Nichtparametrische Modellierung von Zeitreihen mit Infektionshäufigkeiten

Dissertation

zur Erlangung des Grades
eines Doktors der Naturwissenschaften
der Technischen Universität Dortmund

Der Fakultät Statistik
vorgelegt von

Christoph Schürmann
aus Essen

Dortmund 2008

1. Gutachter: Prof. Dr. Ursula Gather

2. Gutachter: Prof. Dr. Katja Ickstadt

Tag der mündlichen Prüfung: 14. Juli 2008

Inhaltsverzeichnis

1 Einleitung

1.1 Ziele der zeitlichen Analyse von Meldefallhäufigkeiten

In einem leistungsfähigen Gesundheitssystem ist die Überwachung und Bewertung der Häufigkeiten, mit denen verschiedene Krankheiten auftreten, eine wesentliche Aufgabe. In Deutschland ist mit dem 2001 in Kraft getretenen Infektionsschutzgesetz (IfSG) eine einheitliche Grundlage für die Erfassung meldepflichtiger Krankheiten gelegt worden. Eine systematische Auswertung der dadurch erhobenen Daten ermöglicht ein zunehmendes epidemiologisches Verständnis dieser Krankheiten und hilft zukünftige Meldefälle zu bewerten, indem man sie mit früheren Beobachtungen vergleicht. Dem zeitlichen Verlauf der Häufigkeiten, mit denen Krankheiten gemeldet werden, kommt dabei eine zentrale Bedeutung zu, da daraus Informationen über mögliche Trends oder saisonale Eigenschaften gewonnen werden können.

Unter diesem Aspekt werden in dieser Arbeit Zeitreihen mit wöchentlich gemeldeten Häufigkeiten bestimmter Infektionskrankheiten untersucht und Dekompositionsmodelle entwickelt, mit denen die Beobachtungen durch die Summe aus Trend-, Saison-, zyklischer und Kalenderkomponente sowie einem zufälligen Fehler modelliert werden. Zur Schätzung der deterministischen Komponenten werden verschiedene nichtparametrische Methoden eingesetzt und miteinander verglichen. Weil hierbei keine Annahmen über die Struktur und Verteilung der Daten vorausgesetzt werden, erlauben diese Methoden eine flexible und datennahe Modellierung. Die betrachteten Verfahren und Modelle sind bislang nicht für die Analyse von Meldefallhäufigkeiten eingesetzt worden. Am Beispiel der Zeitreihen für die Jahre 2001 bis 2006 der in Nordrhein-Westfalen häufigsten meldepflichtigen Infektionskankheiten, Campylobacteriose, Rotavirus-Infektion und Salmonellose, wird hier gezeigt, dass derartige Modelle durch ein sequentielles, nahezu automatisches Verfahren geschätzt werden können und so ein gleichermaßen einfaches wie sinnvolles Ergebnis erreicht wird.

Ein Vergleich mit den Ergebnissen bei Einsatz eines Poissonregressionsmodells verdeutlicht die Vorteile des nichtparametrischen Ansatzes. Für diesen werden abschließend Möglichkeiten zur Vorhersage der Meldefälle aufgezeigt.

1.2 Eigenschaften von Meldefalldatensätzen

Allgemeine Eigenschaften

Die interessierende Variable der Häufigkeit, mit der eine bestimmte Infektionskrankheit innerhalb einer interessierenden Population auftritt, kann in der Regel (ohne unverhältnismäßig hohen Aufwand) nicht beobachtet werden. Zur Ermittlung der Erkrankungs- oder Infektionshäufigkeit in einem interessierenden Zeitabschnitt wäre eine laborgestützte medizinische Untersuchung aller Personen der interessierenden Grundgesamtheit erforderlich, die ggf. durch Alters-, Geschlechts-, räumliche oder sonstige Angaben definiert und eingeschränkt werden kann. Stattdessen kann aber die Häufigkeit beobachtet werden, mit der die Infektionskrankheit bei zuständigen Stellen gemeldet wird.

Systematische Unterschiede zwischen tatsächlichen Infektionshäufigkeiten und der Zahl der gemeldeten Fälle sind durch die folgenden Faktoren gegeben: Eine Infektion kann wegen fehlender Symptome unbemerkt verlaufen, in welchem Fall ein Betroffener keinen Arzt aufsuchen wird, der die Infektion feststellen könnte. Auch bei auftretenden Beschwerden ist ein Arztbesuch nicht zwangsläufig. Selbst wenn sich ein Patient einer ärztlichen Untersuchung unterzieht, kann die korrekte Diagnose ausbleiben. Außerdem ist es möglich, dass der Arzt oder das Labor der Meldepflicht wissentlich oder unwissentlich nicht nachkommt, obwohl eine Infektion festgestellt wurde. Aus diesen Gründen können Meldefälle keinesfalls mit den tatsächlichen Häufigkeiten gleichgesetzt werden; vielmehr ist davon auszugehen, dass sogar die Mehrheit der Erkrankungen unregistriert bleibt. Ein allgemeines, einheitliches und standardisiertes Meldeverfahren ist notwendig, um mit den beobachtbaren Meldefällen eine möglichst proportional zu den tatsächlichen Häufigkeiten verlaufende Variable messen und die Ergebnisse einer zeitlichen Analyse interpretieren zu können.

In dieser Arbeit werden Methoden vorgeschlagen, die für die Anwendung auf Zeitreihen wöchentlich gemeldeter Erkrankungshäufigkeiten oder Inzidenzen geeignet sind. Über den Verlauf der Meldefälle innerhalb eines Jahres ermöglichen sie detailliertere Aussagen als Monats- oder Quartalszahlen. Datensätze mit einer höheren zeitlichen Auflösung sind in der Regel nicht verfügbar und würden auch keinen höheren Informationsgewinn ermöglichen.

Zeitreihen mit wöchentlichen Meldefällen sind insbesondere sogenannten Kalendereffekten ausgesetzt. Bestimmte Wochen können einen lokalen, verzerrenden Einfluss auf die Meldefälle besitzen, wenn sie z. B. einen Feiertag in der Woche enthalten, womit für eine erkrankte Person ein Tag weniger zur Verfügung steht, einen niedergelassenen Arzt zu besuchen. Gerade an Feiertagen, die auf einen Donnerstag oder Freitag fallen, ist es denkbar, dass bei schwachen oder gemäßigten Beschwerden der Arztbesuch nicht oder erst in der darauffolgenden Woche erfolgt. Die theoretisch erfolgte Meldung unterbleibt also. Weitere Störeffekte können Wochen in Schulferien aufweisen. Für diese Zeiträume ist anzunehmen, dass größere Anteile der Bevölkerung als sonst urlaubsbedingt im Ausland sind und damit im Krankheitsfall durch das Meldesystem nicht erfasst werden. Der verzerrende Einfluss auf die Meldezahlen wird in diesem Fall durch Veränderungen in der Grundgesamtheit bewirkt. Solche Veränderungen in der Zusammensetzung der Bevölkerung können nicht nur kurzfristig wie in Ferienzeiten erfolgen, sondern auch langfristig durch Änderungen der Altersstruktur oder der Größe der Bevölkerung. Exakte Zahlen hierzu sind auch über amtliche Stellen wie z. B. statistische Ämter nicht verfügbar, bestenfalls liegen hierfür Schätzungen vor. Auch wenn die Analyse von Inzidenzen (Fälle pro 100 000 Einwohner) als relativer Größe auf den ersten Blick geeignetere Interpretationen erlaubt, werden in dieser Arbeit die absoluten Meldehäufigkeiten betrachtet, um dadurch Fehler zu vermeiden, die durch falsche Bevölkerungsangaben entstehen können.

Allgemeine Aussagen über die Datenqualität derartiger Zeitreihen sind nicht möglich, da sie von der Erhebungsart bzw. der Umsetzung der Meldepflicht und der jeweiligen Kontrolle der Meldedaten abhängen.

Eigenschaften der betrachteten Datensätze

Die Entwicklung eines Modells zur Analyse der wöchentlichen Meldefälle erfolgt in dieser Arbeit am Beispiel der Daten der drei häufigsten meldepflichtigen Infektionskrankheiten in Nordrhein-Westfalen; dies sind Campylobacteriose (CAM), Rotavirus-Infektion (RTV) und Salmonellose (SAL). Diese Infektionen werden meist durch kontaminierte Lebensmittel oder direkten Kontakt von Mensch zu Mensch übertragen und führen nach kurzer Inkubationszeit zu Durchfallerkrankungen mit den entsprechenden Symptomen. Sie treten ganzjährig auf und lösen keine Epidemien aus. Der Abschnitt A.1 im Anhang enthält hierzu einige weitere Informationen.

Eine einheitliche Meldepflicht für diese Krankheiten wurde erst 2001 mit dem seither gültigen Infektionsschutzgesetz (IfSG) eingeführt. Mit diesem wurde das seit 1961 geltende Bundesseuchengesetz (BSeuchG) abgelöst, das erstmalig dazu verpflichtete, Erkrankung und/oder Tod durch bestimmte Krankheiten

der zuständigen Gesundheitsbehörde zu melden. Durch das IfSG ist der Katalog meldepflichtiger Krankheiten erweitert und die eigentliche Meldung durch standardisierte, leichter vergleichbare Falldefinitionen vereinheitlicht worden. Da die vor 2001 erhobenen Daten nicht mit den danach erhobenen vergleichbar sind, werden nur die seit Januar 2001 vorliegenden Meldefälle berücksichtigt.

Die zentrale Erfassung der Meldefälle erfolgt durch das Robert Koch-Institut (RKI). Es registriert alle von den jeweiligen Gesundheitsämtern erfassten und über die zuständigen Landesstellen weitergeleiteten Fälle. Auf diesem Weg findet sowohl bei den Landesstellen wie auch beim Robert Koch-Institut eine genaue Prüfung jeder Einzelmeldung statt. Durch die mehrfache Kontrolle ist die Qualität des Datenmaterials sehr hoch. Falsche zeitliche Angaben zum Meldefall wie auch fehlende Angaben zu patientenbezogenen Größen sind nahezu auszuschließen. Dazu zählen beispielsweise Geschlecht, Alter der betroffenen Person sowie räumliche Angaben, die jedoch bei der rein zeitlichen Betrachtung nicht relevant sind. Die Daten sind über das RKI öffentlich zugänglich und z. B. im Internet unter `www.rki.de/survstat` oder in den Infektionsepidemiologischen Jahrbüchern des RKI zugänglich.

Die Zeitreihen der wöchentlichen Häufigkeiten der Meldefälle, an deren Beispiel die in dieser Arbeit vorgestellten Methoden erläutert werden, sind in den Abbildungen 1.1 – 1.3 graphisch dargestellt. Die Meldewoche 1 ist die erste Woche im Jahr 2001, die Meldewoche 313 ist die letzte (52.) Woche im Jahr 2006. Vertikale Trennlinien markieren die Jahreswechsel. Deutlich erkennbar ist in jeder Abbildung ein saisonaler Verlauf der Meldefälle. Demnach gibt es für jede Krankheit eine Ausbruchsperiode, also einen Zeitraum, in dem besonders viele Meldungen erfolgen. Diesem stehen Perioden mit relativ wenigen Fällen gegenüber. Für Campylobacter und Salmonellen gilt, dass der zeitliche Verlauf eine höhere lokale Variabilität als bei Rotavirus aufweist. Diese unterscheidet sich außerdem durch die deutliche Zunahme an Fällen in den Ausbruchszeiträumen der letzten Jahre. Weiterhin bemerkenswert ist bei den Campylobacterfällen der während des Jahreswechsels beobachtbare Effekt, dass in der letzten Woche eines Jahres auffällig wenig, in den ein bis zwei darauffolgenden Wochen aber auffällig viele Meldungen eingehen.

Für eine weitere deskriptive Beschreibung der Daten listet Tabelle 1.1 für die jeweiligen Krankheiten die Anzahl aller im Beobachtungszeitraum 2001 bis 2006 gemeldeten Fälle, die minimalen und maximalen Meldezahlen unter Angabe der Woche, in der sie beobachtet wurden, sowie weitere Verteilungskennzahlen auf.

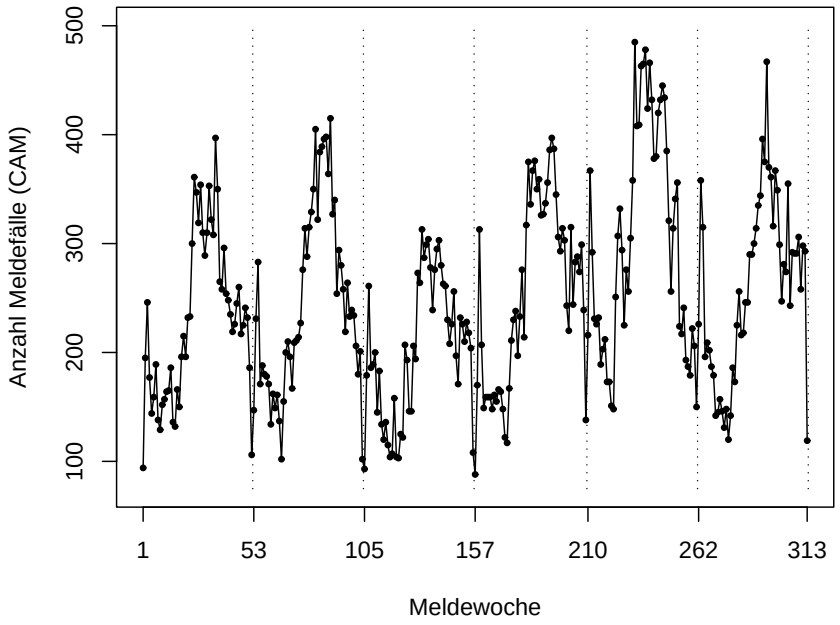

Abb. 1.1: Campylobacteriose (CAM): Meldefälle in Nordrhein-Westfalen von 2001 bis 2006.

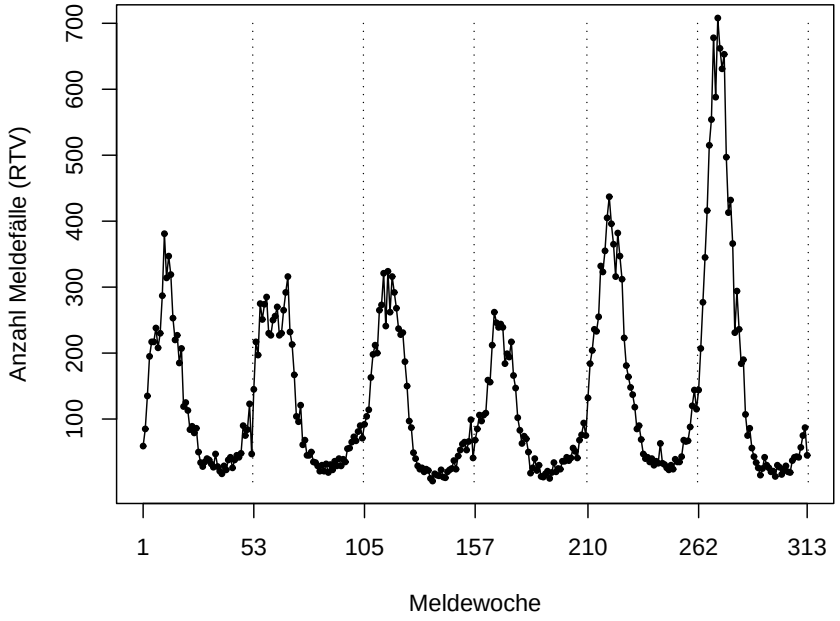

Abb. 1.2: Rotavirus-Infektion (RTV): Meldefälle in Nordrhein-Westfalen von 2001 bis 2006.

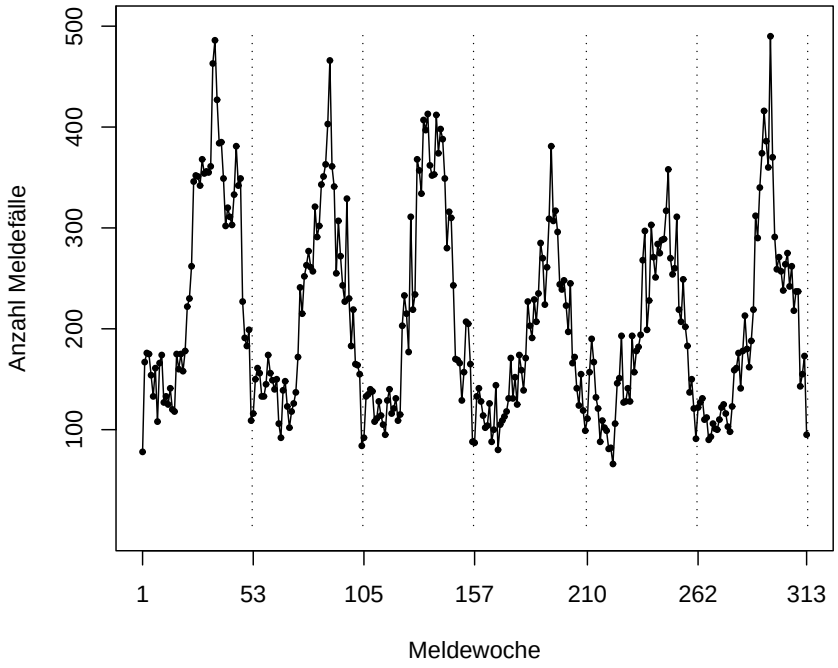

Abb. 1.3: Salmonellose (SAL): Meldefälle in Nordrhein-Westfalen von 2001 bis 2006.

	CAM	RTV	SAL
Summe aller Fälle	77 917	41 506	65 724
Arithm. Mittel	248.9	132.6	210.0
Std.abweichung	88.1	134.2	95.7
Minimum (Woche)	88 (157)	6 (137)	66 (222)
5%-Quantil	120	18	95
25%-Quantil	179	35	131
Median	234	75	178
75%-Quantil	310	217	275
95%-Quantil	405	382	384
Maximum (Woche)	485 (232)	708 (271)	490 (296)

Tab. 1.1: Kennzahlen der betrachteten Zeitreihen.

Woche t	Beginn	Woche t	Beginn
1	1. 01. 2001	157	29. 12. 2003
53	31. 12. 2001	210	3. 1. 2005
105	30.12. 2002	262	2. 1. 2006

Tab. 1.2: Zeitpunkte für den Beginn der jeweils ersten Woche eines Jahres.

Jeder Datensatz von Meldefällen einer Krankheit ist eine Zeitreihe mit 313 Beobachtungen. Jede Beobachtung z_t, $t = 1, \ldots, 313$, entspricht der Anzahl der gemeldeten Fälle dieser Krankheit, die in der Woche t seit Beginn des Jahres 2001 gemeldet wurden. Meldewochen und Kalenderwochen sind identisch, sie beginnen jeweils montags und enden sonntags. Die Woche $t = 1$ beginnt am Montag, 1. Januar 2001 und endet am Sonntag, 7. Januar 2001. Die Woche $t = 313$ beginnt am Montag, 25. Dezember 2006 und endet am Sonntag, 31. Dezember 2006.

Da das Jahr mit 365 bzw. 366 Tagen in keinem Fall eine ganzzahlige Einteilung in Wochen ermöglicht, verschiebt sich der Beginn der jeweils ersten Woche eines Jahres relativ gesehen zum tatsächlichen Jahresbeginn. So beginnt beispielsweise die erste Woche im Jahr 2001 am Montag, 1. Januar, aber die erste Woche des Jahres 2002 beginnt bereits am Montag, 31. Dezember 2001. Eine weitere Besonderheit bildet das Jahr 2004. Einerseits handelt es sich um ein Schaltjahr mit 366 Tagen, andererseits umfasst es 53 Wochen. Tabelle 1.2 listet die jeweils erste Woche eines Jahres mit dem Tag ihres Beginns auf. Bezogen auf den Jahresbeginn am 1. Januar liegen die Meldewochen verschiedener Jahre demnach nicht parallel, sondern um wenige Tage versetzt zueinander. Dies muss bei der späteren Untersuchung saisonaler bzw. periodischer Effekte berücksichtigt werden.

