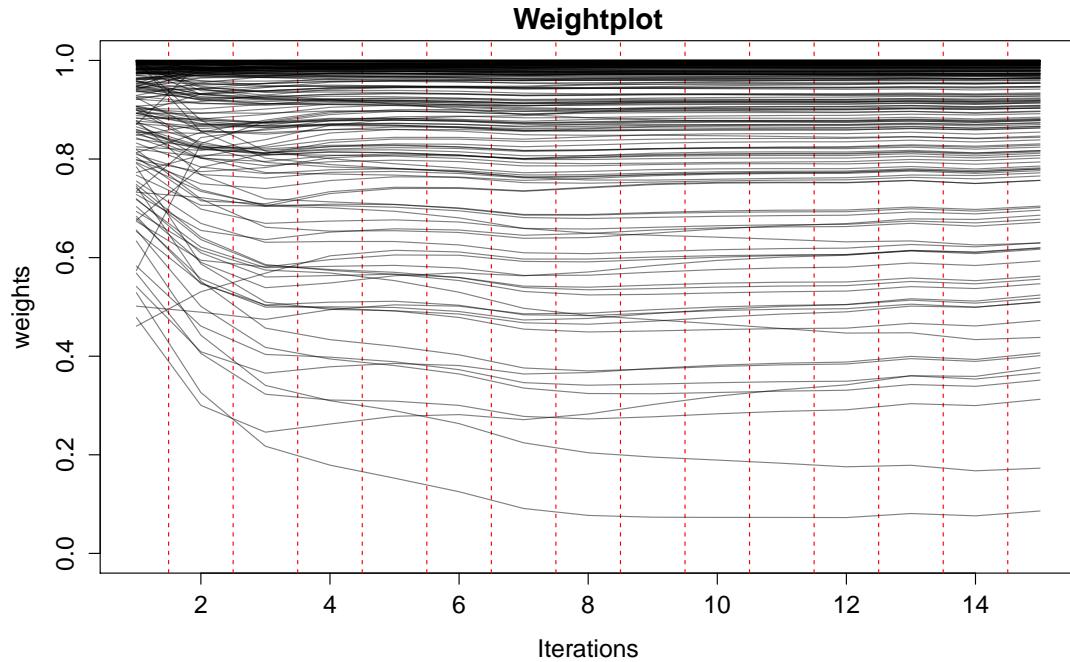


Abbildung 3.10.: Gewichtegrafik mit Parameter wf=8 für die Produktion im Produzierenden Gewerbe in Deutschland

In Abbildung 3.10 erkennt man, dass bei diesem Wert von wf (wf=8) nun keine Werte mehr als komplette Ausreißer eingestuft werden und eine Konvergenz in der Gewichteschätzung, im Gegensatz zu Abbildung 3.8 sichtbar ist. Somit kann mit Veränderung des Parameters wf erreicht werden, dass die Robustheitsgewichte nicht zu geringe Werte annehmen. Außerdem könnte man in diesem Fall, bei erneutem Durchlauf der Zerlegung, wieder nach circa 6-8 Schleifendurchläufen abbrechen (no=7).

Der Parameter wf ist somit stark von den Wünschen des Analysten abhängig. Möchte er die Komponenten möglichst mit allen Werten der Ausgangszeitreihe bestimmen oder dürfen auch viele Ausreißer vorhanden sein? Ist die Annahme von externen Effekten sinnvoll für die Ausgangszeitreihe, oder nicht? Normalerweise wird der Parameter aber nicht verändert.

3.2.8. Computationale Parameter

Da beim Loess-Verfahren für jeden Punkt die jeweiligen Gewichte berechnet werden müssen und eine Regressionschätzung durchgeführt wird, ist die STL-Methode ein sehr zeitaufwendiges Verfahren. Versuche die Laufzeit zu verkürzen, haben gezeigt, dass nicht unbedingt an jedem Zeitpunkt eine Loess-Schätzung nötig ist. In den meisten Fällen genügt es, einzelne Werte zu schätzen und die anderen durch lineare Interpolation zu berechnen. Dies hat kaum negative Auswirkungen auf die Ergebnisse und führt zu einer Verbesserung der Laufzeit. Die Verkürzung der Laufzeit war vor allem bei der Vorstellung des STL-Programms im Jahre 1990, aufgrund der damaligen geringen Rechnerleistungen, von sehr großer Bedeutung.

3. Saison-Trend Zerlegung mithilfe von Loess

Der Abstand der Punkte, die mithilfe von Loess geschätzt werden sollen, können über die Parameter nsjump (Saison), ntjump (Trend) und nljump (low-pass) bestimmt werden. Defaultmäßig wird die jeweilige Spannweite durch zehn geteilt (z.B ns/10) und dann aufgerundet. Ist die Spannweite des Trends mit dem Wert 23 angegeben, so ergibt sich für ntjump der Wert drei. Somit wird eine Loess-Schätzung für den ersten, $1+x^*ntjump$ und den N-ten Wert durchgeführt. Die übrigen Werte erhält man durch lineare Interpolation (Cleveland et al., 1990, vgl. S. 20-22).

$$\hat{y}(x_0) = \hat{y}(x_{-1}) + \frac{\hat{y}(x_1) - \hat{y}(x_{-1})}{x_1 - x_{-1}} \quad (3.6)$$

x_{-1} ist der Zeitpunkt der letzten Loess-Schätzung vor dem Zeitpunkt x_0 .

x_1 ist der Zeitpunkt der ersten Loess-Schätzung nach dem Zeitpunkt x_0 .

Um zu zeigen, dass es bei der Berechnung nicht jedes einzelnen Zeitpunktes, zu keiner Beeinträchtigung in den Ergebnissen kommt, betrachtet man die durchschnittlichen absoluten Abweichungen in den einzelnen Komponenten, für unser Beispiel, der Produktion im Produzierenden Gewerbe in Deutschland.

Art	Saison	Trend	Laufzeit (in s)
Maximale absolute Abweichung	0.1215677	0.4032692	-
Durchschnittliche absolute Abweichung	0.0169900	0.0309778	0.634

Tabelle 3.6.: Abweichung, die durch die computationalen Parameter bei der Zerlegung der Produktion im Produzierenden Gewerbe in Deutschland entsteht

In der Tabelle 3.6 zeigt sich, dass die Unterschiede ziemlich gering sind. So liegen die durchschnittlichen Abweichungen der Trend- und Saisonkomponente unterhalb von einem Prozent. Auch in der Laufzeit stellt sich eine Verbesserung dar, die jedoch aufgrund der allgemein geringen Laufzeit nicht ausschlaggebend ist.

Für die Berechnung der Laufzeit wurde die durchschnittliche Abweichung von zehn Ausführungen genommen. Der geringe Unterschied liegt vor allem an der geringen Anzahl von Zeitpunkten. Würde man sich täglich Daten über mehrere Jahre anschauen, so würden die computationalen Parameter bereits einen deutlich höheren Einfluss auf die Laufzeit haben und die Unterschiede in der Laufzeit würden größer werden.

3.2.9. Multiplikative Umformung

Der additive Ansatz hat oftmals Probleme bei einer sich abschwächenden oder verstärkenden Saison, da bei diesem die Anpassung zu schwach oder stark ausfällt und somit die Werte besonders starker oder schwacher Monate als Ausreißer eingestuft werden. Deswegen besteht die Möglichkeit, die Komponentenschätzung multiplikativ durchzuführen. Dazu muss die Input Zeitreihe logarithmiert werden und nach Durchführung der Zerlegung die Ergebnisse in die gewünschte Form gebracht werden. Erörtert man die mathematischen Vorgänge, die bei dieser Umformung passieren.

$$\log(y(x_0)) = T(x_0) + S(x_0) + R(x_0) \quad (3.7)$$

$$\Rightarrow y(x_0) = \exp(T(x_0) + S(x_0) + R(x_0)) \quad (3.8)$$

$$\Rightarrow y(x_0) = \exp(T(x_0)) * \exp(S(x_0)) * \exp(R(x_0)) \quad (3.9)$$

3. Saison-Trend Zerlegung mithilfe von Loess

Die erhaltenen Ergebnisse müssen also exponiert werden, um ein multiplikatives Ergebnis zu erhalten.

Im multiplikativen Fall stellt eine Komponente eine Referenz dar, von der die anderen Multiplikative abhängig sind. Dies sollte die Trendkomponente sein. Mithilfe dieses Wissens kann man die Ergebnisse interpretieren. Zum Beispiel bei einer monatlichen Zeitreihe impliziert ein Wert von 0.9 der Saisonkomponente, dass der Monat 0.9 mal der Wert Trendkomponente ist. Ein Wert kleiner eins spricht für einen schwachen Monat und ein Wert größer eins für einen starken Monat. Betrachtet man die Zerlegung der Produktion im Produzierenden Gewerbe in Deutschland und die exponierten Werte für das Jahr 2000.

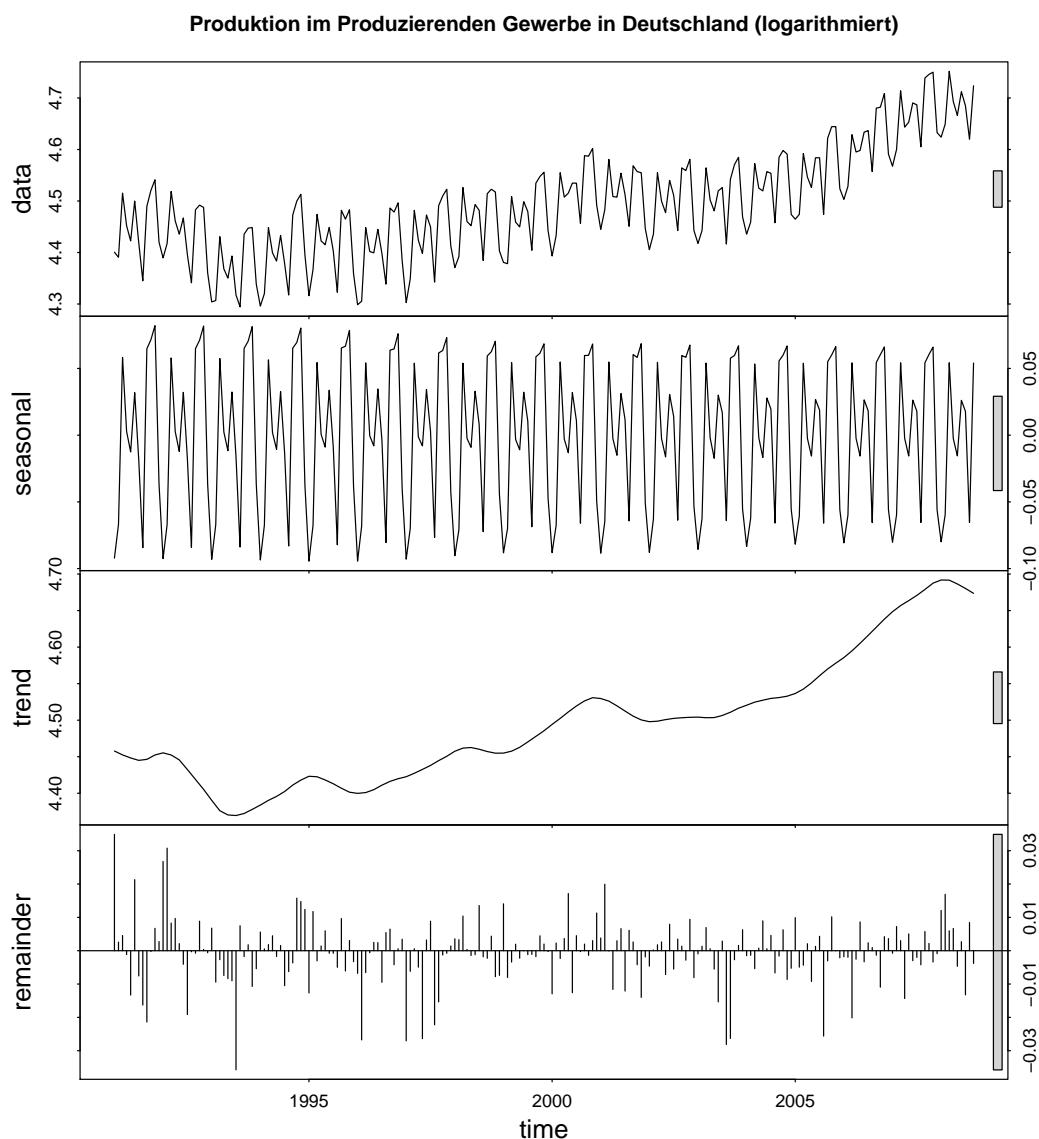


Abbildung 3.11.: Multiplikative Zerlegung der Produktion im Produzierenden Gewerbe in Deutschland