Auch wenn alle Meldewochen 7 Tage umfassen und in dieser Hinsicht identisch sind, unterscheiden sie sich hinsichtlich der Anzahl der enthaltenen Feiertage. Für die praktische Umsetzung der Meldepflicht ist es bedeutsam, ob eine Woche 4 oder 5 Werktage enthält, insbesondere wenn ein Feiertag auf einen Freitag fällt. Da die Krankheitssymptome oft nicht so schwerwiegend sind, dass eine stationäre Behandlung erforderlich ist, ist davon auszugehen, dass Patienten in der Regel einen niedergelassenen Arzt mit eigener Praxis aufsuchen. Ein solcher Arzt kann üblicherweise nicht an einem Feiertag aufgesucht werden. Möglich ist, dass der Patient dann den nächsten Arbeitstag abwartet, um zum Arzt zu gehen, oder entscheidet, überhaupt nicht zum Arzt zu gehen, wenn z. B. die Symptome nachlassen. Aufgrund dieser und ähnlicher Überlegungen ist zu folgern, dass Meldefälle in Wochen mit Feiertagen, insbesondere mit Feiertagen am Ende der Woche wie donnerstags und freitags, anders bewertet werden müssen als sol-

Feiertag	Datum	Feiertag	Datum
Neujahr	1. Januar	Pfingstmontag	wechselnd
Karfreitag	wechselnd	Fronleichnam	wechselnd
Ostermontag	wechselnd	Tag der dt. Einheit	3. Oktober
Maifeiertag	1. Mai	Allerheiligen	1. November
Christi Himmelfahrt	wechselnd	Weihnachten	25. /26. Dezember

Tab. 1.3: Feiertage in Nordrhein-Westfalen und ihre Lage im Jahr. Die Lage wechselnder Feiertage wird bestimmt durch Ostern. Dessen Zeitpunkt wird durch den Mondkalender festgelegt und liegt zwischen dem 22. März und dem 25. April.

che, die in Wochen ohne Feiertage auftreten. Tabelle 1.3 listet die Feiertage in Nordrhein-Westfalen auf.

Entsprechend sind neben Wochen, die Feiertage enthalten, auch solche gesondert zu beachten, die in übliche Ferienzeiten fallen. Dazu zählen die Wochen um Weihnachten und Neujahr sowie die Wochen vor und nach Ostern. Hier ist nicht nur zu bedenken, dass Arztpraxen in diesen Zeiträumen häufiger geschlossen sind, sondern dass dies wegen der Schulferien auch typische Reisezeiten sind. Da ein Teil der Bevölkerung sich dann außerhalb von Nordrhein-Westfalen aufhält, ist zu erwarten, dass die beobachtete Zahl der Erkrankungen und damit der Meldefälle im jeweiligen Zeitraum geringer ist als in den übrigen Wochen. Ein ähnlicher Effekt ist auch für Meldewochen denkbar, die in den sonstigen Wochen mit Schulferien wie Sommer- und Herbstferien liegen. Von den Weihnachtsferien abgesehen sind alle Schulferien beweglich, beginnen also von Jahr zu Jahr in verschiedenen Wochen.

1.3 Überblick über mögliche Analyseverfahren

Am Beginn einer systematischen Analyse der im Rahmen gesetzlicher Bestimmungen erhobenen Infektionshäufigkeiten stehen zunächst explorative Verfahren, die mit dem Ziel der Beobachtung und Überwachung auch Surveillance-Verfahren genannt werden, vgl. z. B. Exner (1997). Sie werden typischerweise eingesetzt zur Deskription des zeitlichen und / oder räumlichen Verlaufs, zur Ermittlung und Quantifizierung weiterer Einfluss- oder Risikofaktoren wie z. B. Alter und Geschlecht sowie zum Auffinden unregelmäßiger Häufungen oder Cluster. Eine jeweils kurze Einführung und Überblick über Methoden, die in diesen Situationen eingesetzt werden, geben u. a. Devine (2004), Brookmeyer (2004) und Waller (2004).

Zur Beantwortung von Fragen, in denen der zeitliche Verlauf von zentraler Bedeutung ist, sind in der Literatur zahlreiche Ansätze bekannt. Die statistische Analyse der Auftretenshäufigkeit von Infektionskrankheiten kann unter vielen verschiedenen Fragestellungen erfolgen. Einige von diesen stellt die folgende Übersicht vor.

Die erste mathematische Formulierung eines Modells, das auch eine zeitliche Betrachtung ermöglicht, geben Kermack und McKendrick (1927). Darin können unter Berücksichtigung des individuellen Krankheitsstatus (anfällig, infiziert, genesen) Übergangswahrscheinlichkeiten für Erkrankungszustände für beliebige Zeitpunkte geschätzt werden. Einen neueren Ansatz, in dem derartige SIR-Modelle (*susceptible, infected, recovered*) betrachtet werden, stellt beispielsweise Haber (1997) vor. SIR-Modelle eignen sich vor allem für Krankheiten, die hauptsächlich durch direkten Kontakt übertragen werden wie z. B. Masern oder Influenza, und bei denen genaue Angaben über den Gesundheitsstatus der Individuen bekannt sind. Sie eignen sich vornehmlich bei kleinen bis mittelgroßen Populationen, wenn auch Übertragungswege genau verfolgt werden können. Werden wie hier größere Bevölkerungsgruppen betrachtet, sind sie weniger geeignet.

Statistische Arbeiten, die Zeitreihen insbesondere hinsichtlich bevorstehender Ausbruchs- oder Epidemiebeginne untersuchen, nutzen dazu typischerweise Schwellwertmodelle. Die Überschreitung einer durch den früheren Verlauf der Zeitreihe gegebenen Grenzwert markiert dabei den Beginn einer Ausbruchsperiode. Die frühzeitige Erkennung von Ausbrüchen kann auf einen lokalen Entstehungsherd hinweisen und darauf aufbauend gezielte Präventionsmaßnahmen ermöglichen. Beispiele für solche Ansätze geben Farrington et al. (1996), Hashimoto et al. (2000), Ranta et al. (2004) oder Heisterkamp et al. (2006). Der Übertragungsweg der Krankheiten wird hier ebenso wie in den im folgenden aufgeführten Modellen nicht explizit berücksichtigt. Damit eigenen sie sich auch zur Analyse von teilweise oder hauptsächlich endemisch auftretenden Krankheiten, d. h. solchen, die weniger durch direkte Übertragung von Mensch zu Mensch als wegen des andauernden Vorhandenseins einer durch z. B. Umweltfaktoren begründbaren Ursache auftreten.

Neben rein zeitlichen Modellen werden zunehmend auch solche betrachtet, die weitere Kovariablen berücksichtigen. Dies können einerseits äußere Faktoren sein wie zeitgleiche Umwelteinflüsse in Form von Luftverschmutzung, Wasserqualität oder Temperatur, die z. B. in Chiogna und Gaetan (2005) oder Ethelberg et al. (2005) einbezogen werden. Andererseits sind personenspezifische Eigenschaften wie Alter, Geschlecht oder Wohnort bedeutsam, wenn Unterschiede

zwischen ihren verschiedenen Ausprägungen und deren Einfluss auf die Erkrankungshäufigkeit untersucht werden. So berücksichtigen beispielsweise Lindbäck und Svenson (2001) auch geschlechtsbedingte Unterschiede und Held et al. (2005) verschiedene Altersgruppen in multivariaten Zeitreihen. Held et al. (2006) stellen ein Modell vor, in dem für jeden Zeitpunkt auch der epidemische Status bekannt ist. Zunehmend werden in den letzten Jahren auch Modelle betrachtet, die sich allein auf die Analyse des räumlich-zeitlichen Verhaltens der Meldefälle konzentrieren, Beispiele hierzu geben MacNab und Dean (2001), Mugglin et al. (2002), Knorr-Held und Richardson (2003) oder Diggle et al. (2005).

In den genannten Arbeiten werden die Zielvariablen mehrheitlich als poisson- bzw. negativ binomial verteilt angesehen. Dabei existieren zur Modellierung der Parameter selbst verschiedene Ansätze, sie können sowohl durch deterministische (z. B. trigonometrische) Funktionen wie auch durch stochastische (z. B. autoregressive) Prozesse gesteuert werden. Häufig erfolgt die Umsetzung dann in einem bayesianischen Kontext, beispielsweise durch hierarchische Modelle. Hier können Markov Chain Monte Carlo (MCMC) Methoden eingesetzt werden, vgl. Gilks et al. (1994), um die analytisch meist nicht berechenbaren Schätzwerte zu bestimmen.

Ein nichtparametrisches Dekompositionsmodell

Ziel dieser Arbeit ist die Modellierung des beobachteten zeitlichen Verlaufs der Häufigkeiten von Krankheiten, die häufig auftreten, jahreszeitliche Eigenschaften haben und eine kurze Inkubationszeit besitzen. Ein allgemeines Verfahren wird entwickelt, mit dem solche wöchentlichen Zeitreihen standardisiert modelliert werden können. Dabei geht die Annahme ein, dass die Beobachtungen durch die Summe aus einer unregelmäßigen, zufälligen Fehler- sowie aus einer glatten, deterministischen Signal-Komponente dargestellt werden können. Um verwertbare, infektionsepidemiologische Erkenntnisse über den Verlauf der Häufigkeiten gewinnen zu können, wird das Signal additiv in die weiteren Komponenten Trend, Saison, zyklische und Kalenderkomponente zerlegt. Sowohl die deterministischen Komponenten wie auch der zufällige Fehler sollen dabei eine möglichst einfache Struktur besitzen, was jeweils durch individuelle Kriterien geeignet spezifiziert werden muss.

Nichtparametrische Ansätze sind zur Dekomposition von Zeitreihen mit Meldefällen von Infektionskrankheiten bislang nicht entwickelt. Das breite Spektrum nichtparametrischer Verfahren, die prinzipiell in der Zeitreihenanalyse eingesetzt werden können, vgl. Härdle et al. (1997), lässt jedoch eine geeignete Modellierung mit diesen Verfahren aussichtsreich erscheinen. Der Verzicht auf die bei klassischen Verfahren üblichen Annahmen bzgl. Verteilung der Zielvariablen

und dem Aufbau der Regressoren, die oft nicht durch den eigentlichen Untersuchungsgegenstand sondern aus modellierungspraktischen Gründen erfolgt, erlaubt zudem eine rein beobachtungsgesteuerte Analyse. Bei der Wahl der Methoden kommen daher ausschließlich nichtparametrische Verfahren zum Einsatz.

Mit ihnen können die jeweiligen Komponenten des Dekompositionsmodells geschätzt werden. Da diese Elemente teils einen globalen, teils einen lokalen Verlauf erklären, sind jeweils verschiedene Arten von Verfahren zu berücksichtigen. Damit wird mehr als nur eine Klasse wie z. B. die von Heiler und Feng (2000) vorgestellte mit polynomiellen und trigonometrischen Funktionen betrachtet. Auch die besonderen kalendarischen Eigenschaften wöchentlicher Daten sowie besondere Effekte einzelner Kalenderwochen dürfen bei der Modellierung nicht vernachlässigt werden.

2 Nichtparametrische Regressionsverfahren

2.1 Nichtparametrische Regressionsmodelle

Für das allgemeine nichtparametrische Regressionsmodell geht man von einem funktionalen Zusammenhang zwischen einer nicht zufälligen Regressorvariable X und einer zufälligen Zielvariable Y aus. Für die Beobachtungen $(x_i, y_i), i = 1, \ldots, n$, wird angenommen, dass $y_i, i = 1, \ldots, n$, Realisierungen von Zufallsvariablen $Y_i, i = 1, \ldots, n$, sind. Der Erwartungswert von Y_i an Stellen x_i, $i = 1, \ldots, n$, die auch Designpunkte genannt werden, ist durch den Wert $\mathrm{E}\, Y_i = f(x_i)$ einer unbekannten Funktion $f(.)$ gegeben, die auch Signal genannt wird. Die zufälligen Eigenschaften von Y entstehen durch additiv wirkende, nicht beobachtbare Fehler ε_i, $i = 1, \ldots, n$, die als unabhängig angenommen werden mit konstantem Erwartungswert $\mathrm{E}\, \varepsilon_i = 0$ und konstanter Varianz $\mathrm{Var}\, \varepsilon_i = \sigma^2$. Die Annahme der Normalverteilung für die Fehler erfolgt zwar oft, ist aber nicht zwingend. Das nichtparametrische Regressionsmodell ist damit durch die charakteristische Gleichung

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

gegeben.

Zur Schätzung der unbekannten Funktion f sind zahlreiche nichtparametrische Methoden bekannt. Je nach Herangehensweise unterscheidet man u. a. zwischen Kernschätzern, lokaler polynommieller Regression, Splines und Penalised-Likelihood-Verfahren. Einen Überblick über die jeweiligen Verfahren und die ihnen zugrunde liegenden Prinzipien geben z. B. Wahba (1990), Fan und Gijbels (1996), Simonoff (1996) oder Eubank (1999). Über die Anwendungsmöglichkeiten dieser Methoden zur Analyse von Zeitreihen informieren u. a. Györfi et al. (1989) und Fan und Yao (2003).

In dieser Arbeit werden zur Schätzung der Zeitreihenkomponenten ein Kernschätzerverfahren mit lokaler Bandbreitenwahl, adaptiv gewichtete Glättungsspli-

nes (*adaptive weights smoothing*) und eine glatte Variante des Verfahrens Straffe Saite (*taut string*) eingesetzt. Eine ausführliche, vergleichende Diskussion dieser Verfahren geben Davies et al. (2008). Zusätzlich wird das Verfahren *singular spectrum analysis* verwendet. Dies ist zwar kein Regressionsverfahren, aber die Ergebnisse können unter bestimmten Bedingungen als geeignete Schätzer des unbekannten Signals angesehen werden.

2.2 Lokale Kernschätzung mit adaptiver Bandbreite

Ausgehend von den Arbeiten durch Nadaraya (1964) und Watson (1964) sind Kernschätzer wichtige und weit verbreitete Methoden in der nichtparametrischen Regressionsschätzung geworden. Allgemein ist ein Kernschätzer an einer Stelle x als ein gewichtetes Mittel benachbarter Beobachtungen gegeben durch

$$\hat{f}(x) = \arg\min_c \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) (y_i - c)^2. \tag{2.2}$$

Die Wahl der Gewichte erfolgt durch eine sogenannte Kernfunktion K in Form einer Dichte und eine Bandbreite h, mit der der Grad der Glättung kontrolliert wird.

Während die Wahl einer Kernfunktion meist weniger relevant ist, kommt einer geeigneten Bestimmung der Bandbreite in der Regel die meiste Bedeutung zu. Statt des häufig gewählten Vorgehens, die Bandbreite global festzulegen, gibt es auch Ansätze, sie lokal zu bestimmen. Vorteilhaft ist dies zum einen, wenn der Abstand der Designpunkte x_i verschieden groß ist: Liegen sie weit auseinander, erscheint die Wahl einer größeren Bandbreite sinnvoll, die zu einer glatteren Schätzung führt, während bei eng liegenden Punkten eine schmalere Bandbreite eine genauere Anpassung an den lokalen Funktionsverlauf ermöglicht. Zudem ermöglicht eine lokale Bandbreitenwahl auch die Schätzung von Signalen in Modellen mit heteroskedastischen Störgrößen: An Stellen mit hoher Variabilität ist eine glattere, weniger variable Schätzung wünschenswert als an Stellen mit weniger Variabilität, wo die Schätzung eine stärkere Anpassung an die Daten liefern soll. Schließlich erlaubt eine lokale Bandbreitenwahl auch eine Anpassung an die Struktur der Regressionsfunktion, wenn sie in Bereichen mit geringer Variabilität des Signals auch eine glatte und in solchen mit hoher Variabilität eine genauere Schätzung ergibt. Die Wahl der Bandbreite erfolgt bei lokalen Verfahren sinnvollerweise adaptiv durch die gegebenen Daten.

Brockmann et al. (1993) haben ein geeignetes Verfahren mit lokaler Bandbreitenwahl entwickelt, das insbesondere wegen seiner Eigenschaft, sich an die lokale Variabilität von f anzupassen, verwendet wird. Ursprünglich nur für Situationen entwickelt, in denen die Designpunkte $x_i, i = 1, \ldots, n$, gleichen Abstand voneinander haben und im Intervall $[0, 1]$ liegen, ist diese Methode, im weiteren Verlauf durch LOK (*local kernel regression*) bezeichnet, durch Herrmann (1997) u. a. auch auf nicht äquidistante Designs erweitert worden. Sie wird im folgenden kurz skizziert.

Für eine jeweilige Stelle x auf einem kompakten Intervall $[a, b]$ und eine zugehörige Bandbreite $h(x)$ ist die geschätzte Signalkomponente im inneren Bereich von $[a, b]$ gegeben durch

$$\hat{f}(x, h(x)) \;=\; \sum_{i=1}^{n} y_i \int_{s_{i-1}}^{s_i} \frac{1}{h(x)} K\left(\frac{x-u}{h(x)}\right) du, \qquad x \in [a+h(x), b-h(x)]. \quad (2.3)$$

Dabei sind die Grenzen des Integrals mit $s_i = 1/2(x_i + x_{i+1})$ die Mittel zwischen den Designpunkten. $K(.)$ ist eine symmetrische Kernfunktion k-ter Ordnung mit Träger $[-1, 1]$ sowie $\int K(u)du = 1$ und $\int uK(u)du = 0$. Weitere eher technische Anforderungen an f und K sind Herrmann (1997) zu entnehmen. Dort werden auch entsprechende Möglichkeiten zur Schätzung $\hat{f}(x, h(x))$ in den Randbereichen aufgezeigt, auf die hier nicht näher eingegangen wird.

Die Güte von $\hat{f}(x, h(x))$ wird lokal durch den mittleren quadratischen Fehler (*mean squared error*, MSE)

$$\mathrm{MSE}(\hat{f}(x, h(x))) \;=\; \mathrm{E}\left(\hat{f}(x, h(x)) - f(x)\right)^2 \quad (2.4)$$

und global durch den mittleren integrierten quadratischen Fehler (*mean integrated squared error*, MISE)

$$\mathrm{MISE}(\hat{f}) \;=\; \mathrm{E} \int_a^b w(x) \left(\hat{f}(x, h(x)) - f(x)\right)^2 dx \quad (2.5)$$

beurteilt, wobei mit einer geeigneten Gewichtungsfunktion $w(.) > 0$ der Einfluss von Randeffekten an den Intervallgrenzen von $[a, b]$ reduziert wird.

Eine hinsichtlich dieser Kriterien asymptotisch MISE-optimale globale Bandbreite h_A ist gegeben durch

$$h_A \;=\; \left(\frac{I_{g,\sigma^2}}{nI_k(f)} C(K)\right)^{\frac{1}{2k+1}}. \quad (2.6)$$

Dabei ist g die diskrete Dichtefunktion der Designpunkte, für die zweifache stetige Differenzierbarkeit auf $[a, b]$ angenommen wird, und deren zugehörige Verteilungsfunktion G die Eigenschaft $G^{-1}((i - 0.5)/n) = x_i$ besitzt. Durch das Funktional $I_{g,\sigma^2} = \int w(x)\sigma^2(x)/g(x)dx$ wird die globale Varianz beschrieben, wobei die unterschiedliche Dichte der Designpunkte berücksichtigt ist. Sie wird in Relation zur globalen Variabilität $I_k(f) = \int w(x)(f^{(k)}(x))^2 dx$ gesetzt, die durch die k-te Ableitung von f erfasst wird. $C(K)$ ist eine durch die Kernfunktion K gegebene Konstante mit

$$C(K) \;=\; \frac{(k!)^2 \int \left(K(x)\right)^2 dx}{2k\left(\int x^k K(x)dx\right)^2}.$$

Für die asymptotisch MSE-optimale lokale Bandbreite $h_A(x)$ erhält man analog

$$h_A(x) \;=\; \left(\frac{\sigma^2(x)}{n(f^{(k)}(x))^2 g(x)} \cdot C(K)\right)^{\frac{1}{2k+1}}, \qquad \text{falls } f^{(k)}(x) \neq 0. \tag{2.7}$$

Um eine stabile Schätzung von h_A und $h_A(x)$ auch an Stellen zu erhalten, an denen $f^{(k)}(x)$ nahe 0 ist, schlägt Herrmann (1997) das folgende iterative Verfahren vor: In einem ersten Schritt wird die Bandbreite durch $\hat{h}_0 = (k-1)(b-a)/n$ global geschätzt. Es folgen $j = 1, \ldots, (k+1)(2k+1)$ weitere Schritte, in denen dieser Schätzer aktualisiert wird. Wie in (2.6) enthält er in jedem Schritt j im Zähler einen Schätzer für die Varianz und im Nenner einen für die k-te Ableitung von f, für deren Berechnung jeweils das Ergebnis $\hat{h}_j$ des vorigen Schrittes eingesetzt wird. Nach der letzten Iteration ist dann ein globaler Schätzer $\hat{h}$ für die Bandbreite gegeben, die auch notwendig ist, um in einem abschließenden Schritt den lokalen Bandbreitenschätzer $\hat{h}(x)$ berechnen zu können. Unter Verwendung von $\hat{h}$ können dann wiederum sogenannte „Plug-In-Schätzer“ für Zähler und Nenner in (2.7) gebildet werden. Die daraus resultierende Schätzung von $h(x)$ wird an den Randbereichen von $[a, b]$ durch eine Linearkombination von $\hat{h}$ und $\hat{h}(x)$ ersetzt, deren Koeffizienten durch eine geeignet zu wählende Gewichtsfunktion gegeben sind. Die vollständige und genaue Darstellung des Algorithmus' findet sich bei Herrmann (1997).

Durch Einsetzen von $\hat{h}(x)$ in (2.3) ist dann eine lokale Schätzung von $f(x)$ an jedem Punkt $x \in [a, b]$ möglich.