3. Saison-Trend Zerlegung mithilfe von Loess

Monat	Trend	Saison	Rest
Januar	89.496	0.9157	0.9871
Februar	89.867	0.9348	1.0023
März	90.240	1.0563	0.9977
April	90.638	0.9970	1.0037
Mai	91.037	0.9869	1.0173
Juni	91.410	1.0325	0.9875
Juli	91.784	1.0109	1.0045
August	92.103	0.9362	0.9997
September	92.423	1.0615	1.0020
Oktober	92.628	1.0616	0.9986
November	92.834	1.0707	1.0030
Dezember	92.780	0.9517	1.0114

Tabelle 3.7.: Werte für das Jahr 2000 der multiplikativen Zerlegung für die Produktion im Produzierenden Gewerbe in Deutschland

Man erkennt in Tabelle 3.7, dass es sich bei März, September, Oktober und November um starke Monate und bei Januar, Februar und August um schwache Monate handelt. Die Restkomponente weist in unserem Fall für das Jahr 2000 nur geringe Abweichungen auf, da die Werte knapp um eins schwanken. Multipliziert man die Saisonkomponente mit der Trendkomponente, erhält man die Werte, die durch das Modell erklärt werden. Die Zeitreihe weist einen positiven Trend auf, da die Realisationen der Trendkomponente mit jedem Monat steigt. Man erkennt aus Abbildung 3.11 und der Tabelle, dass die multiplikative Berechnung ebenso gute Ergebnisse, wie die additive Zerlegung liefert (vgl. Abbildung 1.1). Jedoch ist zu beachten, dass in der Grafik die Werte noch nicht exponiert worden sind und es sich somit in dieser noch um eine additive Zerlegung handelt. Somit ist das additive Verfahren dem multiplikativen Verfahren vorzuziehen, wenn sich keine verstärkende Saison zeigt und da das multiplikative Verfahren immer mit mehr Rechenaufwand verbunden ist.

3.2.10. Vergleich mit dem alten STL-Programm

Wie am Anfang bereits erwähnt, geht das Verfahren auf Cleveland zurück, der bereits ein STL-Programm in R implementiert hat, das auf den selben Algorithmus wie meine Erweiterung des STL-Programms, zurückgreift. Es weist jedoch weniger Einstellungsmöglichkeiten auf und mehrere Diagnosegrafiken sind noch nicht vorhanden.

Neue Möglichkeiten und Parametereinstellungen, die vorgenommen werden können, sind:

- Übergabe von Robustheitsgewichten (rw)
- Einstellung des Gewichtefaktors der Robustheitsschätzung (wf)
- Möglichkeit einer quadratischen Loess-Schätzung
- Bestimmung der Trendspannweite mithilfe des Box-Pierce-Tests
- Diagnoseplot der Saisonkomponente (Saisonuntergruppengrafiken)
- Diagnoseplot der Trendkomponente
- Diagnoseplot für die Entwicklung der Robustheitsgewichte
- Diagnoseplot für die Verteilung der Robustheitsgewichte (Histogramm)

Die neuen Möglichkeiten erleichtern es die optimalen Einstellungen zu finden.

4. Prognose

Bevor näher auf die Prognose eingegangen wird, betrachtet man zuerst, was man unter einem deterministischen und einem stochastischen Prozess versteht.

Ein deterministischer Prozess ist ein Prozess, bei dem jeder Vorgang von vorherigen Vorgängen abhängig ist und insbesondere zukünftige Ereignisse ebenso durch die vorherigen Vorgänge eindeutig bestimmt werden können.

Ein stochastischer Prozess ist ein Prozess, dessen Vorgänge zufälligerweise entstehen. Deswegen können die zukünftigen Ereignisse nur aufgrund von Wahrscheinlichkeitsannahmen, die mithilfe der vorherigen Vorgänge bestimmt worden sind ermittelt werden.

Allgemein wird eine Zeitreihe meistens in einen deterministischen und stochastischen Teil aufgeteilt, um zukünftige Werte prognostizieren zu können.

$$y_t = f(t) + \epsilon_t \quad (4.1)$$

Die Funktion $f(\cdot)$ stellt den deterministischen Teil dar, der Fehlerterm ϵ den stochastischen, die Unsicherheit.

Bei der Prognose für die Saison-Trend Zerlegung müssen für die zukünftigen Werte die einzelnen Komponenten der Zeitreihe geschätzt werden. Den erwarteten Wert zum Zeitpunkt $E[\hat{y}_{t+h}]$ erhält man somit mithilfe folgender Formel.

$$E[\hat{y}_{t+h}] = E[\hat{S}_{t+h}] + E[\hat{T}_{t+h}] + E[\hat{R}_{t+h}] \quad (4.2)$$

Doch welche Komponenten stellen den deterministischen und stochastischen Teil dar und wie bekommt man die prognostizierten Werte für die jeweiligen Komponenten?

Die Saison (\hat{S}_{t+h}) ist in sich möglichst konstant und weist über den Zeitverlauf eine gleichbleibende Struktur auf. Somit ist diese eindeutig über ihre vorherigen Werte definiert und es muss kein Modell geschätzt werden. Deswegen genügt es den letzten Wert der entsprechenden Saisonuntergruppe konstant fortzuschreiben. Wer trotzdem noch die letzten Werte der jeweiligen Saisonuntergruppe mit einfließen lassen will, für den besteht die Möglichkeit eine Loess-Schätzung durchzuführen. So wird an dem zu prognostizierenden Zeitpunkt eine Loess-Schätzung vorgenommen. Bei dieser findet damit noch eine leichte Anpassung an die letzten Werte der Saisonuntergruppe statt. Das genaue Vorgehen wird in Kapitel 4.1.1 erläutert. Jedoch ist auch in diesem Fall der Schätzwert eindeutig über die vorherigen Saisonwerte definiert.

Bei der Saisonkomponente handelt es sich somit um eine deterministische Komponente. Die Prognose nimmt feste Werte an und weist keine Unsicherheiten auf. Daraus folgt, dass sie nur Einfluss auf die Punktschätzung, aber nicht auf die Prognoseintervalle hat.

Die Trendkomponente (\hat{T}_{t+h}) weist hingegen nicht immer die gleiche Struktur (den gleichen Verlauf) auf. Deswegen ist es für die Prognose hilfreich, zuerst ein Modell zu schätzen und damit die zukünftigen Werte der Trendkomponente zu prognostizieren.

Im späteren Verlauf werden zwei Modelle betrachtet: das ARIMA-Modell, vergleiche Kapitel 4.1.3 und Random-Walk mit Drift, vergleiche Kapitel 4.1.2. Diese Modelle bestehen

4. Prognose

wiederum aus einem deterministischen und stochastischen Teil. Im ARIMA-Modell ist der deterministische Teil abhängig von den vorherigen Werten, beim Random-Walk mit Drift vom Zeitverlauf. Für die Prognose der Trendkomponente erhält man somit folgende Formel, wobei $f(\cdot)$ der deterministische Teil des jeweiligen Modells ist und ϵ der stochastische Teil, der sogenannte Fehlerterm (die Abweichungen die nicht über das Modell erklärt werden können).

$$E[\hat{T}_{t+h}] = f(t+h) + E[\hat{\epsilon}_{t+h}] \quad (4.3)$$

Die Annahmen, die für die einzelnen Teile des Modells angenommen werden müssen, werden in den nachfolgenden Kapiteln erläutert.

Die Restkomponente ist der stochastische Teil der Zerlegung. Für sie muss eine Verteilungsannahme getroffen werden. Im einfachen Fall nimmt man ein gaußsches weißes Rauschen mit Erwartungswert null und Varianz σ^2 an ($R_t \sim N(0, \sigma^2)$ mit $t \in 1, \dots, n$). Beim weißen Rauschen sind die einzelnen Realisationen unabhängig voneinander.

Doch wie sinnvoll ist diese Annahme? Wenn bei der Bestimmung des Trendparameters darauf geachtet worden ist, dass die Restkomponente unkorreliert ist, ist die Annahme eines Erwartungswertes von null sehr sinnvoll, da die Restkomponente um den Nullpunkt schwankt. Im anderen Fall sollte darauf geachtet werden, dass eventuell, wie bei der Trendkomponente, ein Modell für die Restkomponente geschätzt wird, da die Möglichkeit besteht, dass in dieser noch Struktur ist und damit vergangene Werte die zukünftigen Realisationen der Restkomponente beeinflussen.

Die Annahme der konstanten Varianz von σ^2 stellt sich als weniger sinnvoll heraus, da wie später im Ausblick erläutert wird, es Zeiträume gibt, zu denen es vermehrt stärkere Abweichungen gibt, als in anderen Zeiträumen. Deswegen kann eine Vorhersage des Parameters σ^2 sehr viel genauere Ergebnisse über die Unsicherheit, die besteht, liefern. Im weiteren Verlauf wird jedoch auf die einfache Annahme des gaußschen weißen Rauschen zurückgegriffen.

Da der Erwartungswert beim gaußschen weißen Rauschen null ist, fließt die Restkomponente nur in die Schätzung der Prognoseintervalle mit ein. Sie verändert nur die Unsicherheit zukünftiger Werte.

Falls die Annahme einer unkorrelierten Restkomponente nicht eintritt, kann auf diese auch ein Modell geschätzt werden. Die Schätzung läuft analog der Schätzung der Trendkomponente ab. Im folgenden Teilkapitel wird auf unser Beispiel der Produktion im Produzierenden Gewerbe zurückgegriffen, wobei die Annahme einer unkorrelierten Restkomponente besteht. Beim empirischen Beispiel trifft die Annahme einer unkorrelierten Restkomponente dagegen nicht ein.

Bei einem schon in R implementierten Verfahren für die Prognose einer Saison-Trend zerlegten Zeitreihe, wird für die Vorhersage die Trend- mit der Restkomponente zusammengelagert. Auf die zusammengelegte Zeitreihe wird ein Modell geschätzt und die Prognosewerte berechnet (Hyndman and Athanasopoulos, 2012, vgl. Kap. 6.6). Im Kapitel 4.3.2 wird ein Vergleich gezogen, welche der beiden Methode bessere Ergebnisse liefert.

4.1. Methoden

4.1.1. Loess-Prognose

Die Prognose mithilfe von Loess findet bei der Saisonkomponente Anwendung. Sie läuft äquivalent der in Kapitel 2.2 erläuterten Vorgänge einer lokal gewichteten Regression ab.

4. Prognose

Der Unterschied besteht darin, dass nur linksseitige Werte für die Schätzung vorhanden sind und zum schätzenden Zeitpunkt keine Realisation vorliegt. So wird der Schätzwert nur über die vorherigen Werte bestimmt. Der Wert von β_0 gibt dann erneut, den zu prognostizierenden Wert an.

Der Vorteil der Loess-Prognose, gegenüber der Verfahrensweise einfach den letzten Saisonuntergruppenwert zu nehmen, besteht darin, dass leichte Schwankungen, die an der rechten Grenze der Saisonuntergruppe entstehen, auch in die Prognose des saisonalen Wertes mit einfließen und somit mit abgebildet werden.

4.1.2. Random Walk

Bei einem Random-Walk wird angenommen, dass die Realisation y_t nur von ihrem vorherigen Wert und einer Unsicherheit (Fehlerterm) abhängig ist. Der Fehlerterm weist zufällige Werte auf, die mithilfe einer Wahrscheinlichkeitsannahme bestimmt werden. Hier betrachtet man die Modellgleichung eines Random-Walk (Foster, a, vgl. S. 5-10).

$$y_t = y_{t-1} + \epsilon_t \quad (4.4)$$

$$\Delta y_t = \epsilon_t \quad (4.5)$$

Für den Fehlerterm wird weißes Rauschen mit Erwartungswert Null und Varianz σ^2 angenommen. Betrachtet man die Zuwachsrate, so erkennt man, dass diese nur von ϵ_t abhängig ist und somit zufällige Werte annimmt.

Der Erwartungswert dieses Modells ist konstant über den Zeitverlauf ($E(y_t) = E(y_{t-1}) = y_0$) und die Varianz ist abhängig von der Zeit ($Var(y_t) = t\sigma^2$). Das heißt, die Varianz ist nicht endlich und mit zunehmenden in der Ferne liegenden Prognosezeitpunkten wird das Prognoseintervall linear größer.

Jedoch ist in den meisten Fällen ein Random-Walk nicht sehr sinnvoll, da der Erwartungswert konstant ist und somit bei der Punktprognose die letzte Realisation konstant weiter geschrieben wird. Deswegen kann der Random-Walk um einen deterministischen Teil erweitert werden, einen Drift.

Die Modellgleichung eines Random-Walk mit Drift sieht folgendermaßen aus (Foster, a, vgl. S. 12-13).

$$y_t = y_{t-1} + \alpha_0 + \epsilon_t \quad (4.6)$$

$$\Delta y_t = \alpha_0 + \epsilon_t \quad (4.7)$$

Man erkennt, dass nun die Zeitreihe zu jedem Zeitpunkt noch um den Wert von α_0 steigt. Wenn α_0 kleiner als null ist, handelt es sich um einen fallenden Trend, wenn α_0 größer null ist, um einen steigenden. Der Erwartungswert ist nun nicht mehr konstant über den Zeitverlauf. Er ist vom Parameter α_0 und der Zeit abhängig ($E(y_t) = t\alpha_0 + y_0$). Somit ist der Erwartungswert beim Random-Walk mit Drift nicht endlich.

Aus der Modellgleichung kann man die Prognose eines Random-Walk mit Drift, in Abhängigkeit zu den vergangenen Werten, berechnen. Man erhält folgende Gleichungen:

$$\hat{y}_{t+h}|y_t, \dots, y_0 = y_t + \alpha_0 h + \sum_{i=t+1}^{t+h} \epsilon_i \quad (4.8)$$

$$E[\hat{y}_{t+h}|y_t, \dots, y_0] = y_t + \alpha_0 h \quad (4.9)$$

$$Var(\hat{y}_{t+h}|y_t, \dots, y_0) = h\widehat{\sigma^2} \quad (4.10)$$

4. Prognose

Die Zuwachsrate α_0 fließt nun auch in das Modell mit ein. Der Term ϵ_t hat keinen Einfluss auf die Punktschätzung, da der Erwartungswert von diesem Null ist. Er beeinflusst nur die Prognoseintervalle, die mithilfe der geschätzten Varianz ($\widehat{\sigma_{t+h}^2}$), für jeden Prognosezeitpunkt, bestimmt werden. Die geschätzte Varianz nimmt für jeden neu prognostizierten Zeitpunkt, um den selben Wert zu. Dieser Wert stellt die geschätzte Varianz des Modells dar ($\widehat{\sigma^2}$) und setzt sich aus der Varianz des Fehlerterms ($Var[\epsilon_t]$) und der Varianz des Parameters ($Var[\alpha_0]$) zusammen.

Beim Random-Walk mit Drift wird somit ein steigender oder fallender Trend prognostiziert und die Breite des Prognoseintervalls nimmt linear zu.

4.1.3. ARIMA-Modell

Das ARIMA-Modell setzt sich aus einem autoregressiven Prozess (AR) und einen Moving-Average Prozess (MA) zusammen.

Das I steht für “Integrated” und bedeutet, dass die Zeitreihe zuerst durch Differenzieren in einen schwach stationären Prozess überführt werden muss, da dies eine Voraussetzung für ein ARMA-Modell ist. Dazu betrachtet man die Definition von schwacher Stationarität (Hyndman and Athanasopoulos, 2012, vgl. 8.1).

Unter schwacher Stationarität versteht man, dass der Erwartungswert über den Zeitverlauf konstant ist (a), die Varianz endlich ist (b) und die Kovarianz nur von ihrem Zeitabstand h abhängig ist (c).

$$a) E[y_t] = E[y_{t+1}] = \dots = E[y_{t+h}] \text{ für } h \in \{0, 1, \dots\} \quad (4.11)$$

$$b) Var[y_{t+h}] < \infty \text{ für } h \in \{0, 1, \dots\} \quad (4.12)$$

$$c) cov[y_0, y_{0+h}] = cov[y_1, y_{1+h}] = \dots = cov[y_t, y_{t+h}] \quad (4.13)$$

Aus (a) erkennt man, dass die Zeitreihe keinen Trend aufweisen darf, da sonst der Erwartungswert über dem Zeitverlauf nicht konstant wäre.

Schwache Stationarität erreicht man durch Differenzieren der Zeitreihe. Für Zeitreihen mit linearem Trend, muss man die Werte einmal differenzieren (Δy_t), für die mit quadratischem Trend, zweimal ($\Delta^2 y_t$).

$$\Delta y_t = y_t - y_{t-1} \quad (4.14)$$

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1} \quad (4.15)$$

$$= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \quad (4.16)$$

Ob die, nach Differenzieren erhaltene Zeitreihe wirklich stationär ist, kann mit einem adjusted Dickey-Fuller Test oder einem KPSS Test überprüft werden. Sollte immer noch keine Stationarität vorliegen, muss eine erneute Differenziation vorgenommen werden.

Nachdem man die Zeitreihe in einen schwach stationären Prozess überführt hat, kann man nun die Modellgleichungen eines AR(p)- und eines MA(q)-Prozesses betrachten (Hyndman and Athanasopoulos, 2012, vgl. 8.3-8.5).

$$AR(p): y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \dots + \alpha_p y_{t-p} + \epsilon_t \quad (4.17)$$

$$MA(q): y_t = \epsilon_t + \beta_1 \epsilon_{t-1} + \beta_2 \epsilon_{t-2} + \dots + \beta_q \epsilon_{t-q} \quad (4.18)$$

Aus der Formel erkennt man, dass im autoregressiven Modell die Realisation y_t , von ihren vorherigen Werten jeweils mit einem bestimmten Anteil (α) abhängig ist. Für den

4. Prognose

Fehlerterm ϵ_t wird gaußsches weißes Rauschen angenommen. So weist dieser einen Erwartungswert von Null auf und besitzt die Varianz von σ^2 . Die einzelnen ϵ_t sind über den Zeitverlauf unabhängig voneinander und identisch verteilt.

Im Moving-Average Modell ist die Realisation von den vergangenen Fehlertermen abhängig. So kann diese als gewichteter gleitender Durchschnitt der letzten q -Fehlerterme angesehen werden. Für die Fehlerterme gilt, wie oben erwähnt, gaußsches weißes Rauschen und sie sind unabhängig identisch voneinander.

Fasst man nun die beiden Gleichungen zu einem ARMA-Modell zusammen, so erhält man.

$$y_t = \alpha_1 y_{t-1} + \dots + \alpha_p y_{t-p} + \beta_1 \epsilon_{t-1} + \dots + \beta_q \epsilon_{t-q} + \epsilon_t \quad (4.19)$$

Die Realisation zum Zeitpunkt t (y_t) ist nun von ihren vergangenen Realisationen und Fehlertermen abhängig.

Kurz schreibt man ARIMA(p,d,q), wobei p die Anzahl der autoregressiven Lags darstellt, d die Ordnung der Differenzierung und q die Anzahl der Lags beim MA-Prozess.

Die Werte für p , d und q können in R über die Funktion `auto.arima` erhalten werden (Hyndman and Athanasopoulos, 2012, vgl. 8.7). Sie ist im package "forecast" enthalten. Die Parameter werden mithilfe des AIC (Akaike Information Criterion) ausgewählt. So bestimmt die Funktion in R zuerst den Parameter d , anhand des adjusted Dickey-Fuller oder KPSS Tests (d ist maximal 2) und anschließend die Parameter p und q . Dazu wird für mehrere verschiedene Parametereinstellungen für p und q , die anhand eines Algorithmus vorgenommen werden, eine Maximum-Likelihood Schätzung (ML-Schätzung) durchgeführt und die Parameterwerte mit den geringsten AIC verwendet. Durch die ML-Schätzung erhält man auch die Werte für die Parameter α_p und β_q .

Nachdem man die Parameterwerte erhalten hat, kann man nun eine Prognose für das ARIMA-Modell erstellen. Für die Punktschätzung sind die zukünftigen ϵ_t nicht von Bedeutung, da deren Erwartungswert null ist. Die vergangenen Fehlerterme fließen in die Prognose mit ein. Damit erhält man folgende Modellgleichung für die Prognose des nächsten und übernächsten Wertes eines ARIMA(2,0,2)-Modells: (Hyndman and Athanasopoulos, 2012, vgl. 8.8)

$$E[\widehat{y_{t+1}}|y_t, \dots, y_0] = \alpha_1 y_t + \alpha_2 y_{t-1} + \beta_1 \widehat{\epsilon}_t + \beta_2 \widehat{\epsilon}_{t-1} \quad (4.20)$$

$$E[\widehat{y_{t+2}}|y_t, \dots, y_0] = \alpha_1 E[\widehat{y_{t+1}}|y_t, \dots, y_0] + \alpha_2 y_t + \beta_2 \widehat{\epsilon}_t \quad (4.21)$$

$$\text{mit } E[\widehat{\epsilon}_{t+1}] = E[\widehat{\epsilon}_{t+2}] = 0 \quad (4.22)$$

Man erkennt, dass für die Punktschätzung des übernächsten Wertes der vorherige geschätzte Wert vonnöten ist. Außerdem fließen, ab einem bestimmten in der Ferne liegenden Zeitpunkt, die geschätzten ϵ_t des Modells nicht mehr direkt in die Prognose mit ein, da deren Erwartungswerte der Schätzungen Null sind.

Für die Schätzung eines Prognoseintervalls spielen die Schätzungen des Parameters ϵ_t jedoch eine wichtige Rolle. So muss ein Wert für die geschätzte Varianz ermittelt werden ($\widehat{\sigma_{t+h}^2}$). Für das erste Prognoseintervall (erster Zeitpunkt nach letzter Realisation) ist diese einfach die geschätzte Varianz des Fehlerterms ($\widehat{\sigma^2}$). Für von der letzten Realisation weiter entfernt liegende Prognosen, ist die Berechnung der Varianz deutlich komplizierter, da in diesem Fall die Varianzen mehrerer Fehlerterme mit einfließen und die Parameter auch noch eine Rolle spielen. Deswegen wird auf deren Berechnung nicht genauer eingegangen. Allgemein gilt für ein ARIMA-Modell, dass die Varianz, für entfernte Zeitpunkte, gegen einen Wert konvergiert. Voraussetzung hierfür ist, dass die Ausgangszeitreihe Stationarität aufweist, was bei einem ARIMA-Modell gegeben sein sollte.

4. Prognose

Eine Erweiterung des ARIMA-Modells stellt das saisonale ARIMA-Modell dar (Hyndman and Athanasopoulos, 2012, vgl. 8.9). Bei diesem kann eine Differenzierung der Saison vorgenommen werden. Außerdem kann auch bestimmt werden, dass die Werte vom letzten Saisonwert oder vom Fehlerterm abhängig sind. Das bedeutet bei monatlichen Daten, dass der zwölften vorherige Wert oder Fehlerterm in das Modell mit aufgenommen wird (z.B. $\alpha_{12}y_{t-12}; \beta_{12}\epsilon_{t-12}$).

4.2. Prognosebeispiel

In diesem Abschnitt wird beispielhaft eine Prognose für die Produktion im Produzierenden Gewerbe in Deutschland für die nächsten zwölf Monate erstellt. Die Zeitreihe beginnt mit dem Januar 1991 und endet im September 2008. Für die Zerlegung der Zeitreihe wurden folgende Parameter benutzt: ns=9; nt=13; isdeg=0; itdeg=1; no=10; ni=1. Zuerst betrachtet man die Punktprognose, danach die Prognoseintervalle.

4.2.1. Punktprognose

In die Punktschätzung fließen nur die prognostizierten Werte der Saison- und Trendkomponente mit ein, da für die Restkomponente ein Erwartungswert von Null angenommen wird.

Zuerst behandelt man die Saisonkomponente. Für ihre Prognose stehen zwei Möglichkeiten zur Verfügung, entweder eine Loess-Schätzung oder das Fortschreiben des letzten Wertes der entsprechenden Saisonuntergruppe.

Die Werte für die Loess-Schätzung erhält man bereits über den Output der Zerlegung. So wird im zweiten Schritt der inneren Schleife bereits ein Wert jeder Saisonuntergruppe prognostiziert. Diese werden für die Saisonprognose genommen und mit dem in Schritt drei erhaltenen Wert der Saisonkorrektur subtrahiert. In der folgenden Tabelle (4.1) ist dargestellt, welche Werte man für die Prognose der Saisonkomponente erhalten würde.