2.3 Adaptive Weights Smoothing

Das Verfahren des *adaptive weights smoothing* (AWS) ist erstmals durch Polzehl und Spokoiny (2000) für den Anwendungsfall der Bildrekonstruktion und

-segmentierung vorgestellt worden. Diese nichtparametrische Methode schätzt die unbekannte Signalfunktion durch abschnittsweise definierte Konstanten, wobei durch ein lokales und adaptives Verfahren die Umgebung, aus der Beobachtungen zur Schätzung an einer Stelle herangezogen werden, datengesteuert ermittelt wird. Um eine einfache Schätzung des Signals zu erhalten, werden die lokalen Umgebungen so groß wie möglich gewählt, solange die resultierende Schätzung in diesem Bereich eine gute Anpassung darstellt. Da die Bestimmung der Schätzer algorithmisch einfach umsetzbar ist und die Methode außerdem an den Rändern des untersuchten Bereichs unverzerrt ist, hat sie vorteilhafte Eigenschaften. Polzehl und Spokoiny (2003) haben das Verfahren erweitert, indem statt Konstanten auch weitere Funktionen wie z. B. Polynome lokal angepasst werden können. Da das geschätzte Signal damit auch durch eine glatte Funktion dargestellt werden kann, ist dieses Verfahren aussichtsreich für die spätere Anwendung bei der Modellierung der Meldefälle. Es wird an dieser Stelle für die eindimensionale Situation vorgestellt.

Zunächst wird ein Regressionsmodell entsprechend (2.1) zugrunde gelegt. Es wird dann jedoch die Annahme getroffen, dass die Funktion f Element einer parametrischen Familie $\mathcal{F} = \{f_{\boldsymbol{\vartheta}}, \vartheta \in \Theta\}$ ist, die durch gegebene Funktionen $\psi_1, \ldots, \psi_p$ und einen $p-$dimensionalen Parametervektor $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_p)'$ definiert ist, so dass gilt

$$f_{\boldsymbol{\vartheta}}(x) \;=\; \vartheta_1\psi_1(x) + \ldots + \vartheta_p\psi_p(x).$$

Ist diese sogenannte globale parametrische Annahme erfüllt, ist ein Schätzer $\hat{\boldsymbol{\vartheta}}$, der den quadratischen Abstand zwischen Y_i und $f(x_i)$, $i = 1, \ldots, n$, minimiert, gegeben durch

$$\begin{aligned}
\hat{\boldsymbol{\vartheta}} &= \arg\min_{\boldsymbol{\vartheta}\in\Theta} \sum_{i=1}^{n} \left(Y_i - f_{\boldsymbol{\vartheta}}(x_i)\right)^2 \\
&= \left(\sum_{i=1}^{n} \Psi_i\Psi_i'\right)^{-1} \sum_{i=1}^{n} \Psi_i Y_i \qquad (2.8) \\
&= \left(\Psi\Psi'\right)^{-1} \Psi Y.
\end{aligned}$$

Für $i = 1, \ldots, n$ ist Ψ_i der Vektor der jeweiligen Funktionswerte $\psi_1(x_i), \ldots, \psi_p(x_i)$ an der zugehörigen Stelle, und Ψ ist die aus der Gesamtheit der Vektoren erstellte Matrix $\Psi = (\Psi_1, \ldots, \Psi_n)$. Um (2.8) lösen zu können, muss $\Psi\Psi'$ invertierbar sein.

Indem die parametrische Annahme $f_{\boldsymbol{\vartheta}} \in \mathcal{F}$ nur für begrenzte Umgebungen und nicht mehr global aufrecht erhalten wird, ist eine lokale Modellierung möglich. Die Schätzung von $f_{\boldsymbol{\vartheta}}$ an einer bestimmten Stelle x_i kann durch Einführung

geeigneter Gewichte $w_i(x)$, die den Beobachtungen Y_i zugewiesen werden, lokal erfolgen. Mit $W(x) = \mathrm{diag}\{w_1(x), \ldots, w_n(x)\}$ ist dann in Analogie zu (2.8)

$$\hat{\boldsymbol{\vartheta}}(x) \;=\; \arg\min_{\boldsymbol{\vartheta}\in\Theta} \sum_{i=1}^{n} w_i(x) \left((y_i - f_{\boldsymbol{\vartheta}}(x_i)\right)^2 \;=\; (\Psi W \Psi')^{-1}\, \Psi W \mathbf{y}. \tag{2.9}$$

Das AWS-Verfahren verwendet dabei Gewichte $w_i(x)$, die aus verschiedenen Straftermen zusammengesetzt sind und in Abhängigkeit einer Bandbreite h gewählt werden. Einen dieser Strafterme stellt die Lokations-Kernfunktion $K_l(\mathbf{l}_i)$ dar, in die als Argument der durch

$$\mathbf{l}_{ij} = \left(\frac{x_j - x_i}{h}\right)^2 \tag{2.10}$$

gegebene, gewichtete und quadrierte Abstand zweier Beobachtungspunkte x_i und x_j eingeht. Die Kernfunktion muss die Bedingung $K_l(0) = 1$ erfüllen und monoton nicht steigend sein. Um den rechentechnischen Aufwand zu reduzieren, werden Kernfunktionen mit kompaktem Träger verwendet wie $K_l(u) = (1-u)\mathbb{I}_{[0,1]}(u)$. Die Wahl von K_l hat jedoch keinen entscheidenden Einfluss auf $\hat{\boldsymbol{\vartheta}}(x)$.

Ein weiterer Strafterm wird durch Vergleich von Schätzungen an zwei Stellen x_i und x_j gewählt: Bei einer lokalen Schätzung von $\boldsymbol{\vartheta}(x)$ muss entschieden werden, aus welchem umgebenden Bereich von x Beobachtungen zur Schätzung herangezogen werden. Das AWS-Verfahren ist adaptiv, indem es ausgehend von einer anfänglichen Bandbreite $h^{(0)}$ das zugehörige lokale Intervall schrittweise erweitert, solange die neu hinzukommenden Beobachtungen geeignet durch eine Funktion $f_{\boldsymbol{\vartheta}} \in \mathcal{F}$ in diesem Intervall modelliert werden kann: Die Beobachtungen um einen Punkt x_i werden dabei durch $W(x_i) = \mathrm{diag}\{w_1(x_i), \ldots, w_n(x_i)\}$, im folgenden mit $W_i = \mathrm{diag}\{w_{i1}, \ldots, w_{in}\}$ bezeichnet, gewichtet. Ist die Funktion f, repräsentiert durch den Vektor $\boldsymbol{\vartheta}$, in einer lokalen Umgebung um zwei Punkte x_i und x_j identisch, so gilt dies auch für die Matrizen der Gewichte W_i und W_j. Daraus folgt, dass die Bandbreite $h^{(k)}$ in einem Schritt k erweitert werden kann, solange der Unterschied zwischen $W_i^{(k)}$ und $W_j^{(k)}$ gering ist. Dies wird mit einem Likelihood-Quotienten-Test untersucht, der die Hypothese $H_0 : \boldsymbol{\vartheta}_i = \boldsymbol{\vartheta}_j$ überprüft. Für diesen ist zunächst für einen Parameterschätzer $\hat{\boldsymbol{\vartheta}}_i$ die Likelihood L von $\boldsymbol{\vartheta}$ gegeben durch

$$L(W_i, \hat{\boldsymbol{\vartheta}}_i, \boldsymbol{\vartheta}) \;=\; \frac{1}{2\sigma^2}(\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta})'\Psi W_i \Psi'(\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}).$$

Der Likelihood-Quotienten-Test wird dann mit einer asymmetrischen Form der Teststatistik

$$T_{ij} \;=\; L(W_i, \hat{\boldsymbol{\vartheta}}_i, \boldsymbol{\vartheta}) - L(W_i, \hat{\boldsymbol{\vartheta}}_j, \boldsymbol{\vartheta}) \;=\; \frac{1}{2\sigma^2}(\hat{\boldsymbol{\vartheta}}_i - \hat{\boldsymbol{\vartheta}}_j)'\Psi W_i \Psi'(\hat{\boldsymbol{\vartheta}}_i - \hat{\boldsymbol{\vartheta}}_j) \tag{2.11}$$

durchgeführt, da diese gegenüber der symmetrischen Variante den Vorteil besitzt, dass die Likelihood von $\hat{\boldsymbol{\vartheta}}_j$ hinsichtlich der durch W_i bestimmten Umgebung um x_i beurteilt wird. Da T_{ij} umso größer wird, je stärker sich die Schätzungen $\hat{\boldsymbol{\vartheta}}_i$ und $\hat{\boldsymbol{\vartheta}}_j$ unterscheiden, wird der Wert der Statistik als Kriterium für eine mögliche Erweiterung der lokalen Umgebung verwendet. Skaliert durch einen weiteren Parameter λ wird dann $\mathbf{s}_{ij} = \lambda^{-1} T_{ij}$ als Strafterm eingesetzt, mit dem Änderungen der Bandbreite im iterativen Verfahren umso stärker bestraft werden, je größer der durch T_{ij} bzw. $\mathbf{s}_{ij}$ quantifizierte Unterschied zwischen den lokalen Modellen ist. Als Einstellung für den Skalierungsparameter wird $\lambda = 6.76$ empfohlen, wenn, wie in dieser Arbeit, als Regressionsfunktionen Polynome zweiten Grades betrachtet werden.

Außer diesem Strafterm wird ein weiterer zur Reduzierung des Einflusses von sogenannten Hebelpunkten eingesetzt. Eine Beobachtung an der Stelle x_j, die einen großen Abstand zu x_i besitzt, kann insbesondere bei polynomialen Regressionsmodellen einen bedeutenden Einfluss auf $\hat{\boldsymbol{\vartheta}}_i$ haben. Bestimmt wird dieser durch

$$\gamma_{ij} \;=\; (trW_i)\Psi_j' \left(\Psi W_i \Psi'\right)^{-1} \Psi_j, \tag{2.12}$$

das die relative Differenz berechnet zwischen einem ursprünglichen lokalen Modell und einem, das durch zusätzliches Einbeziehen der Stelle x_j erweitert ist. Je größer γ_{ij} desto größer ist die Hebelwirkung der Stelle x_j für das Modell $\boldsymbol{\vartheta}_i$. Entsprechend werden kleine Gewichte zugeteilt, selbst wenn die lokalen Modelle $\hat{\boldsymbol{\vartheta}}_i$ und $\hat{\boldsymbol{\vartheta}}_j$ ähnlich sind und folglich $\mathbf{s}_{ij}$ klein ist. Der tatsächlich verwendete Strafterm ist

$$\mathbf{e}_{ij} \;=\; \frac{1}{\tau} \max\{0, \left(\frac{\gamma_{ij}}{\gamma_{ij}^*} - 1\right)\}, \tag{2.13}$$

wobei γ_{ij}^* wie in (2.12) bestimmt wird, jedoch unter der Annahme, dass $\hat{\boldsymbol{\vartheta}}_i$ bei Lokationsänderung unveränderlich bleibt. In diesem Fall ist die Matrix der Gewichte W_i durch W_i^* zu ersetzen, in der die Einträge unter Verwendung von $\mathbf{l}_i$ der vorigen Iteration berechnet werden (s. u.). Weiterhin ist τ ein Parameter, mit dem der Grad der Glättung beeinflusst werden kann. Für quadratische Modelle, wie sie in den folgenden Kapiteln eingesetzt werden, wird $\tau = 13.5$ empfohlen.

Die Wahl der Gewichte $w_{ij}^{(k)}$ in einem Iterationsschritt k wird damit durch die jeweiligen Strafterme $\mathbf{l}_{ij}^{(k)}$, $\mathbf{s}_{ij}^{(k)}$, $\mathbf{e}_{ij}^{(k)}$ gesteuert. Sie gehen unabhängig voneinander bei der Bestimmung des Gewichts ein, das zunächst provisorisch berechnet wird durch

$$\tilde{w}_{ij}^{(k)} \;=\; K_l(\mathbf{l}_{ij}^{(k)}) K_s(\mathbf{s}_{ij}^{(k)}) K_e(\mathbf{e}_{ij}^{(k)}). \tag{2.14}$$

Als Kernfunktionen werden nicht steigende Funktionen verwendet mit $K_l(0) = K_s(0) = K_e(0) = 1$. Vorgeschlagen werden dazu $K_l(u)$ wie oben beschrieben und $K_s(u) = K_e(u) = e^{-u}\mathbb{I}_{[0,6]}(u)$. Anschließend wird das neue Gewicht $w_{ij}^{(k)}$ durch Linearkombination aus dem des vorigen Iterationsschrittes $k-1$ und dem provisorischen Gewicht gebildet durch

$$w_{ij}^{(k)} = \eta w_{ij}^{(k-1)} + (1-\eta)\tilde{w}_{ij}^{(k)}.$$

Mit dem Parameter $\eta \in (0,1)$ kann die Stabilität der gesamten Prozedur kontrolliert werden, als Standard wird hierfür $\eta = 0.5$ verwendet.

Zusammenfassung: Mit den obigen Angaben kann das vollständige Verfahren nun formal wie folgt beschrieben werden:

(1) Anfang. Für jeden Punkt x_i, $i = 1, \ldots, n$, wird eine Diagonalmatrix $W_i^{(0)}$ mit Einträgen $w_{ij}^{(0)} = K_l(\mathbf{l}_{ij}^{(0)}$, $\mathbf{l}_{ij}^{(0)} = ((x_i - x_j)/h^{(0)})^2$ aufgestellt. Damit werden lokale (Anfangs-)Schätzer für jede Stelle bestimmt durch $\hat{\boldsymbol{\vartheta}}_i = (\Psi W_i^{(0)} \Psi')^{-1} \Psi W_i^{(0)} \mathbf{y}$. Als Zähler für die Anzahl der Iterationen wird $k = 1$ gesetzt.

(2) Iteration. Die folgenden Schritte werden für alle $i = 1, \ldots, n$ durchlaufen: Jeder Iterationsschritt besteht aus einer adaptiv erfolgenden Änderung der Matrizen W_i für die Gewichte und der darauffolgenden Berechnung neuer Schätzer $\hat{\boldsymbol{\vartheta}}_i$. Zunächst werden mit der Annahme, dass das Modell bei Lokationsänderung unverändert bleibt, Matrizen $W_i^{(k-1)*} = \mathrm{diag}\{K_l(\mathbf{l}_{i1}^{(k-1)}), \ldots, K_l(\mathbf{l}_{in}^{(k-1)})\}$ erstellt. Durch Einsetzen von $W_i^{(k-1)*}$ in (2.10), (2.11) und (2.13) werden dann für jeden Punkt x_j, $j = 1, \ldots, n \neq i$, aktualisierte Strafterme

$$\begin{aligned}
\mathbf{l}_{ij}^{(k)} &= \left(\frac{x_i - x_j}{h^{(k)}}\right)^2, \\
\mathbf{s}_{ij}^{(k)} &= \frac{1}{2\hat{\sigma}^2\lambda}\left(\hat{\boldsymbol{\vartheta}}_i^{(k-1)} - \hat{\boldsymbol{\vartheta}}_j^{(k-1)}\right)' \Psi W_i^{(k-1)} \Psi' \left(\hat{\boldsymbol{\vartheta}}_i^{(k-1)} - \hat{\boldsymbol{\vartheta}}_j^{(k-1)}\right), \qquad (2.15) \\
\mathbf{e}_{ij}^{(k)} &= \frac{1}{\tau}\max\{0, \left(\frac{\gamma_{ij}^{(k)}}{\gamma_{ij}^{(k)*}} - 1\right)\}
\end{aligned}$$

berechnet, mit denen die provisorischen Gewichte $\tilde{w}_{ij}^{(k)}$ entsprechend (2.14) bestimmt werden, die dann die Elemente der Diagonalmatrix $\tilde{W}_i^{(k)}$ bilden.

Anschließend wird das Modell neu geschätzt, indem mit $W_i^{(k)} = \eta W_i^{(k-1)} + (1-\eta)\tilde{W}_i^{(k-1)}$ eine neue Matrix für die Gewichte gebildet wird, mit der gemäß (2.9) eine aktualisierte Modellschätzung an der Stelle x_i durch

$$\hat{\boldsymbol{\vartheta}}_i^{(k)} = \left(\Psi W_i^{(k)} \Psi'\right)^{-1} \Psi W_i^{(k)} \mathbf{y} \qquad (2.16)$$

erfolgt. Das geschätzte Signal erhält man als $\hat{f}^{(k)}(x_i) = \Psi_i' \hat{\boldsymbol{\vartheta}}_i^{(k)}$.

(3) Abbruch. Nach Durchlauf einer Iteration wird die Zählvariable k um 1 erhöht. Die Bandbreite wird um den Faktor a, für den die Einstellung $a = 1.25$ empfohlen wird, vergrößert und auf $h^{(k)} = ah^{(k-1)}$ heraufgesetzt. Überschreitet $h^{(k)}$ eine vorgegebene maximale Bandbreite $h_{\max}$, so wird die Prozedur abgebrochen, andernfalls folgt eine weitere Iteration. Im Gegensatz zu den übrigen frei wählbaren Parametern kann für $h_{\max}$ keine allgemeingültige Empfehlung gegeben werden. Stattdessen sollte die Wahl kontextbezogen erfolgen.

2.4 Straffe Saite (Taut String)

Im Gegensatz zu den oben vorgestellten Verfahren, in denen die Existenz einer Regressionsgleichung bzw. einer „wahren“ Verteilung der Zielvariablen angenommen wird, basiert das Verfahren der Straffen Saite (*taut string*) auf dem von Davies (1995, 2003) entwickelten Konzept der Datenapproximation. Dieses beurteilt mögliche Modelle allein im Hinblick auf ihre datenbezogene Adäquatheit. Ein statistisches Modell wird dann als geeignet angesehen, wenn es bestimmte, durch den Kontext motivierte, wichtige Eigenschaften des Datensatzes, sogenannte *data features*, aufweist. Innerhalb der Menge aller betrachteten adäquaten Datenapproximationen kann dann mittels weiterer Regularisierungsbedingungen ein bestimmtes Modell ausgewählt werden. Hierbei werden typischerweise Kriterien eingesetzt, die die Komplexität bzw. Einfachheit des Modells bewerten.

Die Adäquatheit eines nichtparametrischen Regressionsmodells mit geschätzter Signalfunktion $\hat{f}(.)$ wird anhand der Residuen $r_i = y_i - \hat{f}(x_i)$, $i = 1, \ldots, n$, beurteilt. Können diese als Realisierung einer unabhängig und identisch normalverteilten Zufallsvariable mit konstanter Varianz σ^2 angesehen werden (Weißes Rauschen, vgl. S. 31), so kann das Modell adäquat genannt werden. Das von Davies und Kovac (2001) vorgestellte Verfahren der Straffen Saite überprüft die Residuen mit Hilfe der Multiresolutionsbedingungen. Diese betrachten für alle Teilintervalle $I \subset \{1, \ldots, n\}$ die betragsmäßige Summe der Residuen $r_i, i \in I$, und überprüfen, ob für eine gegebene Stichprobengröße n

$$\max_{i \in I} \frac{1}{\sqrt{|I|}} \left| \sum_{x_i \in I} y_i - \hat{f}(x_i) \right| \leq \sigma \sqrt{2.5 \log n}, \tag{2.17}$$

erfüllt ist. Bei der Anwendung für ein jeweils gegebenes Modell wird der unbekannte Streuungsparameter σ in (2.17) geschätzt, wofür beispielsweise

$$\hat{\sigma} = \frac{1.48}{\sqrt{2}} \text{median}\{|y_2 - y_1|, \ldots, |y_n - y_{n-1}|\} \tag{2.18}$$

vorgeschlagen wird.

Zur Bestimmung von $\hat{f}(t)$ wird zunächst der integrierte Prozess

$$y_i^\circ = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad i = 1, \ldots, n, \tag{2.19}$$

betrachtet. Um diesen wird ein durch untere und obere Grenzen l_i und u_i definierter Schlauch gelegt mit

$$l_i = y_i^\circ - \frac{c}{\sqrt{n}}, \qquad u_i = y_i^\circ + \frac{c}{\sqrt{n}}. \tag{2.20}$$

Bezeichne S_n° dann eine Funktion, die innerhalb des Schlauchs liegt, d. h. $l_i \leq S_n^\circ(x_i) \leq u_i$, und die am Rand des betrachteten Bereichs durch $S_n(0) = 0$ und $S_n(x_n) = y_n^\circ$ definiert ist. Unter allen Funktionen mit diesen Eigenschaften wird dasjenige $S_n^\circ(x)$ mit minimaler Weglänge $\int_0^1 \sqrt{1 + S_n^\circ(x/n)^2}dx$ als Straffe Saite bezeichnet. Anschaulich entspricht es einer zwischen den Punkten $(0, 0)$ und (n, y_n°) straff gespannten Schnur, die innerhalb des Schlauchs liegt. Ihre Ableitung $s_n^\circ(x)$ ist eine stückweise konstante Funktion, die dann als adäquate Schätzung von f angesehen wird, wenn die Residuen $r_i = y_i - s_n^\circ(x_i)$ die Multiresolutionsbedingungen (2.17) erfüllen. Die Komplexität der Resultate bei verschiedenen Schlauchbreiten wird durch die Anzahl der Extrema von $s_n^\circ(x)$ bewertet. Um unter den adäquaten Schätzungen eine in dieser Hinsicht einfachste, d. h. eine mit den wenigsten Extrema zu erhalten, wird die Straffe Saite ausgehend von großen Werten für die Konstante c in (2.20) für schmaler werdende Schläuche bestimmt. Der breiteste Schauch, für den eine adäquate Schätzung $s_n^\circ(x)$ resultiert, gilt damit als das am wenigsten komplexe Ergebnis $\hat{f}(x)$ des Verfahrens.