Zeitpunkt	Loess-Schätzung	letzter Saisonwert
10.2008	5.8062	5.7807
11.2008	6.8023	6.7900
12.2008	-5.4578	-5.4328
01.2009	-7.5510	-7.5486
02.2009	-5.5623	-5.5759
03.2009	5.0806	5.0567
04.2009	-0.2428	-0.2450
05.2009	-1.6393	-1.6397
06.2009	2.5126	2.5120
07.2009	1.6678	1.6913
08.2009	-6.7817	-6.7730
09.2009	5.3692	5.3631

Tabelle 4.1.: Vergleich und Werte der verschiedenen Prognosemethode der Saisonkomponente für die Produktion im Produzierenden Gewerbe in Deutschland

4. Prognose

Aus der Tabelle 4.1 ist ersichtlich, dass sich die Werte nicht besonders stark unterscheiden. Der maximale Abstand beträgt circa 0.03 und tritt für den Monat März auf. So hat die Loess-Schätzung dort ein höheres Niveau als die konstante Schätzung. Wenn man sich noch einmal die Saisonuntergruppengrafik aus Abbildung 3.2 für den Monat März ins Gedächtnis ruft (Grafik 3), so erkennt man, dass die letzten beiden Werte oberhalb der roten Linie liegen. Deswegen scheint der prognostizierte Wert anhand der Loess-Schätzung näher an der Realität zu liegen, als das konstante Fortschreiben des letzten Saisonwertes. Bei der Loess-Schätzung findet somit eine leichte Anpassung an die letzten Werte statt.

Nun wird auf die Prognose der Trendkomponente eingegangen. Bei dieser muss zuerst ein Modell geschätzt werden. Es stehen zwei Möglichkeiten zur Verfügung: ein ARIMA-Modell oder ein Random-Walk mit Drift.

Zuerst wird ein Random-Walk mit Drift auf die Trendkomponente geschätzt, mit der danach eine Prognose erstellt wird. Man erhält für den Random-Walk mit Drift folgende Modellparameter.

Random-Walk	Drift (c)	Fehlerterm (ϵ)
Wert	0.09887	-
Standardabweichung	0.01892	0.27550

Tabelle 4.2.: Modellparameter eines Random-Walks mit Drift in der Trendkomponente für die Produktion im Produzierenden Gewerbe in Deutschland

Beim Random-Walk nimmt der Trend mit jeder Zeiteinheit um den Wert 0.09887 zu. Das heißt pro Monat steigt die Produktion im Produzierenden Gewerbe in Deutschland um den Preisindex von circa 0.1 an. Die Standardabweichung für den Fehlerterm beträgt 0.27550.

Nun betrachtet man das ARIMA-Modell. Da man bei Betrachtung der Trendkomponenten über den Zeithorizont einen linear steigenden Trend feststellt, würde man sich dafür entscheiden, die Zeitreihe einmal zu differenzieren ($d=1$).

Die Funktion `auto.arima` in R schlägt dann, für die einmal-differenzierte Zeitreihe, als optimale Wahl für die Anzahl der autoregressiven Lags eins ($p=1$) und für die Anzahl der Moving-Average Lags zwei ($q=2$) vor. Man schätzt somit ein ARIMA(1,1,2)-Modell. Dieses besitzt den geringsten AIC und man würde folgende Parameterwerte für das Modell erhalten.

ARIMA	AR1 (α_1)	MA1 (β_1)	MA2 (β_2)	Ft (ϵ)
Wert	0.9478	0.0149	0.6054	-
Standardabw.	0.0213	0.0657	0.0460	0.0580

Tabelle 4.3.: Modellparameter eines ARIMA-Modells in der Trendkomponente für die Produktion im Produzierenden Gewerbe in Deutschland

Man erkennt, dass die letzte Realisation der einmal-differenzierten Zeitreihe einen hohen Einfluss auf die Realisationen und zukünftigen Werte besitzt, da der Parameter α_0 einen Wert sehr nahe eins annimmt. Der letzte Fehlerterm hat einen deutlich geringeren Einfluss, als der vorletzte Fehlerterm. Der Einfluss des letzten Fehlerterms ist nicht signifikant. Um keinen Trend in der Trendkomponente zu haben, wurde die Zeitreihe einmal differenziert.

4. Prognose

Es handelt sich somit um einen linearen Trend. Die Standardabweichung des Fehlerterms beträgt 0.0580 und ist geringer als beim Random-Walk mit Drift.

Nachdem man die beiden Modelle aufgestellt hat, kann man eine Prognose für die Trendkomponente schätzen. Im folgenden werden die berechneten Prognosewerte für die Punkteschätzung des Trendes für die beiden Modelle betrachtet.

Zeitpunkt	Random-Walk mit Drift	ARIMA(1,1,2)
10.2008	107.6652	107.2974
11.2008	107.7641	107.0379
12.2008	107.8630	106.7919
01.2009	107.9618	106.5588
02.2009	108.0607	106.3378
03.2009	108.1596	106.1284
04.2009	108.2585	105.9299
05.2009	108.3573	105.7418
06.2009	108.4562	105.5634
07.2009	108.5551	105.3944
08.2009	108.6540	105.2343
09.2009	108.7528	105.0824

Tabelle 4.4.: Vergleich und Werte der verschiedenen Prognosemethoden der Trendkomponente für die Produktion im Produzierenden Gewerbe in Deutschland

Beim Random-Walk mit Drift ergibt sich ein positiver Trend, der pro Monat linear um den selben Wert steigt, beim ARIMA(1,1,2)-Modell hingegen ist ein negativer Trend feststellbar (vgl. Tab. 4.4). Nach zwölf Monaten ergibt sich ein Unterschied zwischen den beiden Prognosemethoden von circa 3.7 Preisindexpunkten.

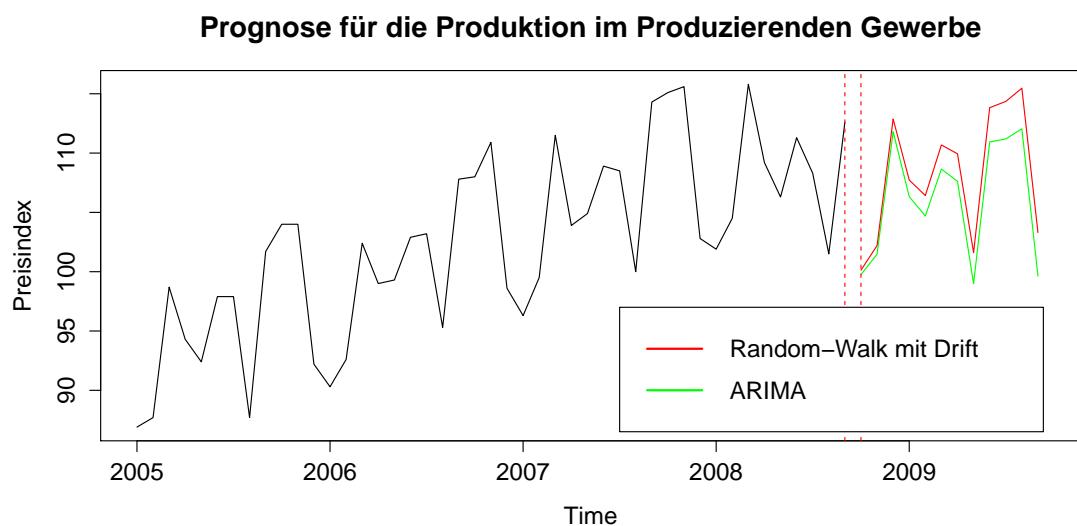


Abbildung 4.1.: Prognose und Vergleich der beiden Modelle für den Trend für die Produktion im Produzierenden Gewerbe in Deutschland

4. Prognose

Um die endgültige Punktschätzung für die Zeitreihe zu erhalten, müssen die Prognose der Saison- und der Trendkomponente addiert werden. In der Abbildung 4.1 wurde für die Schätzung der Saisonkomponenten das Loess-Verfahren benutzt. Man erkennt den unterschiedlichen Verlauf der beiden Prognosen. So weisen die Werte beim Random-Walk mit Drift einen stärkeren Trend auf. Das ARIMA-Modell nimmt die leicht abfallende Kurve im Jahr 2008 zum Anlass einen schwächeren Trend zu schätzen, außerdem ist die saisonale Struktur in der Prognose sichtbar.

4.2.2. Prognoseintervall

Um ein Prognoseintervall erstellen zu können, muss zuerst eine Verteilungsannahme für die zu prognostizierenden Werte angegeben werden. In unserem Fall nimmt man an, dass die zukünftigen Realisationen bedingt auf ihre vergangenen Werte normal-verteilt sind.

$$\hat{y}_{t+h} | y_t, \dots, y_0 \sim N(\mu, \sigma^2) \quad (4.23)$$

Mithilfe dieser Annahme kann man ein Prognoseintervall folgendermaßen berechnen.

$$PI : [E[\hat{y}_{t+h} | y_t, \dots, y_0] \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}_{t+h}^2}] \quad (4.24)$$

$z_{1-\frac{\alpha}{2}}$ stellt das Quantil der Standardnormalverteilung dar. $\hat{\sigma}_{t+h}^2$ die geschätzte Varianz zum Zeitpunkt $t+h$.

In den meisten Fällen wird ein 95%-Prognoseintervall oder ein 90%-Prognoseintervall betrachtet, somit muss man das entsprechende Quantil der Standardnormalverteilung berechnen ($z_{0.975} = 1.96$; $z_{0.95} = 1.64$).

Doch wie erhalte ich die geschätzten Varianzen? Da die Saisonkomponente einem deterministischen Trend folgt und somit dessen Prognose keine Unsicherheit aufweist, hat sie keinen Einfluss auf die Varianz. Man muss also für die Berechnung der Varianz die Trend- und Restkomponente hinzuziehen.

Beim Random-Walk mit Drift, für die Trendkomponente, hat man in Kapitel 4.1.3 gezeigt, dass die Varianz für die Prognose $Var(\hat{y}_{t+h} | y_t, \dots, y_0) = \hat{\sigma}_{t+h}^2 = h\hat{\sigma}^2$ ist. Wobei sich $\hat{\sigma}^2$ aus der Varianz des Fehlerterms plus der Varianz des Parameters für den Drift zusammensetzt. Beim ARIMA-Modell entspricht die Varianz für das erste Prognoseintervall der Varianz des Fehlerterms (ϵ_t). Bei Prognosen, die weiter in die Zukunft reichen, stellt sich die Berechnung der Varianz als komplizierter heraus, deshalb wird auf die genaue Vorgehensweise nicht näher eingegangen. Liegt ein stationäres ARIMA-Modell vor, so konvergiert die Varianz gegen einen bestimmten Wert.

In Tabelle 4.5 vergleicht man die 95%-Prognoseintervalle und die geschätzten Varianzen für die Trendkomponente der beiden verschiedenen Modelle. Man erkennt, dass bis zum Monat März, das geschätzte $\hat{\sigma}^2$ beim Random-Walk mit Drift größer, als beim ARIMA-Modell ist. Mit mittlerer Anzahl von prognostizierten Perioden ist die Unsicherheit beim Random-Walk geringer, als beim ARIMA-Modell. Beim Random-Walk mit Drift steigt die Varianz linear an. Wegen der Stationarität sollte im ARIMA-Modell die Varianz gegen einen bestimmten Wert konvergieren. Für eine kurzfristige Prognose (1-3 Monate) liefert in unserem Beispiel das ARIMA-Modell kleinere Prognoseintervalle, für mittelfristige bis langfristige Prognosen (4-12 Monate), der Random-Walk mit Drift.

4. Prognose

Modell	Random-Walk mit Drift		ARIMA-Modell	
	Zeitpunkt	$\hat{\sigma}^2$	Prognoseintervall	$\hat{\sigma}^2$
10.2008	0.0763	[107.1240;108.2065]	0.0034	[107.1836;107.4111]
11.2008	0.1532	[106.9969;108.5313]	0.0163	[106.7873;107.2885]
12.2008	0.2309	[106.9211;108.8048]	0.0572	[106.3233;107.2605]
01.2009	0.3093	[106.8718;109.0519]	0.1387	[105.8289;107.2887]
02.2009	0.3884	[106.8392;109.2823]	0.2717	[105.3162;107.3594]
03.2009	0.4683	[106.8184;109.5008]	0.4650	[104.7919;107.4649]
04.2009	0.5488	[106.8064;109.7105]	0.7259	[104.2600;107.5998]
05.2009	0.6301	[106.8015;109.9132]	1.0601	[103.7237;107.7598]
06.2009	0.7121	[106.8023;110.1102]	1.4723	[103.1852;107.9417]
07.2009	0.7948	[106.8077;110.3024]	1.9659	[102.6463;108.1426]
08.2009	0.8782	[106.8172;110.4907]	2.5434	[102.1084;108.3601]
09.2009	0.9623	[106.8301;110.6756]	3.2065	[101.5727;109.5922]

Tabelle 4.5.: Prognoseintervalle der beiden Modelle der Trendkomponente für die Produktion im Produzierenden Gewerbe in Deutschland

Für die Restkomponente wird ein weißes gaußsche Rauschen angenommen, mit Erwartungswert Null und Varianz σ^2 . Somit ergibt sich, dass die Varianz zum Zeitpunkt $\hat{\sigma}_{t+h}^2 = \hat{\sigma}^2$ ist. Sie ist somit für jeden Prognosezeitpunkt konstant. In unserem Fall, für die Produktion im Produzierenden Gewerbe in Deutschland, ergibt sich: $\hat{\sigma}^2 = 0.6255$.

Die endgültigen Prognoseintervalle ergeben sich, indem man die Varianzen aus der Trend- und Restkomponente verbindet. Jedoch reicht es nicht, die beiden zu addieren. Man muss auch die Kovarianz zwischen den beiden Parametern betrachten. Diese bekommt man mithilfe der Residuen. Beim Random-Walk mit Drift und beim ARIMA-Modell entsprechen diese dem Fehlerterm. Bei der Restkomponente stellt diese selbst die Residuen dar, da sie ja den Teil, der nicht über die beiden anderen Komponenten abgedeckt werden kann, erklärt.

Modell	Random-Walk mit Drift/Rest	ARIMA-Modell/Rest
Kovarianz	0.00374	-0.00679
Korrelation	0.01740	-0.14738

Tabelle 4.6.: Kovarianzen und Korrelationen zwischen den einzelnen Residuen bei der Produktion im Produzierenden Gewerbe in Deutschland

In der Tabelle 4.6 sieht man, welche Kovarianz sich zwischen der Restkomponente und den Residuen des Trendmodells ergibt, zur besseren Interpretation wird auch die Korrelation betrachtet. Die Residuen des ARIMA-Modells weisen eine geringe negative Korrelation mit der Restkomponente auf, beim Random-Walk ist der Zusammenhang kaum messbar. Für die Schätzung der Kovarianz geht man davon aus, dass die Korrelation konstant bleibt, da diese eine normierte Größe ist. Mithilfe dieses Wissens, kann man die zukünftigen Kovarianzen schätzen. Sie ergeben sich anhand der Korrelation, multipliziert mit den beiden

4. Prognose

Standardabweichungen.

$$\text{allgemein: } \text{cor}_t(T, R) = \frac{\text{cov}_t(T, R)}{\sqrt{\sigma_t^2(T)\sigma_t^2(R)}} \quad (4.25)$$

$$\text{mit: } \text{cor}_t(T, R) = \text{cor}_{t+h}(T, R) \quad (4.26)$$

$$\rightarrow \widehat{\text{cov}_{t+h}(T, R)} = \text{cor}_t(T, R) \sqrt{\widehat{\sigma_{t+h}^2(T)} \widehat{\sigma_{t+h}^2(R)}} \quad (4.27)$$

Nachdem man nun die Kovarianzen berechnet hat, kann man nun das $\widehat{\sigma_t^2}$ berechnen.

$$\widehat{\sigma_{t+h}^2} = \widehat{\sigma_{t+h}^2(T)} + \widehat{\sigma_{t+h}^2(R)} + 2\widehat{\text{cov}_{t+h}(T, R)} \quad (4.28)$$

$$= \widehat{\sigma_{t+h}^2(T)} + \widehat{\sigma_{t+h}^2(R)} + 2\text{cor}_t(T, R) \sqrt{\widehat{\sigma_{t+h}^2(T)} \widehat{\sigma_{t+h}^2(R)}} \quad (4.29)$$

Mithilfe der geschätzten Werte der Varianz, kann man nun die Prognoseintervalle erstellen. Sie ergeben sich, wie oben beschrieben, über die Quantile der Standardnormalverteilung und der geschätzten Varianz.

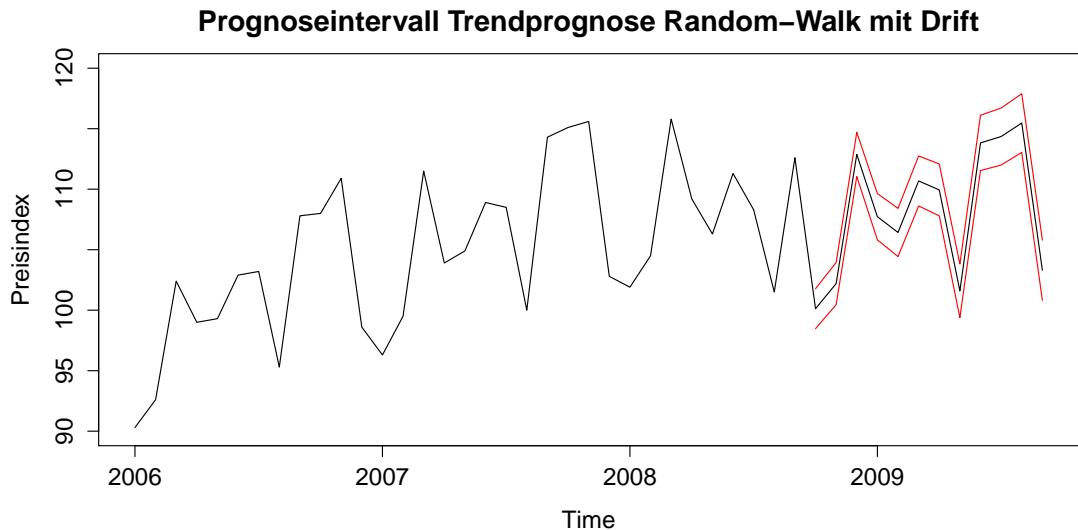


Abbildung 4.2.: Prognoseintervall für die Prognose des Trendes mit Random-Walk mit Drift für die Produktion im Produzierenden Gewerbe in Deutschland

In Abbildung 15 und 16 sind für die beiden Modelle die endgültigen Prognoseintervalle aufgezeigt. Die schwarze Linie im geschätzten Bereich stellt die Punktprognose des jeweiligen Modells dar.

4. Prognose

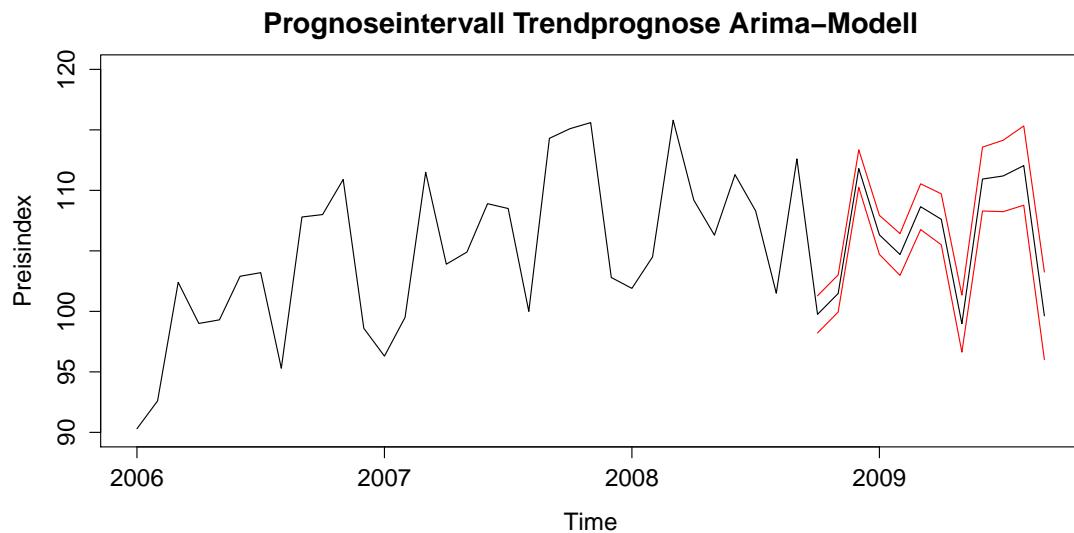


Abbildung 4.3.: Prognoseintervall für die Prognose des Trendes mit Arima-Modell für die Produktion im Produzierenden Gewerbe in Deutschland

4.3. Prognosegüte

Nachdem man die Prognosemöglichkeiten für eine in Saisonkomponente und Trendkomponente zerlegte Zeitreihe aufgezeigt hat, will man feststellen, wie gut diese Methode gegenüber anderen Modellen und Prognosemöglichkeiten ist. Dazu wird die Zeitreihe zu einem früheren Zeitpunkt abgeschnitten, darauf eine Prognose erstellt und mit den danach eingetroffenen Werten verglichen. Der Vergleich, der verschiedenen Prognosemethoden geschieht mithilfe des mittleren quadratischen Fehlers.

Für die Anzahl der Werte, die prognostiziert werden, werden zwei verschiedene Varianten betrachtet, einmal für eine kurzfristige Prognose drei Monate und für eine langfristige Prognose zwölf Monate.

Da der Vergleich, wenn nur zu einem Zeitpunkt abgeschnitten worden ist, keine sinnvolle Lösung darstellt, bestehen zwei Möglichkeiten, entweder ein rekursives Prognosefenster oder eine rollierendes.

Betrachtet man die beiden Möglichkeiten für die Produktion im Produzierenden Gewerbe in Deutschland, so wird beim rekursiven Prognosefenster die Zeitreihe Ende 2000 abgeschnitten und darauf die nächsten zwölf bzw. drei Monate prognostiziert und der mittlere quadratische Fehler berechnet. Daraufhin wird der Wert des Januar 2001 in die Zeitreihe mit aufgenommen und erneut die nächsten zwölf bzw. drei Monate prognostiziert usw.. Beim rollierenden Prognosefenster besteht für den Input immer das gleiche Zeitintervall. So wird bei diesem auch im ersten Schritt Ende 2000 abgeschnitten und daraufhin eine Prognose erstellt und der mittlere quadratische Fehler berechnet. Daraufhin wird erneut der Januar mit in die Ausgangszeitreihe aufgenommen, wobei in diesem Fall auch der erste Wert der Ausgangszeitreihe wegfällt. Diese beginnt nun im Februar 1991. Damit hat man immer die gleiche Anzahl an Werten für die Input-Zeitreihe.

Nachdem im folgenden Teilkapitel die Berechnung des mittleren quadratischen Fehlers aufgezeigt wird, werden im darauffolgenden Teilkapitel folgende Modelle verglichen:

4. Prognose

- Saisonaler ARIMA, Random-Walk mit Drift
- neue Methode: STL+Random-Walk mit Drift und STL+ARIMA
- alte Methode: STL+Random-Walk mit Drift und STL+ARIMA.

Unter der alten Methode ist zu verstehen, dass für die Saison die letzten Saisonwerte konstant weitergeschrieben werden und die Trend- und Restkomponente zusammengefügt werden und darauf ein Modell und eine Prognose geschätzt wird.

4.3.1. Mittlerer quadratischer Fehler

Der mittlere quadratische Fehler stellt ein Maß dar, um die Prognosegüte eines Modells zu beurteilen. Bei diesem werden die Abweichungen des Schätzwertes vom wahren Parameter aufsummiert und durch die Anzahl der vorhergesagten Perioden geteilt.

$$MQF = \frac{1}{h} \sum_{i=t+1}^{t+h} (E[\hat{y}_i] - y_i)^2 \quad (4.30)$$

Um den MQF bestimmen zu können müssen die wahren Werte der Realisationen y_{t+h} bekannt sein. Somit kann der MQF immer erst im Nachhinein berechnet werden. Aus der Formel erkennt man, dass mit zunehmenden Abweichungen der Wert des MQF immer größer wird. Daraus folgt, je kleiner der MQF, desto besser ist die Prognosegüte des Modells.

4.3.2. Vergleich verschiedener Prognosemethoden

Im folgenden Kapitel werden die verschiedenen Prognosemodelle anhand des MQF (mittleren quadratischen Fehler) verglichen. Als Vergleichszeitreihe dient die Produktion im Produzierenden Gewerbe in Deutschland.