Die Straffe Saite liefert eine unstetige Funktion und ist damit nicht in Situationen einsetzbar, in denen Stetigkeit oder Differenzierbarkeit der Schätzung verlangt werden. Majidi (2003) stellt eine Möglichkeit vor, eine glatte Schätzung für $f(t)$ zu erhalten, deren integrierter Prozess Element desselben Schlauchs wie der der Straffen Saite ist. Die Glattheit einer mindestens zweifach differenzierbaren Funktion $g(x)$ auf dem Intervall $[0, 1]$ wird dabei durch

$$\|g(x)\| = \sqrt{\int_0^1 (g^{(2)}(x))^2 dx}$$

definiert. Dann kann gezeigt werden, dass durch eine kubische Splinefunktion eine adäquate und eindeutige Schätzung von $f(x)$ möglich ist, die durch Lösung eines quadratischen Optimierungsproblems berechnet werden kann. Das Ergebnis wird hier im weiteren Verlauf als glatte Straffe Saite (STS, *smooth taut string*) bezeichnet.

2.5 Singular Spectrum Analysis

Wie die Verfahren der Datenapproximation verzichtet *singular spectrum analysis* (SSA) auf Annahmen über die Zufälligkeit vorliegender Daten. Damit ist diese Methode ebenfalls modellfrei, kann jedoch in ihrer Anwendung auf Zeitreihen nicht als eigentliches Regressionsverfahren angesehen werden. Sie ermöglicht die Zerlegung einer Zeitreihe in mehrere additive, zueinander möglichst unähnliche Komponenten, die somit als eigenständige Zeitreihen mit bestimmten charakteristischen Eigenschaften interpretierbar sind. Hierzu zählen z. B. die in dieser Arbeit interessierenden Merkmale von Trend, periodischen Schwankungen und zufälligem Rauschen.

SSA kann als besondere Anwendung der erstmals von Pearson (1901) und Hotelling (1933) vorgestellten Hauptkomponentenanalyse für Zeitreihen angesehen werden. Während sie in im engeren Sinne statistischen Arbeiten bislang wenig benutzt wurde, ist sie in Gebieten wie Signalverarbeitung oder Meteorologie und Klimaforschung ein häufig verwendetes Werkzeug. Beispielsweise seien hier Vautard et al. (1992), Allen und Smith (1996) oder Ghil et al. (2002) als sehr beachtete Veröffentlichungen genannt. Die erstmalige Entwicklung und Präsentation von SSA findet sich in Broomhead und King (1986) und Broomhead et al. (1987).

Mit SSA kann in vier aufeinanderfolgenden Schritten eine Zeitreihe mit äquidistanten Beobachtungen zerlegt und dann neu zusammengesetzt werden. Zuerst wird in einem Einbettung genannten Schritt die Zeitreihe durch eine Übergangsmatrix dargestellt, die alle Teilausschnitte mit vorher gewählter Fensterbreite in den Spalten enthält. Anschließend werden durch Singulärwertzerlegung aus dem Produkt dieser Matrix mit ihrer Transponierten die zugehörigen Eigenwerte und -vektoren bestimmt. Mit dieser Zerlegung wird die Zeitreihe in zwei folgenden Schritten neu rekonstruiert: Im Gruppierungsschritt werden aus den Elementarmatrizen, die man durch die Zerlegung erhält, solche mit ähnlichen Eigenschaften zusammengefasst. Im letzten Schritt werden dann aus den so neu gruppierten Matrizen durch diagonale Mittelbildung neue Übergangsmatrizen und damit neue Zeitreihen gewonnen. Die folgende Darstellung der Einzelschritte orientiert sich an der Beschreibung von Golyandina et al. (2001).

(1) Einbettung. Die Beobachtungen $\mathbf{y} = (y_1, \ldots, y_n)^\top$ werden in Teilzeitreihen der Länge $1 < L < n$ zerlegt und in einer Übergangsmatrix $\mathbf{Y}$ zusammengefasst. Der Parameter L definiert die Fensterbreite und damit auch die Anzahl

$K = n - L + 1$ der Teilzeitreihen. Somit ist die (L, K)-Übergangsmatrix

$$\mathbf{Y} = \mathbf{Y}(L) = \begin{pmatrix} y_1 & y_2 & y_3 & \cdots & y_{n-L+1} \\ y_2 & y_3 & y_4 & \cdots & y_{n-L+2} \\ y_3 & y_4 & y_5 & \cdots & y_{n-L+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_n \end{pmatrix}, \tag{2.21}$$

in Spaltenschreibweise auch $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_{n-L+1})$. Wegen ihrer Eigenschaft, dass auf allen Diagonalen von unten links nach oben rechts dieselben Einträge stehen, handelt es sich um eine Hankel-Matrix.

(2) Singulärwertzerlegung. In diesem Schritt wird die Übergangsmatrix $\mathbf{Y}$ additiv in Elementarmatrizen mit Rang 1 zerlegt. Dazu wird zunächst ausgenutzt, dass für jede Matrix $\mathbf{M}$ eine Singulärwertzerlegung existiert, die als Produkt von zwei orthonormalen Matrizen $\mathbf{U}$ und $\mathbf{V}$ sowie einer Diagonalmatrix $\mathbf{D}$ darstellbar ist als $\mathbf{M} = \mathbf{UDV}$. Für $\mathbf{U}$ und $\mathbf{V}$ gilt, dass sie die aus den Eigenvektoren von $\mathbf{MM}^\top$ bzw. $\mathbf{M}^\top\mathbf{M}$ gebildeten Matrizen sind. Die Spalten von $\mathbf{U}$ werden demnach auch als linke, die Spalten von $\mathbf{V}$ als rechte Eigenvektoren bezeichnet.

Für eine gegebene Übergangsmatrix $\mathbf{Y}$ werden zunächst die Eigenwerte von $\mathbf{S} = \mathbf{YY}^\top$ berechnet, der Größe nach absteigend sortiert und mit $\lambda_1, \ldots, \lambda_L$ bezeichnet. Sie sind nicht notwendigerweise alle größer als 0. Weil allein die echt postiven Eigenwerte im weiteren Verlauf relevant sind, werden nur $\lambda_1, \ldots, \lambda_d$ weiterhin berücksichtigt, wobei $d = \max\{i = 1, \ldots, L | \lambda_i > 0\}$. Als jeweils zugehörige Eigenvektoren werden dann $U_1, \ldots, U_d$ bestimmt. Die Eigenvektoren $V_1, \ldots, V_d$ können damit durch $V_i = \mathbf{Y}^\top U_i / \sqrt{\lambda_i}$, $i = 1, \ldots, d$ berechnet werden. Das Tupel $(\sqrt{\lambda_i}, U_i, V_i)$ wird auch das i-te Eigentripel der Singulärwertzerlegung von $\mathbf{Y}$ genannt. Damit kann $\mathbf{Y}$ auch dargestellt werden als

$$\mathbf{Y} = \mathbf{Z}_1 + \ldots, + \mathbf{Z}_d, \tag{2.22}$$

mit $\mathbf{Z}_i = \sqrt{\lambda_i} U_i V_i^\top$, $i = 1, \ldots, d$. Die $\mathbf{Z}_i$, $i = 1, \ldots, d$ werden wegen ihrer Eigenschaft, den Rang 1 zu besitzen, Elementarmatrizen genannt. Die Darstellung in (2.22) ist genau dann eindeutig, wenn die Eigenwerte verschieden sind.

Eine wichtige Eigenschaft der Singulärwertzerlegung ist, dass die Summe der jeweils ersten $r < d$ Elementarmatrizen im Vergleich zu einer beliebigen Matrix mit Rang r den geringsten Abstand zur Übergangsmatrix $\mathbf{Y}$ aus (2.21) aufweist bezüglich der Matrix-Norm $\|.\|_\mathcal{M}$: Für eine beliebige Matrix $\mathbf{X}$ mit Rang r gilt, dass $\|\mathbf{Y} - \sum_{i=1}^r \mathbf{Z}_i\|_\mathcal{M} \leq \|\mathbf{Y} - \mathbf{X}\|_\mathcal{M}$. Die Norm $\|.\|_\mathcal{M}$ ist für jede Matrix $\mathbf{M}$ definiert durch $\|\mathbf{M}\|_\mathcal{M} = \sqrt{\sum_i \lambda_i^\mathbf{M}}$, wobei mit $\lambda_i^\mathbf{M}$ analog zur obigen Darstellung die

Eigenwerte von $\mathbf{M}\mathbf{M}^\top$ bezeichnet werden. Der Abstand von Elementarmatrizen $\mathbf{Z}_i$ zu $\mathbf{Y}$ ist also umso größer, je kleiner die zugehörigen Eigenwerte sind. Folglich sind bei den folgenden Schritten insbesondere die Matrizen mit den großen Eigenwerten bedeutsam.

(3) Gruppierung. Die Elementarmatrizen aus dem zweiten Schritt werden nun unter bestimmten Gesichtspunkten zusammengefasst. Die Menge der Indizes $I = \{1, \ldots, d\}$ wird dazu in disjunkte, nichtleere Teilmengen $I_1, \ldots, I_m$ zerlegt. Für jede dieser Teilmengen erhält man eine zugehörige Matrix $\mathbf{Z}_{I_k}$, $k = 1, \ldots, m$, die durch Summation der durch die Indexmenge I_k gegebenen Elementarmatrizen berechnet wird als

$$\mathbf{Z}_{I_k} = \sum_{j \in I_k} \mathbf{Z}_j.$$

Folglich gilt

$$\mathbf{Z} = \sum_{k=1}^{m} \mathbf{Z}_{I_k}. \tag{2.23}$$

Der Schritt, in dem die Anzahl m der Indexmengen und ihre Elemente festgelegt werden, wird Gruppierung der Eigentripel genannt. Das grundlegende Ziel hierbei ist, mit einer geeigneten Gruppierung resultierende Matrizen zu erhalten, deren Struktur jeweils möglichst ähnlich zu der einer Hankel-Matrix ist, und die damit als durch einen Einbettungsschritt entstandene Übergangsmatrizen anderer Zeitreihen angesehen werden kann.

(4) Diagonales Mitteln. Auch durch geschickte Gruppierung lässt sich normalerweise nicht erreichen, dass die resultierenden Matrizen $\mathbf{Z}_{I_k}$, $k = 1, \ldots, m$, die Hankel-Struktur aufweisen. Dies kann aber durch die Anwendung des Hankel-Operators $\mathcal{H}$ erreicht werden: Für eine gegebene $(L, K)-$ Matrix $\mathbf{Z}$, $L \leq K$, ist $\tilde{\mathbf{Z}} = \mathcal{H}\mathbf{Z}$ diejenige Hankel-Matrix gleicher Dimension, die zu $\mathbf{Z}$ bezüglich der Matrix-Norm $\|.\|_{\mathcal{M}}$ den geringsten Abstand besitzt. Der Hankel-Operator ersetzt die Einträge auf den Diagonalen von links unten nach rechts oben durch ihr arithmetisches Mittel. Es gilt also für $\tilde{z}_{ij}$, $i = 1, \ldots, L$, $j = 1, \ldots, K$,

$$\tilde{z}_{ij} = \begin{cases} \frac{1}{s} \sum_{l=1}^{s-1} z_{l,s-l} & , \quad 2 \leq s \leq L-1 \\ \frac{1}{L} \sum_{l=1}^{L} z_{l,s-l} & , \quad L \leq s < K+1 \\ \frac{1}{K+L-s+1} \sum_{l=s-K}^{L} z_{l,s-l}\,, & , \quad K+2 \leq s \leq K+L \end{cases} \quad , \; s = i+j. \tag{2.24}$$

In umgekehrter Analogie zum Einbettungsschritt können die in Hankel-Form umgewandelten Matrizen aus dem Gruppierungsschritt, $\tilde{\mathbf{Z}}_{I_1}, \ldots, \tilde{\mathbf{Z}}_{I_m}$, als Darstellungen von Zeitreihen $\tilde{\mathbf{z}}^{(1)}, \ldots, \tilde{\mathbf{z}}^{(m)}$ angesehen werden. Für diese gilt, dass

$$\mathbf{y} = \sum_{k=1}^{m} \tilde{\mathbf{z}}^{(k)} = \left(\sum_{k=1}^{m} \tilde{z}_1^{(k)}, \ldots, \sum_{k=1}^{m} \tilde{z}_n^{(k)} \right). \tag{2.25}$$

Der Vorgang, aus einer Gruppe von Elementarmatrizen durch Addition und Umwandlung in Hankel-Form schließlich eine neue Zeitreihe $\tilde{\mathbf{z}}$ zu erhalten, wird Rekonstruktion genannt.

Wie ersichtlich ist, werden in diesem Verfahren keine modellhaften statistischen Annahmen gemacht. Einfluss auf das Ergebnis haben jedoch der frei wählbare Parameter L im Einbettungsschritt sowie die Art und Weise, nach der die Elementarmatrizen im dritten Schritt gruppiert werden. Allgemein gilt, dass bei umso kleinerer Fensterbreite L und umso mehr gruppierten Elementarmatrizen die Anpassung der daraus rekonstruierten Zeitreihe an die ursprünglichen Daten zunimmt. Im Gegenzug bedeutet dies, dass mit großen Fensterbreiten und wenigen Elementarmatrizen nur allgemeine, weniger variationsreiche Verläufe darstellbar sind.

3 Komponentenmodelle für Zeitreihen mit wöchentlichen Meldefällen

3.1 Klassisches Dekompositionsmodell

Die im letzten Kapitel vorgestellten nichtparametrischen Regressionsverfahren ermöglichen die Schätzung einer unbekannten Funktion f aus Beobachtungen $y_1, \ldots, y_n$ an Stellen $t = 1, \ldots, n$. Insbesondere bei Zeitreihen wie denen der Meldehäufigkeiten von Infektionskrankheiten ist ein genaueres Verständnis des Verlaufs notwendig. Das klassische Dekompositionsmodell, vgl. z. B. Brockwell und Davis (2002), ermöglicht dazu einen ersten geeigneten Ansatz. Es betrachtet eine Zeitreihe z_t, $t = 1, \ldots, n$, kurz $\{z_t\}$, als Beobachtung eines stochastischen Prozesses $\{Z_t\}$, $t \in \mathbb{Z}$, so dass die zeitlich geordneten Daten z_t damit als Realisierungen von Zufallsvariablen Z_t angesehen werden. Durch

$$Z_t = m(t) + s(t) + \varepsilon_t, \qquad t = 1, \ldots, n, \tag{3.1}$$

werden sie als Summe zweier Funktionen $m(t)$ und $s(t)$, die als Trend- und Saisonkomponente bezeichnet werden, und einer zufälligen Fehlerkomponente ε_t dargestellt. $m(t)$ und $s(t)$ sind deterministische Funktionen, während ε_t eine Zufallsvariable ist mit Erwartungswert $\mathrm{E}\,\varepsilon_t = 0$ und Varianz $\mathrm{Var}\,\varepsilon_t = \sigma^2$ für alle $t = 1, \ldots, n$. Dabei ist die Unabhängigkeit der ε_t nicht vorausgesetzt, jedoch besitzen $\varepsilon_1, \ldots, \varepsilon_n$ gemeinsam eine stationäre Verteilung (s. u.). Die einzelnen Komponenten werden im folgenden genauer charakterisiert. Fasst man die Summe von Trend und Saison zu einer Komponente $f(t) = m(t) + s(t)$ zusammen, wird die Analogie zum nichtparametrischen Regressionsmodell deutlich, in dem Beobachtungen durch ein deterministisches Signal $f(t)$ und einen zufälligen Fehler ε_t erklärt werden.

Trend

Obwohl der Begriff Trend seit langem in Arbeiten zur Zeitreihenanalyse verwendet wird, existiert hierzu keine eindeutige und genaue Definition. Im Allgemeinen erfolgt lediglich die vage Erklärung, dass die Trendkomponente eine glatte, sich im Vergleich zur beobachteten Zeitreihe langsam ändernde Funktion ist, die langfristige Änderungen beschreibt. Diese modellunabhängige Idee findet sich z. B. bei Brillinger (1975), Kendall und Ord (1990), Brockwell und Davis (2002) oder Fan und Yao (2003), und auf sie wird im weiteren Verlauf der Arbeit Bezug genommen. Für Anderson (1971) kann die Trendkomponente je nach Modellansatz entweder eine glatte, sich nicht regelmäßig über die Zeit ändernde Funktion oder im Gegensatz dazu auch eine periodische Funktion sein. Abweichend von diesen Eigenschaften wird die Trendkomponente gelegentlich auch konkret als Erwartungswert der Zeitreihe $\mathrm{E}\, Y_t$ an einem Punkt $t = 1, \ldots, n$ definiert wie z. B. durch Diggle (1990). Da in den letzten beiden Fällen keine Unterscheidung von periodischen Saisoneffekten und nicht periodischem Trendverlauf erfolgt, werden sie nicht dem klassischen Dekompositionsmodell zugeordnet.

Um bei der späteren Auswahl unter verschiedenen Trendkomponentenschätzungen nicht willkürlich ein bestimmtes Resultat zu bevorzugen, und weil in der Literatur keine präzisen mathematischen bzw. statistischen Definitionen verwendet werden, die als allgemein verbindlich erachtet werden können, wird für diese Arbeit das folgende Kriterium formuliert: Unter möglichen Trendschätzungen, die die oben genannten Eigenschaften besitzen und damit mögliche Kandidaten darstellen, ist diejenige als Schätzung $\hat{m}(t)$ zu wählen, die die geringste Krümmung aufweist, d. h. die

$$\gamma(\hat{m}(t)) \;=\; \frac{1}{n-2} \sum_{t=2}^{n-1} \left(\hat{m}(t-1) - 2\hat{m}(t) + \hat{m}(t+1)\right)^2 \tag{3.2}$$

minimiert. Dieser Ansatz folgt der Idee, die Glattheit differenzierbarer Funktionen über ihre zweite Ableitung zu definieren. Da die Glattheit zu- und die Krümmung abnimmt, je langsamer bzw. geringer variierend eine geschätzte Funktion $\hat{m}(t)$ ist, ist dies als Kriterium für die Auswahl der Trendkomponente geeignet.

Saison

Eine Zeitreihe besitzt eine periodische Komponente, wenn ihr Verlauf in aufeinander folgenden Intervallen ähnlich ist, vgl. z. B. Box et al. (1994). Dieses Verhalten kann durch eine Saisonkomponente abgebildet werden in Form einer periodischen Funktion, deren Werte sich in festen Abständen d wiederholen, d. h. $s_t = s_{t+d}$, für alle $t \in \mathbb{Z}$. Typischerweise werden dadurch jahreszeitlich bedingte Effekte wie Monats- oder Quartalseffekte beschrieben, wenn die Daten einer

Zeitreihe über mehrere Jahre erhoben wurden. Treten mehrere regelmäßige Effekte mit unterschiedlichen Periodenlängen z. B. d_1 und d_2 auf, so können sie durch eine Saisonkomponente abgebildet werden, deren Länge das kleinste gemeinsame Vielfache von d_1 und d_2 beträgt. Bei Zeitreihen mit einer jährlichen Saisonstruktur darf die Periodenlänge nicht größer als ein Jahr sein, weil ansonsten keine jährliche Periodizität erreicht wird.

Ergebnisse verschiedener Verfahren, die eine jährliche Periodizität aufweisen, können als Kandidatenfunktionen für die Saisonkomponente angesehen werden. Als geschätzte Saison wird dasjenige Resultat verwendet, das im Vergleich mit allen anderen Kandidaten die meiste Variation erklärt, d. h. die geringste quadratische Abweichung

$$RSS(\hat{s}(t)) = \sum_{t=1}^{n} (y_t - \hat{m}(t) - \hat{s}(t))^2 \tag{3.3}$$

zu den trendbereinigten Beobachtungen $y_t - \hat{m}(t)$ besitzt.

Fehler

Das Dekompositionsmodell zerlegt die Beobachtungen in eine deterministische Komponente $f(t) = m(t) + s(t)$, die durch die Summe von Trend- und Saisonkomponente gebildet wird. Die Darstellung $Y_t = f(t) + \varepsilon_t$, $t = 1, \ldots, n$, gleicht dem nichtparametrischen Regressionsmodell. Während dort allerdings für den Fehler in der Regel unabhängig und identisch verteilte Störgrößen angenommen werden, wird im Zeitreihenmodell die weniger strenge Anforderung der Stationarität an die Fehler gestellt.

Ein Prozess $\{Z_t\}$ wird stationär (stationär im weiteren Sinne, schwach stationär oder stationär zweiter Ordnung) genannt, wenn die folgenden Bedingungen erfüllt sind, vgl. z. B. Brockwell und Davis (1991):

$$\begin{aligned} \mathrm{E}\,|Z_t|^2 &< \infty, \qquad \forall t \in \mathbb{Z} \\ \mathrm{E}\,Z_t &= c, \qquad \forall t \in \mathbb{Z} \\ \gamma_Z(r,s) &= \gamma_Z(t+r, t+s), \qquad \forall r,s,t \in \mathbb{Z}. \end{aligned} \tag{3.4}$$

Dabei bezeichnet $\gamma_Z(r,s) = \gamma_Z(r-s,0) = \gamma_Z(r-s)$ die Autokovarianzfunktion der Zufallsvariablen Z_r und Z_s mit

$$\gamma_Z(r-s) = \mathrm{Cov}(Z_{r-s}, Z_0), \qquad r,s \in \mathbb{Z}, \tag{3.5}$$

die also nur vom zeitlichen Abstand $s-r$ zweier Zeitpunkte und nicht von ihrer Position t abhängt. Für $h = r - s$ wird statt $\gamma_Z(h)$ häufig auch die Auokorrelationsfunktion (*autocorrelation function*, acf)

$$\rho_Z(h) = \frac{\gamma_Z(h)}{\gamma_Z(0)}, \qquad h \in \mathbb{Z}, \tag{3.6}$$

verwendet, wobei wegen der Symmetrie $\rho_Z(h) = \rho_Z(-h)$ meist nur $h \geq 0$ betrachtet wird. Die Autokorrelationsfunktion ist ein wichtiges Hilfsmittel zur Charakterisierung der Abhängigkeitsstruktur von stochastischen Prozessen und Zeitreihen.

Ein einfaches Beispiel für einen stationären Prozess ist das Weiße Rauschen (*white noise*), $\{X_t\} \sim WN(0, \sigma^2)$. Die Zufallsvariablen X_t sind dabei identisch verteilt und unkorreliert, aber nicht notwendigerweise unabhängig, und besitzen konstanten Ewartungswert $\mathrm{E}\, X_t = 0$ und gleiche Varianz $\mathrm{Var}\, X_t = \sigma^2$. Folglich gilt für die Autokorrelationsfunktion $\gamma_Z(h) = \sigma^2 \mathbb{1}_{\{0\}}(h)$, $h \in \mathbb{Z}$.

Weitere typische Beispiele sind gleitende Durchschnitte (moving average, MA) und autoregressive (AR) Prozesse: $\{Z_t\}$ ist ein moving average Prozess MA(q) der Ordnung $q > 0$, falls

$$Z_t \;=\; X_t + \theta_1 X_{t-1} + \ldots + \theta_q X_{t-q}, \qquad t \in \mathbb{Z}, \tag{3.7}$$

wobei $\{X_t\} \sim WN(0, \sigma^2)$ und $\theta_1, \ldots, \theta_q$ Konstanten sind. Die Autokorrelationsfunktion ist in diesem Fall

$$\rho_Z(h) \;=\; \sum_{i=0}^{q-|h|} \theta_i \theta_{i+|h|} \mathbb{1}_{[-q,q]}(h), \qquad h \in \mathbb{Z}, \tag{3.8}$$

mit $\theta_0 = 1$ und $\theta_i = 0$ für $i > q$. Dadurch, dass die Zufallsvariable Z_t ein gewichtetes Mittel aus den zufälligen, jeweils unabhängigen Einflüssen X_t der letzten $t = 1, \ldots, q$ Zeitpunkte ist, verschwindet die Autokorrelationsfunktion für Abstände $|h| > q$.

Ein autoregressiver Prozess AR(p) der Ordnung $p > 0$ hingegen ist direkt von den vorherigen Zufallsvariablen $Z_1, \ldots, Z_p$ abhängig und wird definiert als

$$Z_t \;=\; X_t + \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p}, \qquad t \in \mathbb{Z}, \tag{3.9}$$

mit $\{X_t\} \sim WN(0, \sigma^2)$ und Konstanten $\phi_1, \ldots, \phi_p$. Der Prozess ist stationär, wenn keine Lösung der Gleichung $1 - \phi_1 z - \ldots - \phi_p z^p = 0$ existiert für eine komplexe Zahl z mit $|z| = 1$. Für einen AR(p)-Prozess ist die Autokorrelationsfunktion gegeben durch Lösung der Yule-Walker-Gleichungen

$$\begin{aligned} \rho_Z(0) &= 1 \\ \rho_Z(h) &= \sum_{i=1}^{p} \phi_i \rho_Z(h-i) \\ \rho_Z(h) &= \rho_Z(-h). \end{aligned}$$

Die charakteristische Eigenschaft von $\rho_Z(h)$ ist, dass ihre Werte für größer werdendes h mit geometrischer Geschwindigkeit abnehmen.

Prozesse, die sowohl über eine AR- wie über eine MA-Komponente verfügen, bilden die Klasse der ARMA-Prozesse. $\{Z_t\}$ folgt einem ARMA(p,q)-Prozess, wenn die Darstellung

$$Z_t - \phi_1 Z_{t-1} - \ldots - \phi_p Z_{t-p} \;=\; X_t + \theta_1 X_{t-1} + \ldots + \theta_q X_{t-q}, \qquad \forall t \in \mathbb{Z}, \tag{3.10}$$

gilt, wobei $\{X_t\}$ Weißes Rauschen ist und $\phi_1, \ldots, \phi_p$ und $\theta_1, \ldots, \theta_q$ Konstanten sind, durch die die jeweiligen AR- und MA-Anteile bestimmt werden. Die oben aufgeführten MA- und AR-Prozesse können also als Spezialfälle von ARMA-Prozessen angesehen werden.

3.2 Erweitertes Dekompositionsmodell

Für die Analyse von Meldefallhäufigkeiten ist es sinnvoll, das einfache Dekompositionsmodell in (3.1) zu erweitern. Denn nach Schätzung einer deterministischen Trend- und Saisonkomponente können die verbleibenden Residuen hier noch nicht als stationäre Zeitreihe angesehen werden. Das klassische Dekompositionsmodell ist damit nicht ausreichend. Mögliche Gründe sind einerseits, dass neben Trend- und Saisoneffekten auch weitere lang- und kurzfristig wirkende, strukturierte Einflüsse existieren können, die den Verlauf der Zeitreihe bestimmen. Daneben können auch einzelne Meldewochen aufgrund ihrer besonderen Position im Jahr einen zusätzlichen Einfluss auf die Zahl der gemeldeten Fälle haben. Hierzu zählen Wochen, die Feiertage enthalten oder sich im Zeitraum von Schulferien oder typischen Urlaubszeiten befinden. Aus diesen Gründen ist die Erweiterung des klassischen Dekompositionsmodells um eine hier sogenannte zyklische Komponente c_t und eine Kalenderkomponente k_t sinnvoll. Damit ist das erweiterte oder verallgemeinerte Dekompositionsmodell gegeben durch

$$Z_t \;=\; m(t) + s(t) + c(t) + k(t) + \varepsilon_t, \qquad t = 1, \ldots, n. \tag{3.11}$$

Zyklische Komponente

Es ist möglich, dass außer Trend- und Saisoneffekten wie oben beschrieben weitere systematische Effekte den Verlauf der Meldefälle beeinflussen. Denkbar sind einerseits periodische Effekte mit niedriger Frequenz, so dass die Periodenlänge größer als ein Jahr ist. Andererseits sind auch kurzfristig auftretende Effekte möglich, die eine Änderung des durch Trend- und Saisoneffekt gegebenen Erwartungswertes bewirken. Damit beschreibt die zyklische Komponente die

Differenz zwischen dem Erwartungswert der Zeitreihe und der Summe der Komponenten für Trend und Saison sowie Kalendereffekten. Sie kann als Funktion interpretiert werden, die im Gegensatz zum Fehler nicht den zufälligen, sondern einen deterministischen Rest beschreibt, also alle übrigen systematischen Einflüsse, die inhaltlich nicht durch Trend- oder Saisonkomponente erklärt werden. Außer ggf. Stetigkeit und Differenzierbarkeit sind damit keine weiteren Bedingungen an die Form der zyklischen Komponente gegeben. Ziel ist es damit, eine einfache wie auch adäquate Schätzung des in den trend- und saisonbereinigten Daten verbliebenen Signals zu finden.

Diese Anforderungen können unter Verwendung des Konzepts der Datenapproximation formuliert werden. Davies (2003) nennt ein Modell eine adäquate Approximation, wenn ein durch das Modell erzeugter Datensatz wie die tatsächlichen Beobachtungen „aussieht“. Dies kann durch zuvor festgelegte *data features* (Davies, 1995) überprüft werden. Ein data feature stellt ein sachwissenschaftlich motiviertes Kriterium dar, anhand dessen bestimmte Eigenschaften eines statistischen Modells bewertet werden. Die Multiresolutionsbedingungen, mit denen der Taut String-Schätzer bestimmt wird, vgl. Abschnitt 2.4, sind dafür ein Beispiel: Die Taut String-bereinigten Daten bzw. Residuen, die die Multiresolutionsbedingungen erfüllen, können als unabhängige und normalverteilte Fehler angesehen werden. Das durch den Taut String gegebene Modell sieht somit aus wie die Daten und wird daher als adäquate Approximation hinsichtlich dieses data features bezeichnet. Ein geeignetes data feature zur Beurteilung einer geschätzten zyklischen Komponente ist die Bedingung, dass die um diese Schätzung bereinigten Daten eine stationäre Verteilung besitzen. Die Überprüfung des data features für eine jeweilige Schätzung muss durch eine geeignete Operationalisierung ermöglicht werden. Dies wird später in den Abschnitten 4.5 und 4.7 erläutert.

Adäquate Approximationen stellen geeignete Kandidatenfunktionen für die zyklische Komponente dar. Unter diesen wird diejenige als die Schätzung ermittelt, die ein vorgegebenes Maß für die Komplexität minimiert. Im Beispiel des Taut Strings wird die Komplexität über die Anzahl der Extrema definiert: Das Schätzergebnis ist also diejenige Funktion mit den wenigsten Extrema, die adäquat ist bzgl. der Multiresolutionsbedingungen. Da für die zyklische Komponente keine inhaltlich eindeutige Anforderung existiert, wird als Kriterium zur Beurteilung der Komplexität die totale Variation gewählt. In diesem Sinn stellt die geschätzte zyklische Komponente die einfachste Schätzung für die nicht durch Trend und Saison modellierte Variation des Erwartungswertes dar.

Kalenderkomponente

Die Anzahl der Fälle, die in Wochen mit besonderen kalendarischen Eigenschaften wie Feiertagen, Ferien etc. gemeldet werden, kann vom durch Trend-, Saison- und zyklischer Komponente bestimmten Erwartungswert abweichen. Als Ursachen sind Unterschiede im Meldeverhalten oder in einer gegenüber anderen Wochen veränderten Funktionsweise des Meldesystems möglich, wie sie z. B. in Abschnitt 1.2 genannt werden. Demnach sollten derartige Kalendereffekte bei der Analyse saisonaler Daten durch eine gesonderte Komponente berücksichtigt werden.

Harvey et al. (1997) und Cleveland und Scott (2007) stellen Methoden vor, mit denen kalendarische Effekte in Zeitreihenmodellen mit wöchentlichen Daten bei der Schätzung des saisonalen Verhaltens berücksichtigt werden. Für eine explizite Unterscheidung zwischen saison- und kalenderbedingten Effekten, wie sie mit einem Dekompositionsmodell erfolgt, sind sie jedoch nicht geeignet.

In dieser Arbeit wird daher ein eigener Ansatz verwendet, in dem grundsätzlich davon ausgegangen wird, dass der Effekt einer bestimmten Woche mit kalendarischer Besonderheit über die Jahre konstant bleibt, weil Änderungen von Jahr zu Jahr bereits durch Trend- und zyklische Komponente erklärt werden können. Zwischen verschiedenen Feiertagen bzw. Wochen mit Ferien ist dagegen von unterschiedlich großen Effekten auszugehen. Für eine in der Weihnachtszeit liegende Woche ist eine größere Wirkung auf die Meldezahlen denkbar als für eine beliebige Woche mitten im Jahr mit nur einem Feiertag wie z. B. dem 3. Oktober. Die Schätzung des Kalendereffekts sollte daher separat für passend zu wählende Gruppen erfolgen. Eine Schätzung für jede betreffende Woche einzeln durchzuführen würde gerade bei kurzen bis mittellangen Zeitreihen eine deutliche Überanpassung für diese Wochen bewirken und ist daher nicht angemessen.

3.3 Transformation

Die in (3.1) und (3.11) genannten Dekompositionsmodelle sind in allen Komponenten additiv, was eine vereinfachende und nicht überprüfbare Annahme darstellt. Ebenso denkbar wäre auch ein multiplikativer Zusammenhang zwischen Y_t und den Komponenten. Darüber hinaus können auch Situationen, in denen sich die Varianz der stationären Fehlerkomponente ε_t abhängig von t oder $\mathrm{E}\, Z_t$ ändert, durch diese Modelle nicht abgebildet werden. Mit Hilfe einer geeigneten Transformation $g : \mathbb{R}^+ \to \mathbb{R}$ können die Beobachtungen $z_1, \ldots, z_n$ hinsichtlich der Varianz stabilisiert werden. Box und Cox (1964) diskutieren am Beispiel linearer

Modelle Transformationsfunktionen der Form

$$g(z,\lambda) \;=\; z^{(\lambda)} \;=\; \begin{cases} \frac{z^{\lambda}-1}{\lambda} & , \quad \lambda \neq 0 \\ \log z & , \quad \lambda = 0 \end{cases} \;, \tag{3.12}$$

die auch häufig bei der Analyse von Zeitreihen verwendet werden. Auf diesen Ansatz beziehen sich z. B. Box und Jenkins (1970) in ihrer wegweisenden Arbeit zur Analyse von Zeitreihen durch ARMA-Prozesse. Bei Wahl eines passenden Parameters λ kann angenommen werden, dass die transformierte Zeitreihe $\{Y_t\} = \{g(Z_t, \lambda)\}$ durch das Modell (3.1) oder (3.11) geeignet dargestellt werden kann. Dies impliziert insbesondere auch, dass die Verteilung des Fehlers ε_t im transformierten Modell unabhängig von der Summe $f(t) = m(t)+s(t)+c(t)+k(t)$ der deterministischen Komponenten ist und damit die Additivität der Komponenten eine vertretbare Annahme darstellt.

4 Schrittweise nichtparametrische Komponentenschätzung

4.1 Transformation

In diesem Kapitel werden Möglichkeiten für die Modellierung der Komponenten in (3.1) bzw. (3.11) am Beispiel der Zeitreihen der Infektionsdaten $\{z_t\}$ entwickelt und die Ergebnisse interpretiert. Wie in Kapitel 3 gezeigt wurde, ist vor der Schätzung der einzelnen Komponenten zu überprüfen, ob die strukturelle Annahme von (3.11), insbesondere die Additivität der Komponenten und die stationäre Verteilung des Fehlerterms ε_t als erfüllt angesehen werden können. Letzteres impliziert u. a., dass die Varianz der ε_t konstant ist für $t = 1, \ldots, 313$. Diese Eigenschaft ist aber nur überprüfbar, wenn die unbekannten Fehler geschätzt werden können, was wiederum voraussetzt, dass eine Schätzung $\hat{f}(t)$ des Signalterms $f(t)$, also der Summe der nicht zufälligen Komponenten vorliegt. Der Fehler kann in dem Fall durch die Residuen als $\hat{\varepsilon}_t = z_t - \hat{f}(t)$ geschätzt werden. Weist die Zeitreihe der Fehler dann keine konstante Varianz auf, kann diese Eigenschaft in vielen Fällen durch eine varianzstabilisierende Transformation der Beobachtungen z_t, $t = 1, \ldots, 313$, entsprechend (3.12) erreicht werden. Der ursprüngliche Ansatz von Box und Cox (1964) verlangt dafür ein lineares Modell für den Signalterm $f(t)$. Ohne Schätzung $\hat{f}(t)$ kann ein geeigneter Parameter durch den von Jenkins (1979) vorgestellten Range-Mean-Plot gefunden werden, mit dem jedoch nur eine ungefähre Bestimmung von λ möglich ist. Außer dieser Ungenauigkeit hat das Verfahren den Nachteil, dass es nicht robust gegenüber wenigen ungünstig liegenden Beobachtungen ist.

Aus diesen Gründen wird hier der Ansatz verfolgt, zunächst $f(t)$ geeignet zu schätzen und darauf aufbauend durch ein anderes Verfahren einen geeigneten Wert λ zu finden. Da keine strukturellen Annahmen an den Signalterm gemacht werden, also kein parametrisches Modell betrachtet wird, kann $f(t)$ nur mit einem

nichtparametrischen Verfahren geschätzt werden. Viele nichtparametrische Verfahren setzen allerdings insbesondere unabhängige Fehler voraus, was bei Zeitreihen wegen der typischen Abhängigkeiten aufeinanderfolgender Beobachtungen und Fehler im Allgemeinen nicht als erfüllt angesehen werden kann. Die Art der Abhängigkeit ist außerdem unbekannt und kann auch ohne zusätzliche Kenntnis oder Annahmen über $f(t)$ nicht bestimmt werden. Problematisch ist, dass die Verletzung dieser Annahmen zu Verzerrungen bei der Signalschätzung führt, worauf z. B. Herrmann et al. (1992) für Kernschätzer oder Opsomer et al. (2001) auch für Splines hinweisen. Wichtig ist an dieser Stelle jedoch festzuhalten, dass die Schätzung von $f(t)$ nicht endgültig für das weitere Vorgehen ist und keinen direkten Einfluss auf die später folgende Schätzung der Komponenten besitzt. Es handelt sich lediglich um eine Interim-Schätzung, um eine Bewertung des Fehlers und damit eine Transformationsempfehlung für die Daten zu erhalten.

Unter der vorübergehenden Annahme der Gültigkeit des nichtparametrischen Regressionsmodells (2.1) wird das durch $z_t = f(t) + \varepsilon_t$ charakterisierte Signal $f(t)$ für die beobachteten Anzahlen der Meldefälle $z_1, \ldots, z_{313}$ mit den verschiedenen in Kapitel 2 vorgestellten Methoden geschätzt. Abb. 4.1 zeigt die Schätzungen $\hat{f}(t)$ der Signalkomponente für die Meldefälle von Campylobacter und Rotavirus.