Im ersten Fall betrachtet man eine kurzfristige Prognose für das rekursive Prognosefenster. Im ersten Schritt geht die Ausgangszeitreihe bis Januar 2001, im letzten Schritt bis Mai 2008. Somit werden für jedes Modell insgesamt 91 mittlere quadratische Fehler berechnet. Aus diesen 91 mittleren quadratischen Fehlern wird noch einmal der Mittelwert, der Median und das 75%-Quantil gebildet um die Prognosegüte der Modelle vergleichen zu können. Es können ebenso die Boxplots betrachtet werden, um die verschiedenen Prognosemethoden zu vergleichen.

Aus der Tabelle 4.7 erkennt man, dass bei einer kurzfristigen Prognose (drei Monate) die alten STL-Prognosemethoden die besten Ergebnisse liefert, wenn man den Mittelwert betrachtet. Die einzige Methode, die die saisonale Struktur nicht beachtet, der Random-Walk mit Drift, liefert mit Abstand die schlechtesten Ergebnisse. Beim 75%-Quartil hingegen schneiden die neuen STL-Prognosemethoden besser als die alten ab. Bei der Prognose der Komponenten zerlegten Zeitreihe, ist die ARIMA Prognose der Random-Walk mit Drift Prognose vorzuziehen. Das saisonale ARIMA-Modell ergibt bei der Überprüfung nie die niedrigsten Werte für den MQF.

Ein Analyst, bei dem die mittlere Abweichung möglichst gering sein sollte, entscheidet sich über den Mittelwert und wählt die alte STL-Prognosemethode mit ARIMA-Modell. Einer, bei dem die Mehrheit der vergangenen MQF möglichst gering sein soll geht nach dem Median oder 75%-Quartil und entscheidet sich somit wahrscheinlich für die neue STL-Prognosemethode mit Random-Walk mit Drift.

Als nächsten betrachtet man eine langfristige Prognose (12 Monate). In diesen Fall sollte

4. Prognose

Modell	saisonaler ARIMA	Random-Walk m. D.	neue Methode STL+ARIMA	neue Methode STL+R-W m. D.
Mittelwert MQF	2.2424	61.2827	2.3307	2.1504
Median MQF	1.5154	40.1041	1.4636	1.3670
75%-Quartil	2.5606	81.4530	2.7477	2.5260
Modell	alte Methode STL+ARIMA		alte Methode STL+R-W mit Drift	
Mittelwert MQF	2.0182		2.1926	
Median MQF	1.2458		1.3053	
75%-Quartil	2.5598		2.8394	

Tabelle 4.7.: Vergleich der verschiedenen Prognosemethoden für die Produktion im Produzierenden Gewerbe in Deutschland (kurzfristige Prognose; rekursives Prognosefenster)

der Wert des mittleren quadratischen Fehlers größer sein, da mit zunehmendem Vorhersage Fenster die Unsicherheit steigt, wie man an den im vorherigen Kapitel gezeigten Prognoseintervallen erkennen kann. In diesem Fall hat man 82 mittlere quadratische Fehler und man vergleicht die Boxplots der verschiedenen Prognosemethoden. Der Random-Walk mit Drift und die alte STL-Methode mit Random-Walk mit Drift werden nicht betrachtet, da diese die schlechtesten Ergebnisse besitzen.

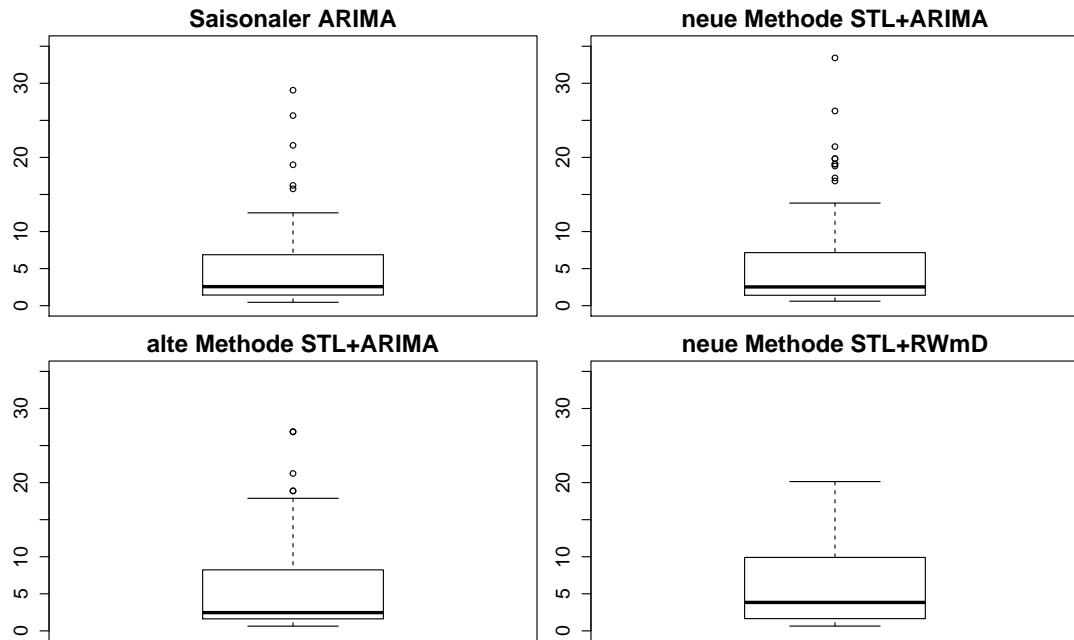


Abbildung 4.4.: Vergleich der verschiedenen Prognosemethoden für die Produktion im Produzierenden Gewerbe in Deutschland (langfristige Prognose; rekursives Prognosefenster)

Bei Betrachtung der Abbildung 4.4 für die langfristigen Prognose, liefert der saisonale ARIMA die besten Ergebnisse. Wenn man nur den Mittelwert betrachtet, ist die alte STL

4. Prognose

Methode besser, als die neue. Bei Median und Quartil hingegen, liefert die neue bessere Ergebnisse ab. Somit weist diese wenige starke Ausreißer beim MQF auf. Bei der zerlegten Zeitreihe liefert das ARIMA-Modell bessere Ergebnisse, als der Random-Walk mit Drift. Die neue STL-Prognosemethode mit Random-Walk mit Drift weist als einzige keine hohen Ausreißer aus. Es kommt somit seltener vor, dass die Prognose vollkommen daneben liegt.

Für das durch rekursives Prognosefenster festgestellte beste Modell, muss zuerst entschieden werden, ob man eine kurzfristige oder langfristige Prognose benötigt. Bei der kurzfristigen Prognose würde man sich für die alte STL-Prognosemethode mit dem ARIMA-Modell entscheiden, bei der langfristigen für einen saisonalen ARIMA.

Nachdem man die Prognose ausgehend vom rekursiven Fenster, betrachtet hat, geht man nun auf ein rollierendes Fenster ein. Zuerst werden erneut die mittleren quadratischen Fehler einer kurzfristigen Prognose betrachtet.

Modell	saisonaler ARIMA	Random-Walk m. D.	neue Methode STL+ARIMA	neue Methode STL+R-W m. D.
Mittelwert MQF	2.2535	61.3659	2.4331	2.2711
Median MQF	1.6030	40.3966	1.2684	1.2844
75%-Quartil	2.8436	80.3580	2.4506	2.6643
Modell		alte Methode STL+ARIMA	alte Methode STL+R-W mit Drift	
Mittelwert MQF		1.9342	2.0737	
Median MQF		1.4706	1.3852	
75%-Quartil		2.4500	2.6393	

Tabelle 4.8.: Vergleich der verschiedenen Prognosemethoden für die Produktion im produzierenden Gewerbe in Deutschland(kurzfristige Prognose, rollierendes Prognosefenster)

Die kurzfristige Prognose eines rollierenden Prognosefenster liefert ungefähr genau so gute Ergebnisse, wie bei einem rekursiven Prognosefenster (vgl. Tab. 4.8). Betrachtet man den Mittelwert, so liefert die alte STL-Prognosemethode die besten Ergebnisse. Beim Median hingegen hat die neue Prognosemethode bessere Ergebnisse. Der saisonale ARIMA ist, außer beim Mittelwert, am schlechtesten. Die STL-Methode mit ARIMA liefert bessere Werte, als die mit Random-Walk mit Drift, außer beim Mittelwert für die neue Methode. Beim 75%-Quartil sind beide Methoden fast identisch.

Nun schaut man sich die Ergebnisse des mittleren quadratischen Fehlers für eine langfristige Prognose (12 Monate) und ein rollierendes Prognosefenster an. Man betrachtet erneut die Boxplots, diesmal ausschließlich für die neue und alte STL-Prognosemethode.

Bei der langfristigen Prognose liefert das rollierende Prognosefenster deutlich bessere Ergebnisse ab, als beim rekursiven (vgl. Abb. 4.4/4.5). Eine Ausnahme bildet das saisonale ARIMA-Modell, es schneidet am schlechtesten ab und wird deswegen nicht in den Boxplots betrachtet. Wenn man den Mittelwert vergleicht, liefert die alte STL-Prognosemethode die besten Ergebnisse, beim Median die neue Prognosemethode. Die STL-Methode mit den Random-Walk mit Drift hat einen geringeren Wert des MQF, als die ARIMA-Methoden.

Nachdem man die vier unterschiedlichen Prognosevorgehensweisen erläutert hat und ihre mittleren quadratischen Fehler berechnet hat, würde man sich für die kurzfristige Pro-

4. Prognose

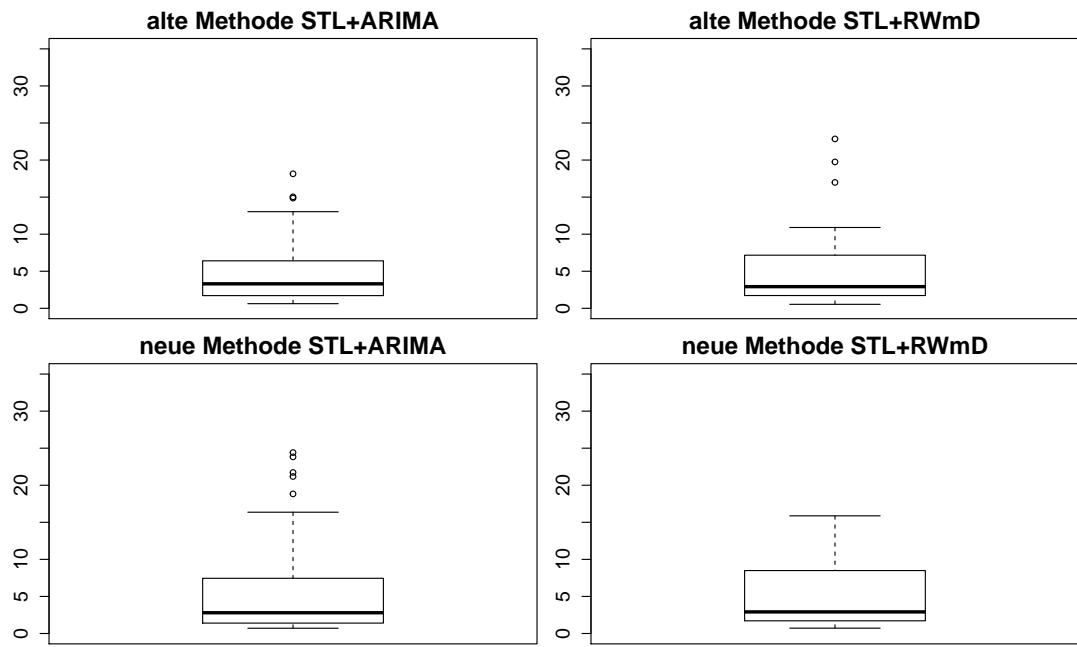


Abbildung 4.5.: Vergleich der verschiedenen Prognosemethoden für die Produktion im Produzierenden Gewerbe in Deutschland (langfristige Prognose; rollierendes Prognosefenster)

gnose für die alte STL-Prognosemethode mit dem ARIMA-Modell entscheiden. Für das Prognosefenster würde man sich, je nach Wunsch des Analysten sich für ein rekursives oder rollierendes entscheiden. Der Unterschied zwischen beiden ist eher gering und beim rekursiven ist der Mittelwert geringer, beim rollierenden der Median.

Bei der langfristigen Prognose würde man auf jeden Fall ein rollierendes Prognosefenster vorziehen und die alte STL-Prognosemethode mit einem Random-Walk mit Drift als Modell für die Trendkomponente.

5. empirisches Beispiel

Arbeitslosigkeit in Deutschland

Nachdem man die Zerlegungs- und Prognosemethoden näher erläutert hat, will man anhand eines empirischen Beispiels, die einzelnen Schritte, um eine passende Saison-Trendzerlegung und damit eine gute Prognose zu bekommen, erneut zeigen.