Für die Meldefälle von Campylobacter liefert die Methode mit lokaler Kernschätzung (LOK) eine unterglättete Schätzung: Die jährlichen Maxima werden hier oft zweigipflig geschätzt, was wegen der schmalen Zeitspanne, in der sie auftreten, jedoch überangepasst erscheint. Im Bereich des Jahreswechsels schwankt die Schätzung kurzfristig sehr stark und kann damit die kalendereffektbedingte Unregelmäßigkeit bei den Meldefällen nicht glätten. Ähnliche Anpassungsdefizite können auch bei den Ergebnissen durch *adaptive weights smoothing* (AWS, $h_{\max} = 26$) und glatte Straffe Saite (STS) beobachtet werden. Bedenklich an dem Ausschlag zu Jahresende ist dabei weniger die Nicht-Glattheit bzw. Nicht-Differenzierbarkeit der Schätzung als die Tatsache, dass diese in den Wochen vor dem Jahreswechsel nach unten, die in den Wochen nach dem Jahreswechsel nach oben verzerrt ist. Verfahren, die in dieser Situation robusteres Verhalten zeigen, wären daher zu bevorzugen. Diese Eigenschaft zeigt unter den betrachteten Verfahren nur die *singular spectrum analysis* (SSA). Für die Fensterbreite $L = 52$, d. h. also einer Breite von der Länge eines Jahres, und die Berücksichtigung der ersten fünf Eigentripel ist die daraus erstellbare Rekonstruktion weder unterglättet (keine überflüssigen Maxima oder hohe Volatilitäten innerhalb kleiner zeitlicher Abschnitte) noch überglättet, denn die glättende Funktion geht bis in die Spitzen und Senken der Beobachtungen. Es ist daher sinnvoll, sie als vorläufige Schätzung

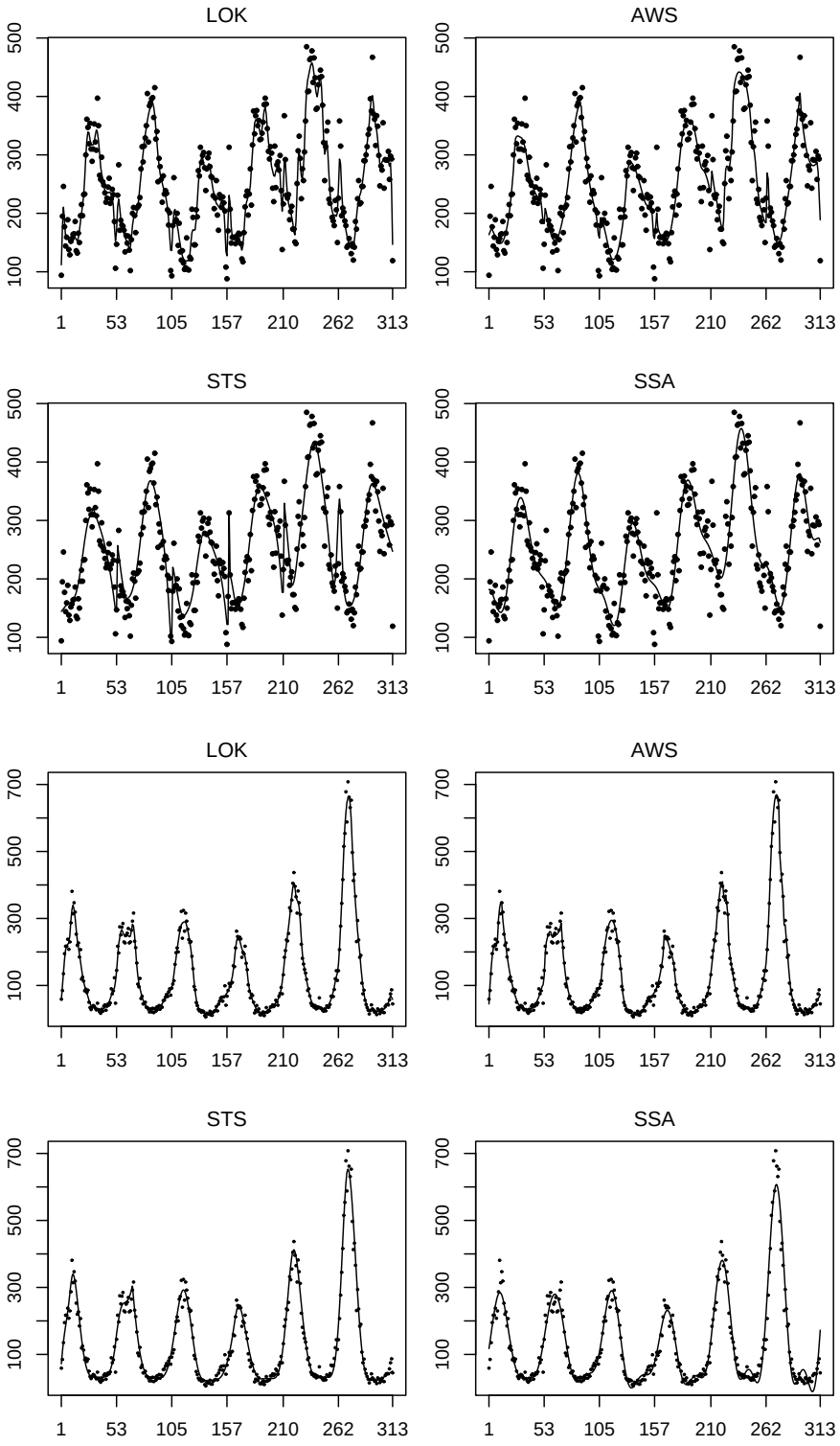

Abb. 4.1: Meldefälle und geschätztes Signal durch LOK, AWS ($h_{max} = 26$), STS und SSA ($L = 52$, 5 Eigentripel) für CAM (oben) und RTV (unten).

von $f(t)$ beim Finden einer geeigneten Transformation zu verwenden.

Da der Verlauf der Rotavirus-Meldefälle eine sichtlich geringere lokale Streuung aufweist, sind die Abweichungen der geschätzten Funktionen weniger groß als bei Campylobacter. Hier ist jedoch das durch SSA erzielte Ergebnis unter Verwendung derselben Einstellungen $L = 52$ und der ersten fünf Eigentripel im letzten Jahr nicht überzeugend. Dort werden die letzten Wochen mit auffälliger Abweichung nicht korrekt angepasst. Auch die Schätzung durch STS ist in den Jahren 2002 bis 2004 in der jeweils zweiten Hälfte nicht passend, da der Verlauf der Meldefälle trotz sehr geringer Varianz hier nicht durch die geschätzte Funktion nachvollzogen wird. Die durch das LOK- bzw. AWS-Verfahren bestimmten Schätzer sind einander sehr ähnlich, die Anpassung an die Meldefälle ist gut und die bei den anderen Verfahren beobachtbaren Abweichungen sind nicht vorhanden. Aus diesen Gründen könnten die Ergebnisse beider Verfahren verwendet werden. Da die summierte quadratische Abweichung von den Beobachtungen beim AWS-Verfahren mit 179 054 aber größer als die von LOK mit 178 757 ist, wird letzteres in der weiteren Analyse verwendet.

Um eine geeignete Transformation der Beobachtungen zu finden, wird die von Box und Cox (1964) vorgeschlagene Klasse von Funktionen (3.12) betrachtet. Diese Transformationen sind, obwohl ursprünglich im Kontext linearer Modelle entwickelt, auch in der Zeitreihenanalyse häufig eingesetzt worden. Der ursprüngliche Likelihood- oder Bayes'sche Ansatz zum Auffinden bzw. Schätzen des Parameters λ kann hier im nichtparametrischen Kontext nicht verfolgt werden, da für Z_t keine Verteilungsannahme gemacht wird und Unabhängigkeit der Beobachtungen ebenfalls nicht vorausgesetzt werden kann. Darüberhinaus existiert keine strukturelle Annahme an $f(t)$, wie sie bei Box und Cox (1964) durch das lineare Regressionsmodell eingeht. Aus diesem Grund wird an dieser Stelle, in Anlehnung an eine Idee von Guerrero (1993), ein modellfreier, heuristischer Ansatz vorgeschlagen: Für den gegebenen Beobachtungszeitraum $T = \{1, \ldots, 313\}$ und einen Parameter $k \in \{1, \ldots, 26\}$ sei zunächst durch $\mathcal{P}^k$ eine Partition von T in zwei Mengen P_1^k und P_2^k gegeben derart, dass

$$P_1^k = \bigcup_{i \in \{1,53,105,157,210,262\}} [i + k, i + 25 + k], \tag{4.1}$$

und $P_2^k = T \backslash P_1^k$. Damit enthalten P_1^k und P_2^k zeitliche Bereiche oder Fenster der Breite 26, was der Anzahl der Wochen eines halben Jahres entspricht. Eine Ausnahme bilden diejenigen Fenster, die in das Schaltjahr 2004 fallen. Hier beträgt die Fensterbreite in P_2^k bzw. der Abstand der Fenster in P_1^k einmalig 27. Aufeinanderfolgende Fenster gehören jeweils verschiedenen Mengen an. Der Parameter k bewirkt eine Verschiebung der Fenster zur Jahresmitte hin; dadurch

entstehen für $k > 0$ in P_2^k zwei am Rand von T liegende Fenster mit kürzerer Länge. Unabhängig von k ist die Anzahl aller Elemente in den Mengen von P_1^k und P_2^k mit 156 bzw. 157 nahezu gleich. Parameterwerte $k > 26$ müssen nicht betrachtet werden, da $P_1^k = P_2^{k-26}$ und $P_2^k = P_1^{k-26}$, und damit in solchen Fällen durch Umbenennung der Mengen dieselben Partitionen wie in (4.1) betrachtet würden.

Unter der Annahme der Gültigkeit des durch (3.1) oder (3.11) gegebenen Modells ist die Varianz σ^2 des Fehlers ε_t konstant für alle $t \in T$. Für einen erwartungstreuen Schätzer $\hat{f}(t)$ folgt dann, dass die Residuen $\hat{\varepsilon}_t = z_t - \hat{f}(t)$ ebenfalls gleiche Varianz besitzen. Insbesondere gilt für eine beliebige Partition $\mathcal{P}^k$, $k \in \{1, \ldots, 26\}$, dass die empirischen Varianzen $\mathrm{Var}\{\hat{\varepsilon}_t | t \in P_1^k\} = \mathrm{Var}\{\hat{\varepsilon}_t | t \in P_2^k\}$ übereinstimmen. Indem statt der empirischen Varianz der mad (median absolute deviation) verwendet wird, ist durch

$$U_\lambda(\mathcal{P}^k) \;=\; \max\left\{\frac{\mathrm{mad}\{\hat{\varepsilon}_t | t \in P_1^k\}}{\mathrm{mad}\{\hat{\varepsilon}_t | t \in P_2^k\}}, \frac{\mathrm{mad}\{\hat{\varepsilon}_t | t \in P_2^k\}}{\mathrm{mad}\{\hat{\varepsilon}_t | t \in P_1^k\}}\right\}, \qquad k \in \{1, \ldots, 26\} \quad (4.2)$$

eine robuste Kennzahl für den Unterschied der empirischen Varianzen der Residuen in den beiden Fensterbereichen der Partition $\mathcal{P}^k$ gegeben. Je mehr sich das Maximum von $U_\lambda(\mathcal{P}^k)$ für alle $k = 1, \ldots, 26$ dem Wert 1 nähert, desto weniger spricht das gegen die Annahme der konstanten Varianz des Fehlers. Um einen geeigneten Transformationsparameter zu finden, kann $U_\lambda(\mathcal{P}^k)$ für alle $\mathcal{P}^k$ und λ bestimmt werden. Für denjenigen Parameter λ_0, für den $U_\lambda(\mathcal{P}^k)$ unabhängig von $\mathcal{P}^k$ den kleinsten Wert annimmt, passen die damit transformierten Beobachtungen am besten zu den Modellannahmen bzgl. der konstanten Varianz. Also wird

$$\lambda_0 \;=\; \min_{\lambda \in [0,1]} \max_{k \in \{1,\ldots,26\}} U_\lambda(\mathcal{P}^k) \quad (4.3)$$

gewählt.

Die analytische Bestimmung von λ_0 ist kompliziert, da dabei auch die Schätzung $f_\lambda(t)$ eingeht. Betrachtet man $U_\lambda(\mathcal{P}^k)$ für mehrere Werte $\lambda \in [0, 1]$, stellt man fest, dass diese für die Campylobacter- wie für die Rotavirus-Fälle einen einfachen Verlauf mit einem deutlichen globalen Minimum aufweisen. Für diese Daten zeigt Abb. 4.2 die Werte von $\max_k U_\lambda(\mathcal{P}^k)$ für $\lambda = \{0, 0.01, 0.02, \ldots, 1\}$. Das Minimum wird an der Stelle $\lambda_0 = 0.78$ angenommen und kann im Folgenden als Parameter für die Transformation verwendet werden. Auf eine genauere Untersuchung von U_λ in einer Umgebung z. B. $[0.77, 0.79]$ von λ_0 kann verzichtet werden, weil die zugrunde liegende Schätzung $\hat{f}_\lambda(t)$ nicht als tatsächliche Schätzung des Signals angenommen wird. Außerdem sind die Änderungen von

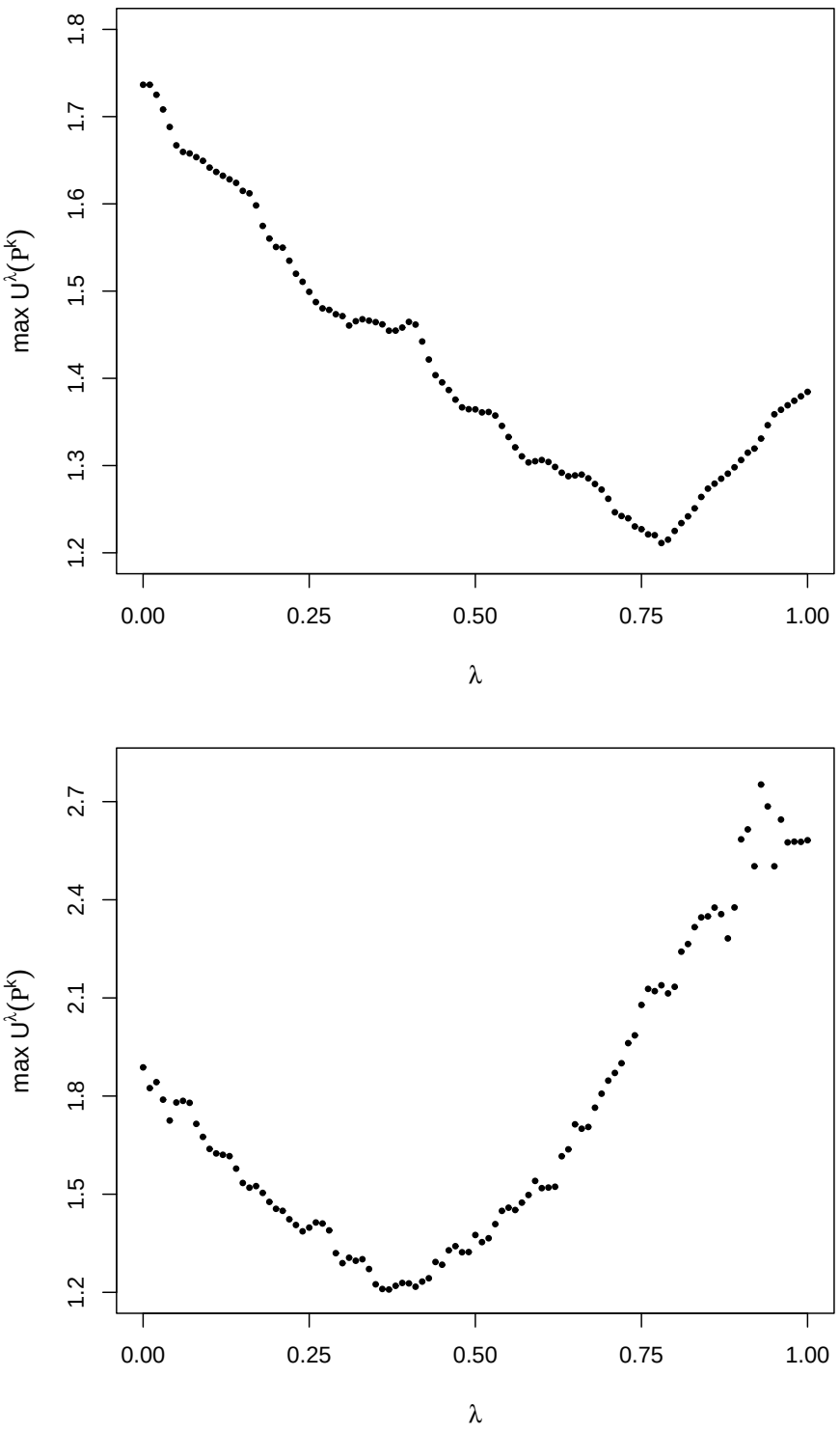

Abb. 4.2: Maximum des Verhältnisses U^{λ} von MAD von Residuen bei disjunkter Zerlegung des Beobachtungszeitraums in 2 gleichgroße Teilbereiche mit jeweils 6 Intervallen gleicher Länge für verschiedene Werte von λ für CAM (oben) und RTV (unten).

U_λ in diesem Bereich so gering, dass das Auffinden des exakten Minimums keinen relevanten Einfluss auf die weitere Modellierung haben würde.

4.2 Auswahl von Wochen mit Kalendereffekten

Vor der Schätzung der einzelnen Komponenten ist zu überlegen, wie Anzahlen zu behandeln sind, die in Wochen mit Feiertagen oder Schulferien gemeldet worden sind. Aus den in Abschnitt 1.2 aufgeführten Gründen ist es möglich, dass der Erwartungswert der Meldefälle $\mathrm{E}\, Z_t$ in einer Woche t auch davon abhängt, ob diese Woche zwischen Montag und Freitag einen Feiertag enthält oder in die Zeit der Schulferien fällt. Derart kalenderbedingte Effekte sind in (3.11) durch die Funktion $k(t)$ berücksichtigt. Diese kann jedoch nicht ohne die Kenntnis der übrigen deterministischen Komponenten bestimmt werden, und deswegen erfolgt ihre Schätzung erst nach der von $m(t)$, $s(t)$ und $c(t)$. Die Schätzung von $m(t)$, $s(t)$ und $c(t)$ soll wiederum nicht durch Wochen mit Kalendereffekten verzerrt werden. Sie müssen daher vor der Schätzung dieser Komponenten identifiziert werden, und die zugehörigen Beobachtungen sollten bei den jeweiligen Schätzungen unberücksichtigt bleiben.

Dass es tatsächlich sinnvoll ist, Wochen mit Feiertagen durch eine gesonderte Komponente $k(t)$ zu analysieren, verdeutlicht Abb. 4.3. Hier werden die Residuen $\hat{\varepsilon}_t = y_t - \hat{f}(t)$ gezeigt, nachdem das Signal $f(t)$ für Campylobacter durch SSA-Zerlegung (mit $L = 52$, erste 5 Eigentripel, wie in 4.1) und für Rotavirus mit der LOK-Methode geschätzt worden ist. In beiden Fällen sind die Beobachtungen y_t in Wochen mit Feiertagen durch Interpolation der Meldezahlen der nächstliegenden Wochen ohne Feiertage ersetzt worden, so dass ein Einfluss von Kalendereffekten auf die Schätzung $\hat{f}(t)$ ausgeschlossen ist. Die Residuen sind aber in jedem Fall für die tatsächlichen und nicht für die interpolierten Beobachtungen berechnet. Für eine genauere Analyse können die betreffenden Meldewochen in drei Gruppen eingeteilt werden: Die Weihnachts-/Neujahrsperiode umfasst die vier Wochen seit Weihnachten, Ostern die beiden um das Osterwochenende liegenden Wochen. In beiden Zeiträumen liegen auch in jedem Jahr die jeweiligen Schulferien. Alle übrigen Wochen, die lediglich einen einzelnen Feiertag enthalten und in der Regel nicht Teil der Schulferien sind, werden in einer dritten Gruppe zusammengefasst.

Wie aus Abb. 4.3 ersichtlich ist, sind die Residuen in jeder dieser Gruppe auffällig: In der Weihnachts-/Neujahrsperiode liegen die Residuen deutlich abseits des Mittelwertes 0. Bei Campylobacter sind die jeweils ersten Residuen deutlich negativ, die letzten hingegen deutlich positiv. Dieses Muster kann

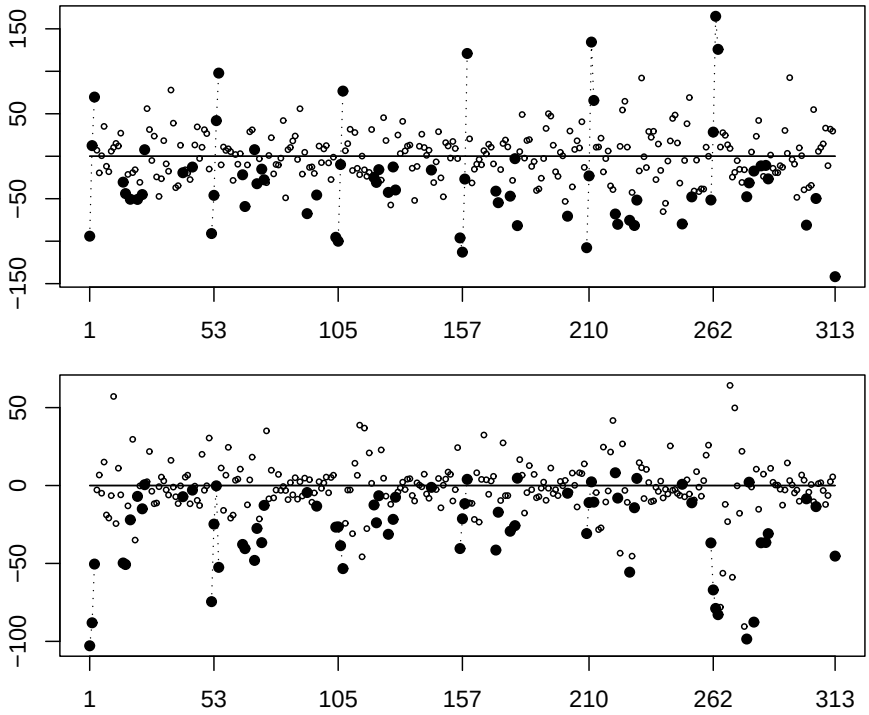

Abb. 4.3: Residuen nach Schätzung der Signalkomponente $f(t)$ durch SSA-Zerlegung ($L = 52$, 5 Eigentripel) für CAM (oben) und Residuen nach Schätzung der Signalkomponente $f(t)$ durch LOK für RTV (unten). Beobachtungen aus Wochen mit Kalendereffekten wurden bei der Schätzung durch interpolierte Werte ersetzt. Die Residuen für diese Wochen sind gesondert markiert (•), zeitlich aufeinanderfolgende Residuen in der Weihnachts-/Neujahrperiode sind zusätzlich durch Linien verbunden.

bei Rotavirus zwar nicht erkannt werden, doch auch hier liegen die betreffenden Residuen nah beieinander und abseits der übrigen Werte. Residuen aus den Osterzeiträumen sind bei beiden Krankheiten stets deutlich negativ. Und auch die Residuen aus der dritten Gruppe weichen mehrheitlich nach unten ab. Während die Wochen der Weihnachts-/Neujahrsperiode und um Ostern damit eindeutige Kalendereffekte aufweisen, sollten die Wochen aus der dritten Gruppe nur dann in die Kalenderkomponente eingeschlossen werden, falls sich die Abweichung von 0 konfirmatorisch nachweisen lässt. Dazu wird ein einseitiger Wilcoxon-Vorzeichen-Rangtest mit der Hypothese, dass die Verteilung der zugehörigen Residuen den Erwartungswert 0 besitzt, zum Niveau 0.05 durchgeführt. Da eine Abweichung der Residuen in den Wochen erwartungsgemäß nur nach unten erfolgen kann, wird das einseitige Testproblem betrachtet. Für beide Situationen resultiert ein so geringer p-Wert, dass die Hypothese abgelehnt werden kann (p-Wert (CAM) = $2.4 \cdot 10^{-6}$, p-Wert (RTV) = $5.0 \cdot 10^{-5}$). Daher sollten auch Wochen mit einzelnen Feiertagen als Träger möglicher Kalendereffekte behandelt werden. Die Menge der Vereinigung aller Wochen mit Kalendereffekten wird mit T_1, die Menge der Wochen ohne Kalendereffekte als T_0 bezeichnet.

Anmerkung: Gleichbedeutende Ergebnisse erhält man auch bei Verwendung anderer Verfahren zur Schätzung des Signals. Im Anhang (Abb. A.2) sind entsprechende Abbildungen der Residuen nach Schätzung des Signals durch LOK für die Campylobacter- und durch AWS für die Rotavirus-Meldefälle aufgeführt. Diese führen zu denselben Schlussfolgerungen.

Die naheliegende Vermutung, dass die unterdurchschnittlichen Meldezahlen in Wochen mit Kalendereffekten durch überdurchschnittliche Anzahlen in den darauffolgenden Wochen gewissermaßen ausgeglichen werden, kann nicht bestätigt werden. Die Hypothese, dass die Verteilung der Residuen in Wochen, die direkt auf solche mit Kalendereffekten folgen, den Erwartungswert ≤ 0 besitzt, kann in beiden Fällen mit dem Wilcoxon-Vorzeichen-Rangtest zum Niveau 0.05 nicht widerlegt werden (p-Wert (CAM) = 0.953, p-Wert (RTV) = 0.932). Wie auch Abb. 4.4 zeigt, streuen die Residuen für diese Wochen unregelmäßig um Null und sind betragsmäßig nicht auffällig von den Residuen der anderen Wochen verschieden.

Ebenso können Wochen mit Schulferien auf auffällige Meldezahlen überprüft werden. Da die Weihnachts- und Osterferien wegen der enthaltenen Feiertage bereits berücksichtigt sind, bleiben nur die Wochen mit Sommer- und Herbstferien zu untersuchen. Bei der Bewertung der Residuen ist zu berücksichtigen, dass die Sommerferien im jährlichen Zeitraum mit den meisten Fällen liegen und schon deswegen mehrheitlich positiv sind. Sofern es einen durch Schulferien bedingten

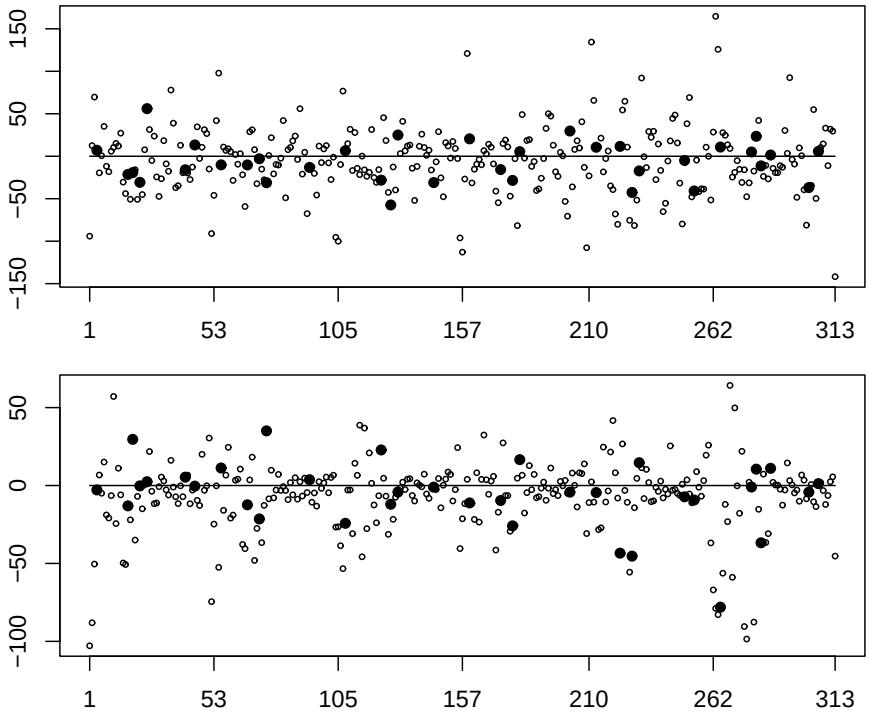

Abb. 4.4: Residuen nach Schätzung der Signalkomponente $f(t)$ durch SSA-Zerlegung ($L = 52$, 5 Eigentripel) für CAM (oben) und Residuen nach Schätzung der Signalkomponente $f(t)$ durch LOK für RTV (unten). Beobachtungen aus Wochen mit Kalendereffekten wurden bei der Schätzung durch interpolierte Werte ersetzt. Die Residuen in Wochen ohne Feiertage, die auf Wochen mit Feiertagen folgen, sind gesondert markiert (•).

Einfluss gibt, wird er durch die saisonale Komponente überlagert und ist daher auch wegen der Länge der Ferien von sechs Wochen nicht bestimmbar. Auffälligkeiten wie bei den oben betrachteten Wochen mit Feiertagen sind weder hier noch in den Herbstferien zu erkennen, weswegen diese Zeiträume nicht bei der Schätzung der Kalenderkomponente $k(t)$ einbezogen werden.

Im weiteren Verlauf werden, sofern nicht anders vermerkt, stets die transformierten Meldefälle von Campylobacter

$$y_t = g_{0.78}(z_t) = \frac{z_t^{0.78} - 1}{0.78}, \qquad t = 1, \ldots, 313, \tag{4.4}$$

bzw. die transformierten Meldefälle von Rotavirus

$$y_t = g_{0.37}(z_t) = \frac{z_t^{0.37} - 1}{0.37}, \qquad t = 1, \ldots, 313, \tag{4.5}$$

betrachtet. Das Dekompositionsmodell (3.11) wird also für die transformierten Zeitreihen $y_1, \ldots, y_{313}$ angenommen.

4.3 Trend

Häufig verwendete Ansätze zur Trendschätzung postulieren ein lineares Modell, in dem als Regressor eine einfache Funktion wie z. B. eine Exponentialfunktion oder ein Polynom geringen Grades eingesetzt wird. Im Gegensatz zu einem solchen modellgestützten, parametrischen Ansatz stellt das Glätten bzw. die Trendschätzung durch Bildung linearer Filter eine modellfreie Alternative dar. Häufig werden für eine Zeitreihe $z_1, \ldots, z_n$ gleitende Durchschnitte w_t der Länge $2q + 1$, $q \in \mathbb{N}$,

$$w_t = \frac{1}{2q+1} \sum_{i=-q}^{q} z_{t-i}, \qquad t = 1, \ldots, n,$$

verwendet, wobei z. B. $z_t = z_1$ für $t < 1$ und $z_t = z_n$ für $t > n$ gesetzt werden, damit w_t auch für die Ränder der Zeitreihe berechnet werden kann. Die Verwendung gleitender Durchschnitte ist aber für die vorliegenden Meldefalldaten nicht sinnvoll. Es muss $2q + 1$ mindestens größer als 104 gewählt werden, da die resultierende Trendschätzung andernfalls nur die Veränderung von einem Jahr auf das folgende beschreibt und damit nicht global ist. Allerdings ist die Schätzung dann im ersten und letzten Jahr wegen der breiten Randbereiche nicht aussagekräftig.

Auf Kernschätzern aufbauende Verfahren haben wie lineare Filter für die Trendschätzung den Nachteil, dass bei Verwendung schmaler Bandbreiten nur ein lokaler und kein globaler Verlauf bestimmt werden kann. Zwar existieren z. B.

bei Hart (1991) für den von Gasser und Müller (1979) eingeführten Kernschätzer Vorschläge für eine geeignete Bandbreitenwahl, doch wird damit ebenfalls keine sinnvolle Schätzung an den Randbereichen der Zeitreihe möglich. Geeignete Kandidaten können damit durch das auf Kernschätzern basierende LOK-Verfahren nicht gefunden werden. Weil sie ebenfalls nur eine lokale Anpassung bewirken, können außerdem das AWS- und das STS-Verfahren nicht eingesetzt werden. Zwar kann bei AWS eine hohe maximale Bandbreite durch $h_{\max}$ gewählt werden, doch dies führt nicht zu einer globalen Funktion, die vom starken saisonalen Effekt unbeeinflusst bleibt.

Eine den in Abschnitt 3.1 für die Trendkomponente aufgestellten Vorgaben entsprechende Schätzung ist durch SSA-Zerlegung möglich. Hier müssen zunächst eine passende Fensterbreite und die Anzahl der eingehenden Eigentripel bestimmt werden. Für die Meldefälle von Campylobacter wie auch von Rotavirus gilt, dass bei Fensterbreiten $L \geq 52$ nur das jeweils erste Eigentripel eine als Trend interpretierbare Komponente und damit eine grundsätzlich geeignete Kandidatenfunktion liefert. Die übrigen Eigentripel führen entweder zu Verläufen mit mehreren Extrema, sind sogar periodisch, oder sie weisen zu geringe Amplituden auf und sind daher unbedeutend. Veranschaulicht wird dies am Beispiel der Campylobacter-Fälle für die Fensterbreiten $L = 156$ und $L = 104$ in Abb. 4.5, die die aus den ersten bis zu 4 Eigentripeln erzeugten Zeitreihen zeigt.

Der Vergleich verschiedener Fensterbreiten zeigt, dass eine aus dem jeweils ersten Eigentripel rekonstruierte Komponente für Werte von L nahe 52, 104 oder 156 jeweils wenige Extrema aufweist. Je mehr die Fensterbreite von einem dieser Werte abweicht, d. h. je näher L an 78 oder 130 kommt, desto mehr lokale Extrema und Wendepunkte entstehen und entsprechend größer wird die lokale Variabilität. Abb. 4.6 zeigt die so erhaltenen Kandidaten für einige ausgewählte Fensterbreiten L im ursprünglichen Maßstab. Wegen ihrer Monotonie und geringen Krümmung sind die Resultate nach $L = 104$ und $L = 156$ prinzipiell geeignet. $L = 104$ führt aber bei ansonsten nahezu konstantem Verlauf zu einer plötzlichen Niveauänderung von 2003 auf 2004. Der gleichmäßige Anstieg, der für $L = 156$ beobachtbar ist, entspricht dagegen eher den gewünschten Eigenschaften einer Trendkomponente und ist daher vorzuziehen. Parameterwerte wie $L = 154, 155$ führen dabei zu sehr ähnlichen Ergebnissen.

Die Entscheidung für eine bestimmte Trendschätzung erfolgt jedoch nicht durch graphische Beurteilung, sondern unter Berücksichtigung des Wertes der Krümmung γ aus (3.2). Bei der Wahl von $L = 156$ wird hierfür das kleinste Ergebnis erreicht; Tabelle 4.4 enthält γ für einige Kandidaten bei Wahl verschiedener Parameter. Da sowohl durch γ wie auch durch graphischen Vergleich derselbe

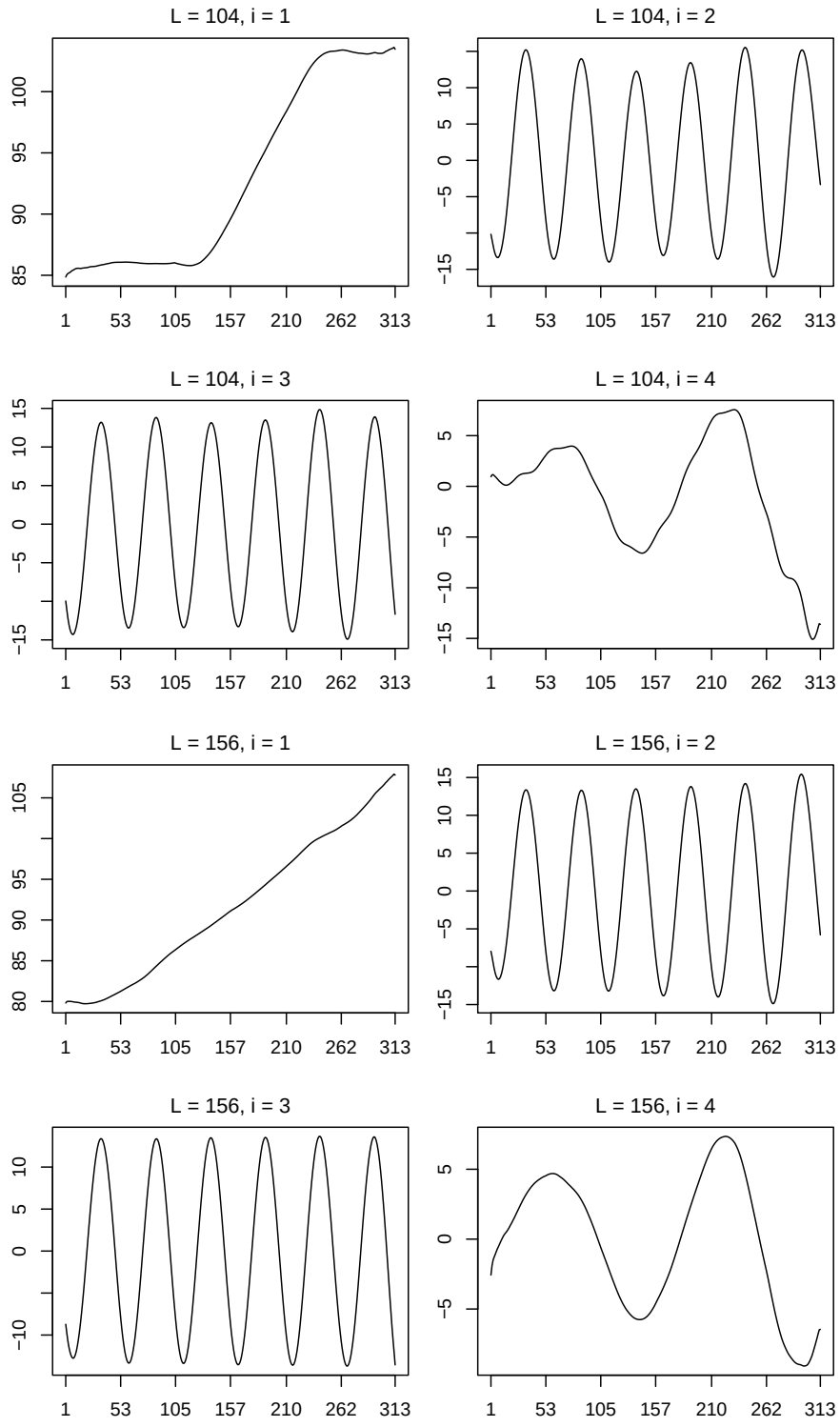

Abb. 4.5: Komponeten durch Eigentripel $i = 1, \ldots, 4$ nach SSA-Zerlegung mit Fensterbreite $L = 104$ (oben) und $L = 156$ (unten) für die transformierten Beobachtungen (CAM).

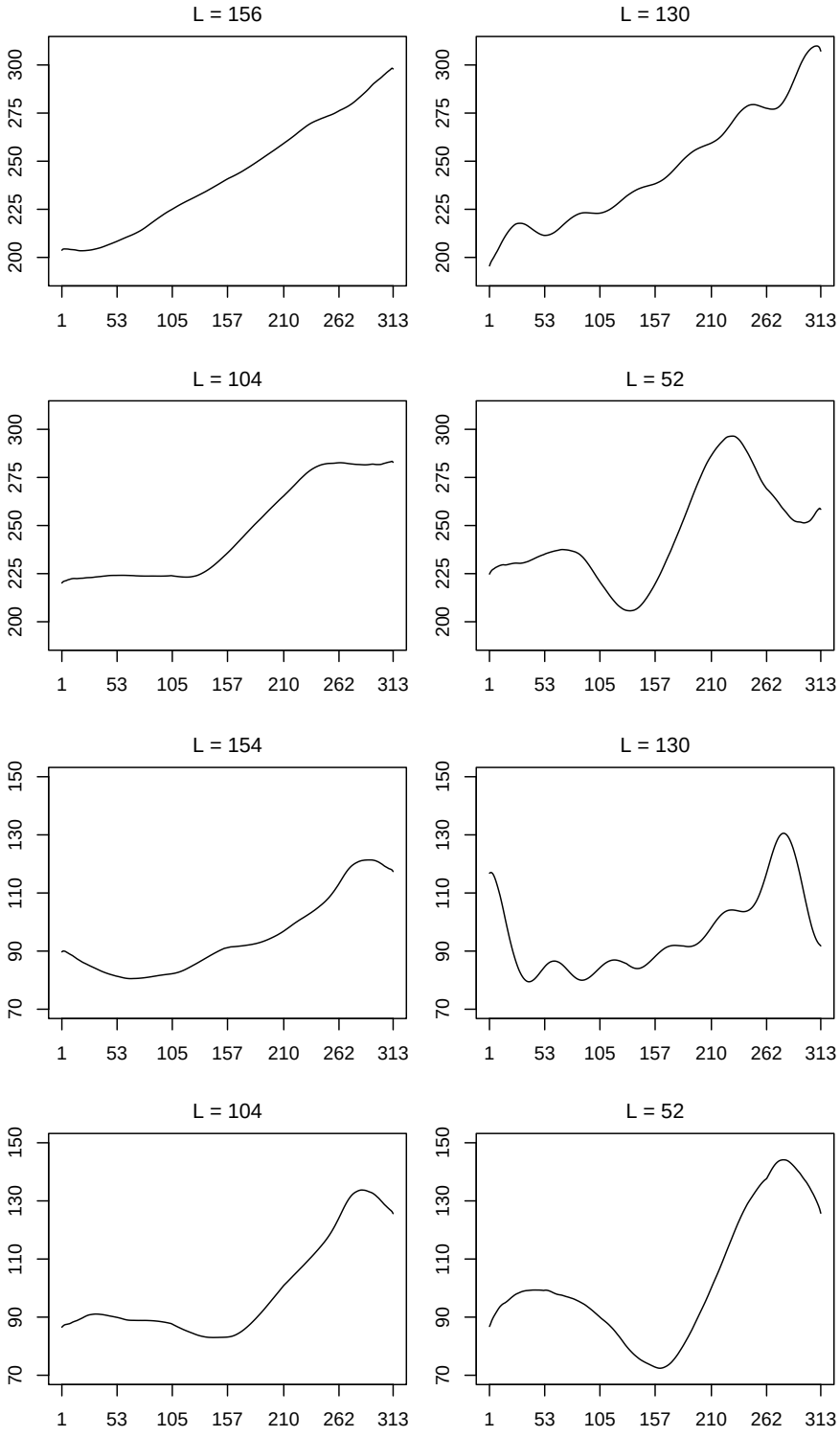

Abb. 4.6: Kandidatenfunktionen für die geschätzte Trendkomponente nach Rekonstruktion aus dem jeweils ersten Eigentripel einer SSA-Zerlegung bei verschiedenen Fensterbreiten L für CAM (oben) und RTV (unten).

L	γ [CAM]	γ [RTV]
52	$28.6 \cdot 10^{-5}$	$4.4 \cdot 10^{-6}$
78	$53.2 \cdot 10^{-5}$	$24.9 \cdot 10^{-6}$
104	$5.2 \cdot 10^{-5}$	$1.3 \cdot 10^{-6}$
130	$23.4 \cdot 10^{-5}$	$11.2 \cdot 10^{-6}$
154	$16.0 \cdot 10^{-5}$	$1.0 \cdot 10^{-6}$
156	$3.1 \cdot 10^{-5}$	$1.2 \cdot 10^{-6}$

Tab. 4.4: Krümmung γ von Kandidatenfunktionen für die geschätzte Trendkomponente nach Rekonstruktion aus dem jeweils ersten Eigentripel einer SSA-Zerlegung bei verschiedenen Fensterbreiten L.

Parameter $L = 156$ als der geeignetste gefunden wird, kann dies als weiterer Beleg für die sinnvolle Definition des Krümmungskriteriums γ angesehen werden.