Dazu betrachtet man die Entwicklung der Arbeitslosigkeit in Deutschland von Januar 1991 bis Oktober 2013 (Statistik der Bundesagentur für Arbeit, 2013). Zuerst wird auf die Zerlegung der Zeitreihe eingegangen.

Bevor die Daten für die Zerlegung verwendet werden können, muss erst eine Bereinigung dieser vorgenommen werden, da sich aufgrund neuer Berechnungsgrundlagen der Arbeitslosenstatistik, zu bestimmten Zeitpunkten, Sprünge in der Zeitreihe ergeben haben (Statistik der Bundesagentur für Arbeit, 2009, vgl. Kap. 6). So gelten seit Januar 2004 Personen, die an einer Weiterbildung der Bundesagentur für Arbeit teilnehmen, nicht mehr als arbeitslos. Dies führte zu einer Verringerung der Arbeitslosenzahlen um circa 100.000 Personen. Eine weitere Änderung tritt seit Januar 2005 auf. Zu diesem Zeitpunkt wurde die Arbeitslosenhilfe mit der Sozialhilfe zusammengelegt. Dies hatte einen Zuwachs von 380.000 Personen in der Arbeitslosenstatistik zur Folge. Diese Änderungen wurden absolut auf die alten Monatswerte hinzugerechnet, um Sprünge in der Input-Zeitreihe zu verhindern.

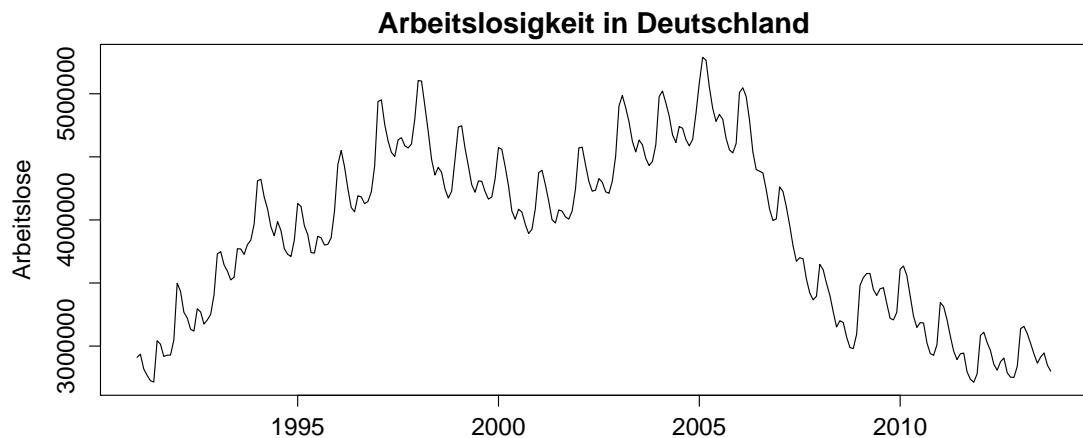


Abbildung 5.1.: Arbeitslosigkeit in Deutschland von 1991 bis Oktober 2013

In Abbildung 5.1 erkennt man, dass die Arbeitslosigkeit bis 1998 angestiegen ist. Nachdem sie danach leicht gesunken ist, kam es wieder zu einem Anstieg, der im Jahre 2005 seinen Höhepunkt erreicht hat. Danach kam es zu einem rasanten Verfall, der im Jahre 2008 geendet hat. Seitdem liegt die Arbeitslosigkeit in Deutschland konstant bei circa drei Millionen Personen.

Für die Zerlegung muss man erneut die Parameter bestimmen (ns, nt). Da bei Betrach-

5. empirisches Beispiel

tung der Arbeitslosigkeit normalerweise keine außergewöhnlichen Ausreißer auftreten, da es bei dieser äußerst selten externe Effekte gibt, die zu außergewöhnlich hohen oder negativen Zuwächsen führen, hat man sich dazu entschieden keine Robustheitsschätzung vorzunehmen. Außerdem hat sich gezeigt, dass ein Saisonparameter von 13 ($ns=13$), die plausibelsten Ergebnisse liefert. Mithilfe dieser Parameter würde mir die Saison-Trend Zerlegungsmethode, anhand des Box-Pierce Tests, zu einem Trendparameter von sieben raten ($nt=7$), da erst ab diesem Unkorreliertheit in der Restkomponente vorliegt. Eine erste Zerlegung wurde somit mit den Parameter $ns=13$ und $nt=7$ durchgeführt.

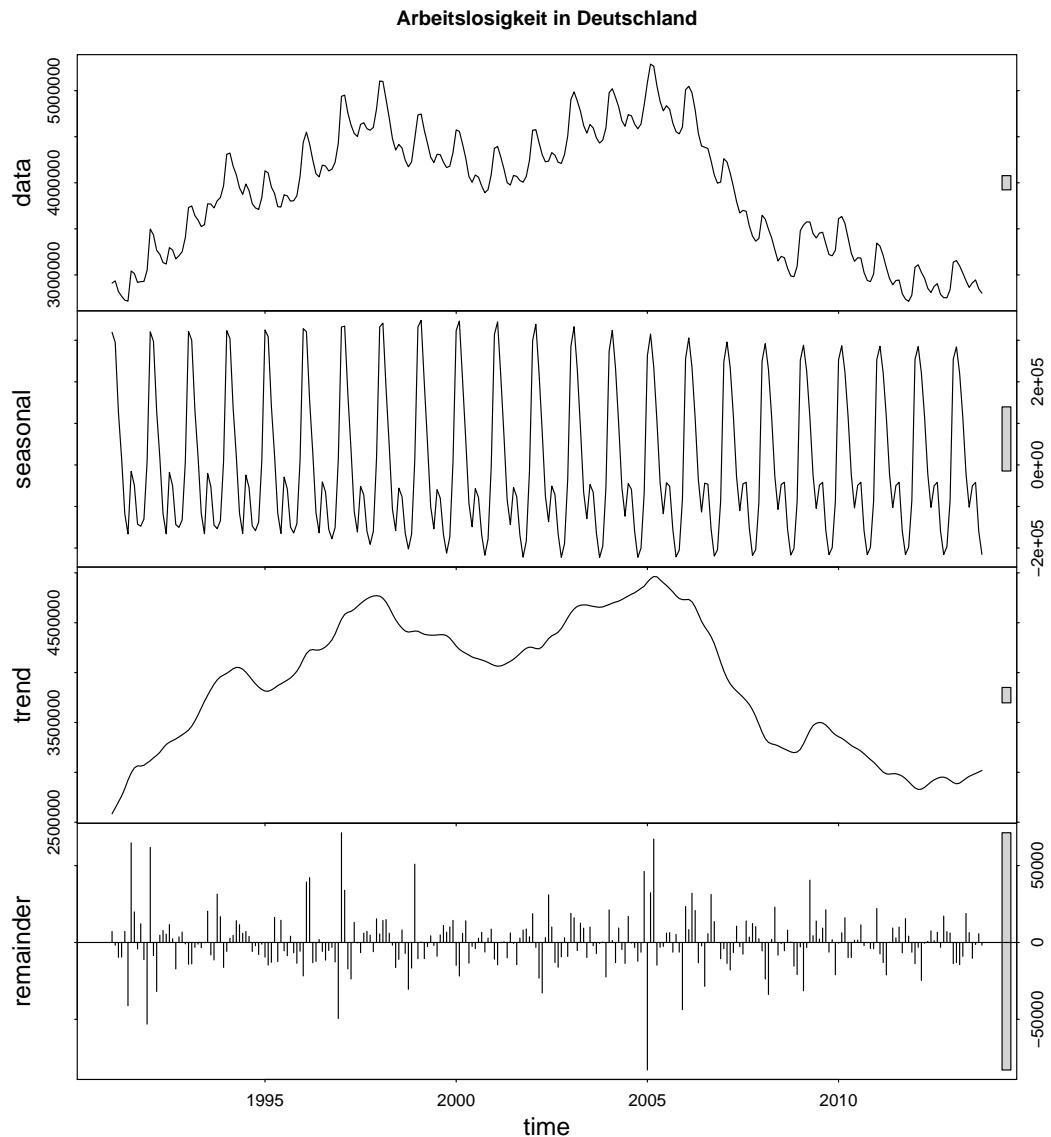


Abbildung 5.2.: Zerlegung der Arbeitslosigkeit in Deutschland mit $ns=13$ und $nt=7$

In Abbildung 5.2 erkennt man im Verlauf der Trendkomponente, dass diese noch sehr viel Schwankungen aufweist. So ist mit einem Trendparameter von sieben die Unkorreliertheit der Restkomponente gewährleistet, jedoch weist die Trendkomponente keine glatte Linie auf. Wie in Kapitel 3.2.3 gesagt, sollte die Trendspannweite auf jeden Fall größer als der

5. empirisches Beispiel

Saisonparameter np sein. Somit ist in diesem Fall davon abzuraten einen Trendparameter von sieben zu nehmen. Es ist hingegen von Vorteil, diesen größer zu machen, um für den Analysten ein interessantes und gewünschtes Resultat der Trendkomponente zu bekommen.

Ausgehend von der Trendspannweite von sieben wurde diese sukzessive erhöht, daraufhin die Trenddiagnosegrafik betrachtet und auf ein zufriedenstellendes Ergebnis überprüft. Letztendlich würde man sich dafür entscheiden eine Spannweite von 15 für die Trendkomponente zu benutzen ($nt=15$). Da in diesem Fall keine Autokorrelation der Restkomponente mehr vorhanden ist, sollte für die Prognose auf die alte STL-Prognose-methode zurückgegriffen werden, da in dieser keine Unkorriertheit der Restkomponente vonnöten ist. Bevor man auf die Prognose eingeht betrachtet man das Ergebnis der zerlegten Zeitreihe mit einer Trendspannweite von 15 und Saisonspannweite von 13.

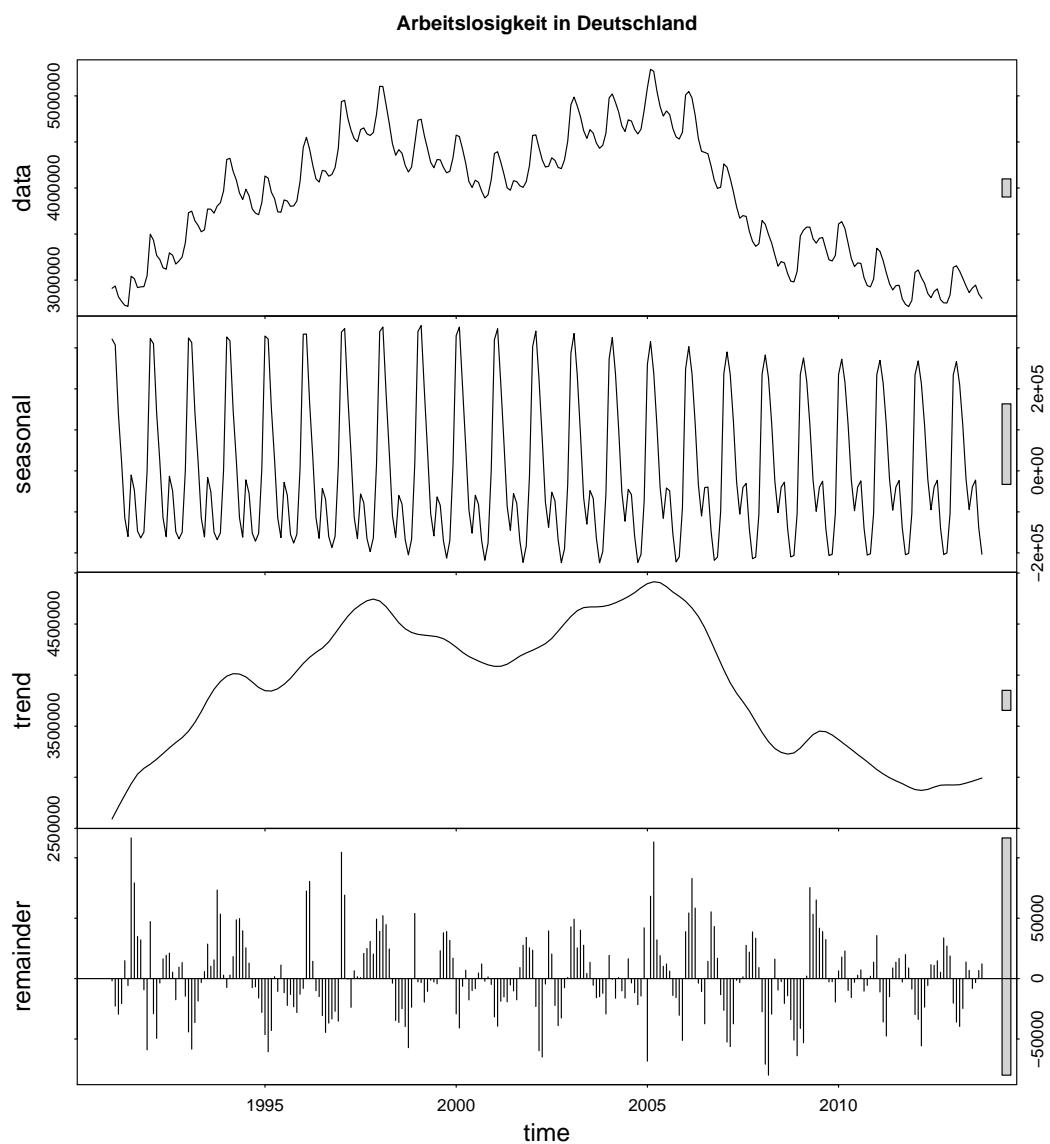


Abbildung 5.3.: Zerlegung der Arbeitslosigkeit in Deutschland mit $nt=15$ und $ns=13$

5. empirisches Beispiel

In Abbildung 5.3 erkennt man nun, dass der Verlauf der Trendkomponente deutlich weniger schwankend ist, er weist eine glattere Linie auf. Es ist jedoch offensichtlich, dass nun in der Restkomponente Struktur vorhanden ist. So treten in dieser vermehrt Zeitpunkte mit positiver oder negativer Abweichung auf, die zeitlich hintereinander liegen. Es scheint, als ob die Restkomponente einen sinus-förmigen Verlauf aufweist. Trotzdem wird die folgende Prognose mithilfe der letzten Saison-Trend Zerlegung vorgenommen.

Bei der zerlegten Zeitreihe weist die Saisonkomponente eine konstante Struktur auf. Der stärkste Monat liegt um circa 300.000 Personen über dem zu diesem Zeitpunkt angenommenen Trend, der schwächste Monat liegt um circa 200.000 Personen darunter. Der Trend weist den unter Abbildung 5.1 beschriebenen Verlauf auf. Die Restkomponente besitzt eine maximale absolute Abweichung von circa 100.000 Personen, was einer prozentualen Abweichung von circa 3% entspricht.

Zur Prognose der Entwicklung der Arbeitslosigkeit in Deutschland nimmt man die Vorhersage der nächsten zwölf Monate vor. Man erhält damit Prognosewerte von November 2013 bis Oktober 2014.

Wie vorher bereits erwähnt, wird die Prognose mithilfe der mit Trendspannweite 15 zerlegten Zeitreihe erstellt. Bei dieser weist die Restkomponente keine Unkorriertheit auf, deswegen wird auf die alte STL-Prognosemethode zurückgegriffen. Außerdem schätzt man auf den Zusammenschluss der Trend- und Restkomponente ein ARIMA-Modell.

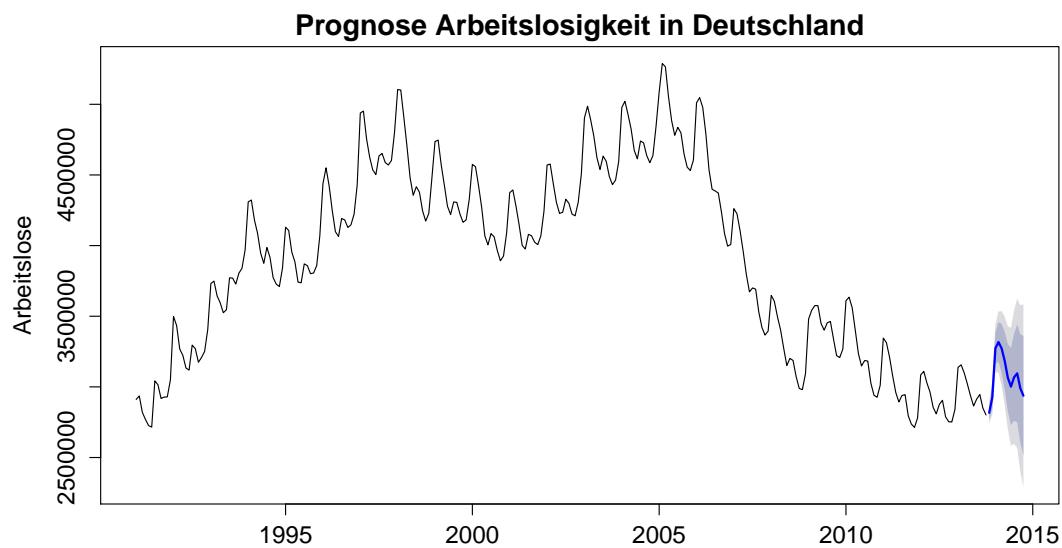


Abbildung 5.4.: Prognosewerte und -intervalle der Arbeitslosigkeit in Deutschland

Die Prognose gibt einen leicht positiven Trend wieder. So steigt die Arbeitslosigkeit in den ersten acht Monaten in 2014 über 3 Millionen. Aufgrund des Saisoneffekts liegt sie im September und Oktober unter 3 Millionen, aber absolut circa 100.000 Personen über den Vorjahreswerten.

6. Schluss

6.1. Zusammenfassung

Beim STL-Zerlegungsverfahren mithilfe von Loess handelt es sich um ein additives Zerlegungsverfahren und es beruht auf einen iterativen Algorithmus. Durch die Loessschätzung erhalten Werte, die nahe des zu schätzenden Zeitpunktes liegen, höheren Einfluss auf die Schätzung, als weiter entfernt liegende Punkte. Mithilfe der Robustheitsschätzung werden die Werte, die nur schlecht über die beiden Komponenten erklärt werden können, bei der erneuten Komponentenschätzung nicht mehr berücksichtigt. Beim STL-Verfahren besteht nicht die Möglichkeit eine Konjunkturkomponente zu erhalten.

Die erzielten Prognoseergebnisse, die auf den durch das STL-Verfahren gewonnenen Komponenten aufbauen, liefern bessere Ergebnisse ab, als wenn eine Prognose direkt auf die Ausgangszeitreihe geschätzt wird. Somit kann mithilfe des STL-Zerlegungsverfahrens eine bessere Prognose der zukünftigen Werte erreicht werden.

6.2. Ausblick

Im Ausblick sollen mögliche Erweiterungen des STL-Zerlegungsverfahren vorgestellt werden.

Die Saison- Trend Zerlegung hat oftmals Schwierigkeiten, wenn ein Strukturbruch in den Daten vorliegt. Mithilfe der Restkomponente könnte versucht werden, ein Kriterium zu ermitteln, um solche Strukturbrüche zu erkennen. So treten vor den Strukturbruch mehrere starke positive bzw. negative Abweichungen auf und nach diesem mehrere starke negative bzw. positive Abweichungen auf. Mithilfe eines Maßes könnte nun versucht werden die Strukturbrüche zu erkennen, zu bereinigen und erneute eine Komponentenzerlegung durchzuführen.

Eine weitere Verbesserung könnte man bei der Robustheitsschätzung vornehmen. So könnte man diese dadurch bestimmen, wie stark die einzelnen Werte die jeweiligen Regressionsschätzungen beeinflussen. Ein Kriterium, was dies messen würde, stellt die Cook's Distance dar. Sie gibt an, wie stark die Regressionsschätzung von einzelnen Werten beeinflusst wird. Der Vorteil von dieser Robustheitsschätzung könnte darin liegen, dass vor allem lokale Maximas und Minimas besser abgebildet werden, wobei dies von mir noch nicht abschließend geprüft worden ist, ob dies so sei.

Auch bei der Prognose könnten noch Verbesserungen vorgenommen werden. So könnte bereits bei der Zerlegung den Analysten die beste Prognosemethode anhand des mittleren quadratischen Fehlers und vorangegangenen Prognosen vorgeschlagen werden.

Wie bereits in einen vorherigen Kapitel erwähnt, treten in der Restkomponente oftmals Zeiträume ein mit vermehrt starken Abweichungen und Zeiträume mit vermehrt schwachen Abweichungen. Dies ist ein Hinweis darauf, dass in der Restkomponente kein weißes Rauschen vorliegt und die Varianz nicht konstant über den Zeitverlauf ist. Eine Vorhersa-

6. Schluss

ge der Varianz würde somit genauerer Prognoseintervalle liefern. Modelle, die die Varianz vorhersagen nennt man GARCH-Modelle.

A. Beiliegende CD-Rom

Die beiliegende CD-ROM enthält die benutzten Datentabellen und den gesamten verwendeten R-Code für die vorliegende Arbeit (Stand: 16. Dezember 2013).

- Produktion im Produzierenden Gewerbe in Deutschland von 1991-2013, kalenderbereinigt und in konstanten Preisen
- Arbeitslosigkeit in Deutschland von 1991-2013, absolute Werte
- R-Code für das STL-Programm
- R-Code für die Diagnoseplots
- R-Code für die neue Prognosemethode
- R-Code für die Abbildungen in der Arbeit
- R-Code für die Simulation der Prognosegüte

Literaturverzeichnis

Cleveland, R. S. and W. S. Cleveland (1990). decompose time series into trend + seasonal + remainder. URL: <http://www.netlib.org/a/stl> Letzter Seitenaufruf (15.12.2013).

Cleveland, R. S., W. S. Cleveland, J. E. McRae, and I. Terpenning (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics* 6(1), 3–77. URL: <http://cs.wellesley.edu/~cs315/Papers/stl%20statistical%20model.pdf> Letzter Seitenaufruf (15.12.2013).

Deutsche Bundesbank (2013). Produktion im Produzierenden Gewerbe / Deutschland / Produzierendes Gewerbe einschließlich Bau (B - F) / nur kalenderbereinigt. URL: http://www.bundesbank.de/Navigation/DE/Statistiken/Zeitreihen-Datenbanken/Makroökonomische_Zeitreihen/its_details_value_node.html?tsId=BBDE1.M.DE.W.BAA1.A2P000000.G.C.I10.A Letzter Seitenaufruf (15.12.2013).

Foster, N. Lecture 8, models with a trend. Technical report, Universität Wien. URL: <http://homepage.univie.ac.at/neil.foster/TEACHING/ECONO/Lec08%20-%20Stochastic%20and%20Deterministic%20Trends.pdf> Letzter Seitenaufruf (15.12.2013).

Foster, P. The local polynomial regression estimator. Technical report, University of Manchester. URL: <http://www.maths.manchester.ac.uk/~peterf/MATH38011/NPR%20local%20Linear%20Estimator.pdf> Letzter Seitenaufruf (15.12.2013).

Hyndman, R. J. and G. Athanasopoulos (2012). Forecasting: principles and practice. URL: <https://www.otexts.org/fpp> Letzter Seitenaufruf (15.12.2013).

Ripley, B. Seasonal decomposition of time series by loess. URL: <http://stat.ethz.ch/R-manual/R-devel/library/stats/html/stl.html> Letzter Seitenaufruf (15.12.2013).

Statistik der Bundesagentur für Arbeit (2009). Qualitätsbericht: Statistik der arbeitslosen und arbeitsuchenden. URL: <http://statistik.arbeitsagentur.de/Statischer-Content/Grundlagen/Qualitaetsberichte/Generische-Publikationen/Qualitaetsbericht-Statistik-Arbeitslose-Arbeitsuchende.pdf> Letzter Seitenaufruf (15.12.2013).

Statistik der Bundesagentur für Arbeit (2013). *Arbeitslosigkeit im Zeitverlauf*. Nürnberg. URL: http://statistik.arbeitsagentur.de/nn_31892/SiteGlobals/Forms/Rubrikensuche/Rubrikensuche_Form.html?view=processForm&resourceId=210368&input_=&pageLocale=de&topicId=17722&year_month=aktuell&year_month.GROUP=1&search=Suchen Letzter Seitenaufruf (15.12.2013).

Treiber, D. M. (2010). Statistik 2: Vorlesung. Technical report, Technische Universität Dresden. URL: http://vwitme011.vkw.tu-dresden.de/~treiber/statistik2/statistik_download/folien10.pdf Letzter Seitenaufruf (15.12.2013).

Erklärung der Urheberschaft

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Ort, Datum

Unterschrift