Bei den Ergebnissen für Rotavirus können ähnliche Beobachtungen getroffen werden. Die Krümmung ist am geringsten, wenn die Fensterbreite durch $L = 154$ gewählt wird. Das entsprechende Ergebnis wird demnach hier für die Schätzung der Trendkomponente verwendet.

Nach Rücktransformation $g^{-1}(\hat{m}(t))$ der geschätzten Komponente auf den ursprünglichen Maßstab kann die Bedeutung des Trends bewertet werden: Bei Campylobacter ist so eine gleichmäßige Zunahme von etwa 80 wöchentlichen Fällen über den Zeitraum aller untersuchten Jahre festzustellen. Die Zunahme ist bei Rotavirus mit etwa 40 Fällen geringer, außerdem besitzt die Komponente zwei wenn auch schwach ausgeprägte Extrema, die einen weniger regelmäßigen Verlauf bewirken.

4.4 Saison

Für die Saisonkomponente gilt die charakteristische Eigenschaft, dass $s(t_1) = s(t_2)$ erfüllt ist für zwei Zeitpunkte $t_1, t_2 \in [0, 313]$, deren zeitlicher Abstand genau der Länge eines Jahres oder eines ganzzahligen Vielfachen davon entspricht. Im betrachteten Zeitraum gibt es jedoch keine zwei Zeitpunkte mit dieser Eigenschaft, die zusätzlich Ende einer Meldewoche sind und für die damit eine zugehörige Anzahl Meldefälle existiert, denn ein Jahr mit 365 oder 366 Tagen besteht in keinem Fall aus einer ganzzahligen Anzahl Wochen. Die naheliegende Idee, die Saisonkomponente durch Bildung des arithmetischen Mittels derjenigen Beobachtungen zu schätzen, die genau einen oder einen mehrjährigen Abstand voneinander besitzen, kann also nicht umgesetzt werden, weil es keine Paare oder

Gruppen von Beobachtungen mit dieser Eigenschaft gibt.

Um die Saison als periodische Funktion geeignet schätzen zu können, ist es stattdessen sinnvoll, die Meldefälle zeitlich neu gemäß ihrer jeweiligen Distanz zum Jahresbeginn anzuordnen. Diese Entfernung kann aus den ursprünglichen Meldewochen durch

$$\delta(t) \;=\; \begin{cases} \frac{7t}{365} \bmod 1, & t \leq 156 \\ \frac{7(t-156)-3}{366}, & 156 < t \leq 209 \\ \frac{7(t-208)+2}{365} \bmod 1, & 209 < t \end{cases}, \tag{4.6}$$

ermittelt werden. Im neu sortierten Datensatz $(\delta(t), y_t)$ sind damit gewissermaßen alle Meldezahlen auf ein Jahr parallel übereinander gelegt, da $\forall t \in [0, 313]\ \delta(t) \in [0, 1]$. Wird für

$$y_t - \hat{m}(t) \;=\; f(\delta(t)) + \varepsilon_t, \qquad t \in T_0, \tag{4.7}$$

das Signal $f(.)$ der trendbereinigten, neu sortierten Zeitreihe geschätzt, kann das Ergebnis als jahrestypischer und damit saisonaler Verlauf interpretiert werden. Durch $f(x) = f(x - \lfloor x \rfloor)$, $x \in I\!R$, wird f in (4.7) zu einer periodischen Funktion erweitert. Gemäß Abschnitt 4.2 werden Beobachtungen aus Wochen $t \in T_1$, für die Kalendereffekte anzunehmen sind, nicht berücksichtigt, da an diesen Zeitpunkten Saison- und Kalendereffekte überlagert sind.

In die Schätzung von f in (4.7) an einem Punkt t gehen auch links und rechts von t liegende Beobachtungen ein. Damit auch für t nahe 0 und 1 Nachbarpunkte auf beiden Seiten gegeben sind, wird der durch $\delta(t)$ auf $[0, 1]$ beschränkte Bereich üblicherweise periodisch erweitert durch $((\delta(t) - 1, \delta(t), \delta(t + 1)), (y_t - \hat{m}(t), y_t - \hat{m}(t), y_t - \hat{m}(t)))$. Dann sind bei geeigneter Bandbreitenwahl (s. u.) die Umgebungen von 0 und 1 keine Randbereiche mehr. Für $t \in T_1$ können anschließend durch lokale Interpolation ebenfalls geschätzte Funktionswerte $\hat{f}(t)$ berechnet werden, so dass dann eine jeweilige Kandidatenfunktion für die geschätzte Saisonkomponente durch $\hat{f}(\delta(t))$ für alle Zeitpunkte $t = 1, \ldots, 313$ gegeben ist.

Für die Schätzung von f können nur Verfahren verwendet werden, die auch nicht-äquidistante Abstände zwischen den Beobachtungen erlauben, denn das durch $\delta(t)$ für $t = 1, \ldots, 313$ auf $[0, 1]$ erzeugte Gitter ist nicht regelmäßig. Deswegen können nur das AWS- und das LOK-Verfahren eingesetzt werden. Während letzteres vollständig automatisch ist, muss für AWS eine maximale Bandbreite $h_{\max}$ gewählt werden. Da die Zeitreihe der ursprünglichen Meldefälle (vgl. Abb. 1.1 bzw. 1.2) einen sinusähnlichen Saisonverlauf erwarten lässt, für den bei einer Modellierung durch Splines mindestens zwei abschnittsweise definierte Polynomfunktionen zweiten Grades notwendig sind, sollte $h_{\max}$ mindestens $1/2$ betragen.

Andererseits kann die Bandbreite eine Länge von 1 nicht übersteigen, ohne dass wegen der Periodizität Beobachtungen zur Schätzung an einem Punkt mehrfach eingehen, weswegen $h_{\max}$ nicht größer als 1 gewählt werden sollte. Durch sonstige substanzwissenschaftliche Überlegungen ist ein konkreter Wert nicht zu motivieren, so dass die Bandbreitenwahl adaptiv erfolgen kann: Der Signalschätzer $\hat{f}(t)$ wird für verschiedene Parametereinstellungen von $h_{\max}$ berechnet. Als bestes Ergebnis und möglicher Kandidat wird dasjenige angesehen, das stetig ist und minimalen Abstand zu den Beobachtungen aufweist. Als Maß für den Abstand wird die Summe der quadrierten Residuen

$$RSS(\hat{f}(\delta(t))) = \sum_{t \in T_0} \left(y_t - \hat{f}(\delta(t)) \right)^2 \tag{4.8}$$

verwendet. Gibt es für keine Einstellung eine stetige Lösung, wird derjenige Parameter gewählt, für den die Summe der quadrierten Abstände an Sprungstellen von $\hat{f}$ minimal ist.

Für die Campylobacter-Fälle kann diesen Vorgaben entsprechend $h_{\max}$ zur Schätzung von f in 4.7 einen beliebigen Wert aus $[0.82, 1]$ erhalten. Für jeden solchen Parameter ist die Schätzung $\hat{f}$ identisch; sie ist außerdem stetig und besitzt mit $RSS = 41\ 092$ die kleinste Quadratsumme der Residuen im Vergleich zu Schätzungen mit Parametern $h_{\max} < 0.82$. Der Abstand beträgt für die mit der LOK-Methode erhaltene Kandidatenfunktion 41 543, weswegen diese nicht weiter berücksichtigt wird. Wie auch Abb. 4.7 (oben) zeigt, sind die Unterschiede der Ergebnisse der verschiedenen Schätzverfahren gering. Die geschätzte Saisonkomponente ist damit gegeben durch $\hat{s}(t) = \hat{f}(\delta(t))$, $t = 1, \ldots, 313$, wobei $\hat{f}$ das Ergebnis der Schätzung durch AWS mit $h_{\max} = 0.9$ ist.

Die geschätzte Saisonkomponente zeigt einen sinusähnlichen Verlauf. Mit Beginn des Jahres sinken die Meldefälle leicht bis etwa zur 9. - 11. Woche und steigen danach steil an. Während des Zeitraums mit den meisten Fällen, dem sogenannten Ausbruchszeitraum in den Sommerwochen 26 - 38, in dem die Kurve das Maximum erreicht, flacht der Verlauf ab. Der anschließende Rückgang verläuft bei etwa gleichbleibender Abnahme wiederum bis ins Folgejahr. Die Zeitpunkte mit den wenigsten und meisten Fällen liegen im Bereich der 12. bzw. 32. Woche. Der Unterschied zwischen Minimum und Maximum ist auch abhängig von den übrigen Komponenten, deren Additivität nur im Modell für die durch (4.4) transformierten Beobachtungen gilt. Unter alleiniger Berücksichtigung der geschätzten Trendkomponente sind die erwarteten Extrema für das Jahr 2001 131 und 317; für das Jahr 2006 sind es 201 und 410 Fälle.

Bei den Rotavirus-Fällen resultiert für jedes $h_{\max} \in [0.5, 1]$ dieselbe, nicht

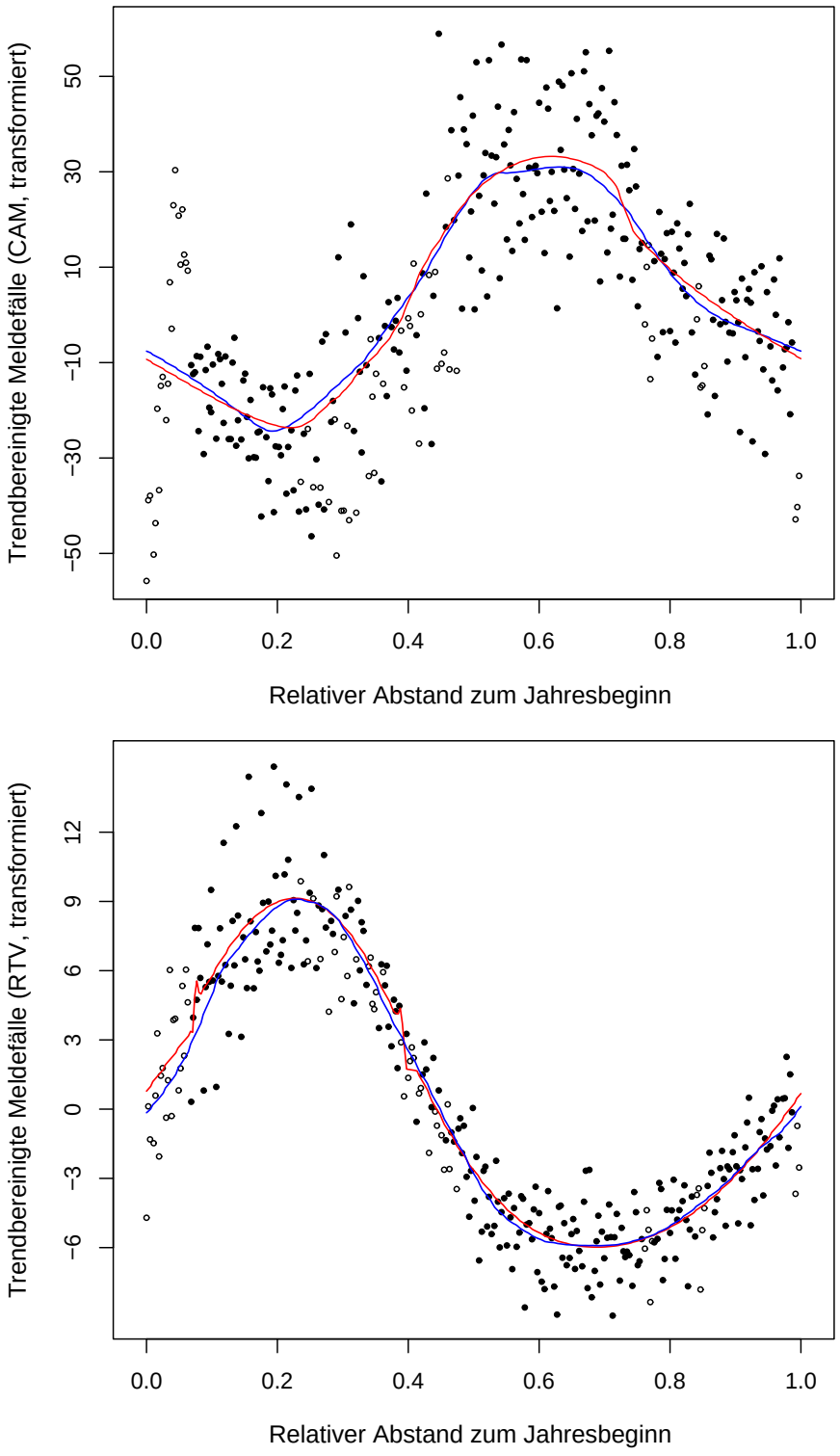

Abb. 4.7: Trendbereinigte Meldefälle $y_t - \hat{m}(t)$ (•), angeordnet entsprechend dem relativen Abstand der Meldewochen zum Jahresbeginn $\delta(t)$, $t = 1, \ldots, 313$, und Kandidaten durch AWS ($h_{max} = 0.9$, rot) und LOK (blau) für die geschätzte Saisonkomponente für CAM (oben) und RTV (unten). Meldefälle aus Wochen mit Kalendereffekten (○) sind dabei nicht berücksichtigt.

stetige Signalschätzung. Daher kann für $h_{\max}$ ein beliebiger Wert entsprechend der Vorgabe verwendet werden. Wegen dieser Unstetigkeit ist die durch LOK erhaltene Kandidatenfunktion die geschätzte Komponente $\hat{s}(t)$, auch wenn der summierte quadratische Abstand zu den Beobachtungen mit $RSS = 821$ etwas größer als der nach AWS (800) ist.

Auch ihr Verlauf ist sinusähnlich und scheint noch regelmäßiger als der der Campylobacter-Fälle. Der Ausbruchszeitraum liegt in der ersten Jahreshälfte, etwa in den Wochen 5 - 18, das Maximum in der 12. Woche. Die weitere Abnahme bis zum Minimum in der etwa 40. Woche und der darauffolgende Zuwachs verlaufen gleichmäßig und mit relativ geringer Krümmung. Damit nimmt der Bereich der Senke einen größeren Zeitraum ein als der um den Höhepunkt. Der maximale Unterschied zwischen den Extrema der geschätzten Saisonkomponente beträgt, nach Rücktransformation, zwischen 301 (Jahr 2002) und 371 (Jahr 2006) Fälle.

Abb. 4.8 zeigt die (nicht transformierten) Beobachtungen mit der Summe der rücktransformierten geschätzten Trend- und Saisonkomponente $g^{-1}(\hat{m}(t) + \hat{s}(t))$. Bei Campylobacter sind insbesondere in den Zeiträumen der Wochen 53 bis 157 und 210 bis etwa 270 deutliche Abweichungen von der Schätzung zu erkennen. Bei Rotavirus sind systematische Unter- bzw. Überschätzungen besonders in den Jahren 2004 und 2006 festzustellen. Mit der Schätzung einer zyklischen Komponente $c(t)$ und einer Kalenderkomponente $k(t)$ wird im folgenden versucht, diese zu vermeiden.

4.5 Zyklische Komponente

Die zyklische Komponente $c(t)$ beschreibt das nach Trend- und Saisonextraktion in den so bereinigten Daten $y_t - \hat{m}(t) - \hat{s}(t)$ verbleibende Signal. Dies kann z. B. aus regelmäßig wiederkehrenden Verläufen mit sich ändernder Periodenlänge bestehen oder lokalen Trends bzw. Änderungen des Niveaus. Um eine Verzerrung durch Kalendereffekte zu vermeiden, werden die Beobachtungen y_t, $t \in T_1$, bei der Schätzung der Komponente nicht beachtet. Das Regressionsmodell ist damit durch

$$y_t - \hat{m}(t) - \hat{s}(t) \;=\; f(t) + \varepsilon_t, \qquad t \in T_0, \tag{4.9}$$

gegeben, wobei die Signalkomponente als die zyklische Komponente interpretiert wird.

Die Güte einer Schätzung $\hat{f}(t)$ wird entsprechend einer datenapproximativen Herangehensweise über die Adäquatheit des Modells für die Fehler beurteilt.

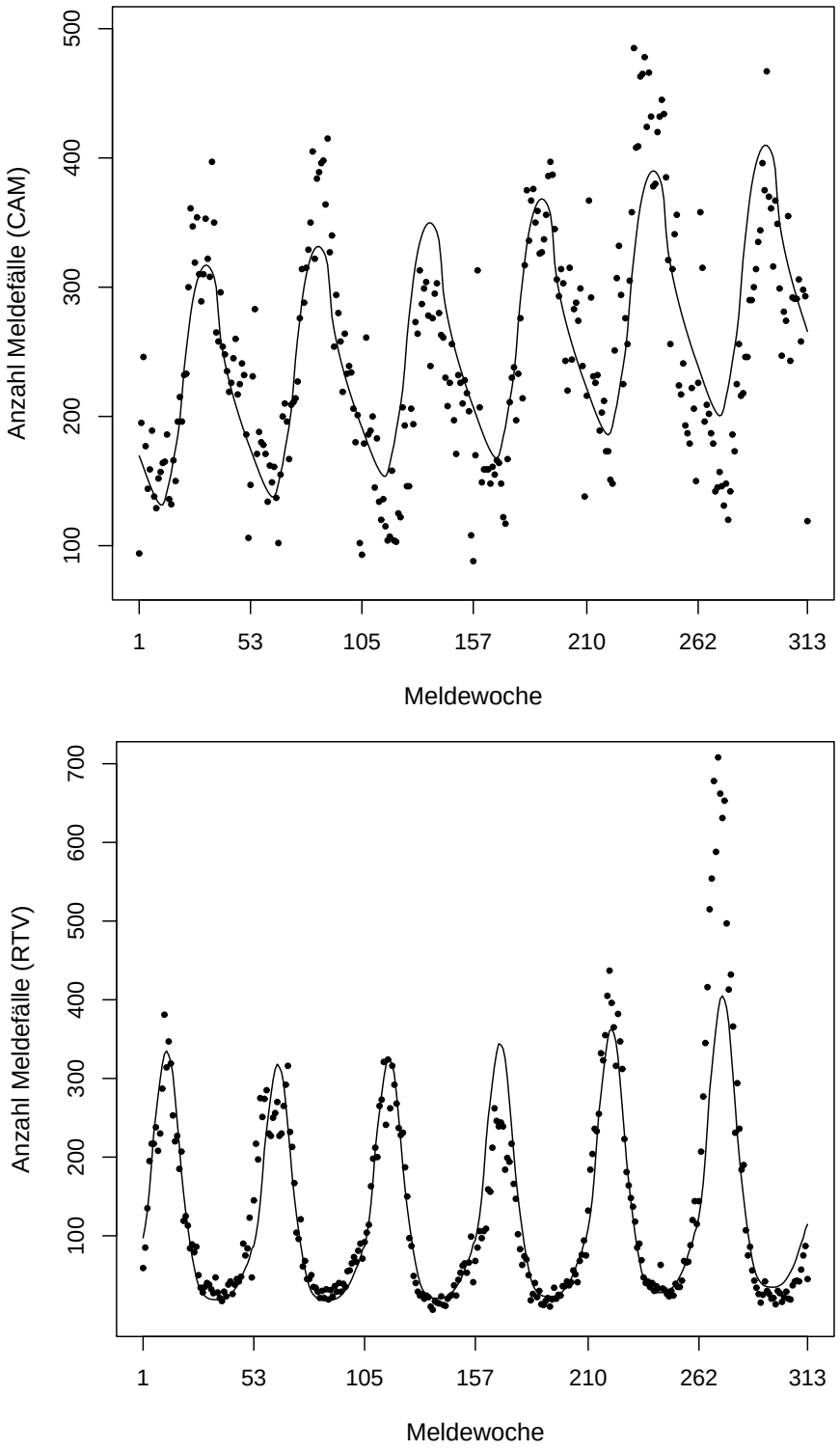

Abb. 4.8: Meldefälle y_t und der nur durch Trend- und Saisonkomponente geschätzte Verlauf $\hat{m}(t) + \hat{s}(t)$ für CAM (oben) und RTV (unten).

Folgen die Residuen $r_t = y_t - \hat{m}(t) - \hat{s}(t) - \hat{c}(t)$ einem stationären Prozess, genauer einem einfachen ARMA(p, q)-Prozess, wird das Modell als adäquat angesehen. $\hat{c}(t)$ ist dabei eine Schätzung $\hat{f}(t)$ in (4.9). Die Überprüfung der Adäquatheit erfolgt jedoch erst nach erfolgter Schätzung der Kalenderkomponente und wird in Abschnitt 4.7 ausführlich erläutert.

Es ist zu erwarten, dass bei Einsatz der zur Verfügung stehenden Verfahren aus Kapitel 2 mehrere Schätzungen $\hat{f}(t)$ im Modell (4.9) existieren, die das *data feature* der stationären Zeitreihe der Residuen aufweisen. Unter solchen Lösungen werden zunächst nur die stetigen beibehalten und als mögliche Kandidaten betrachtet. Weitere inhaltliche Nebenbedingungen wie sie z. B. bei der Saisonkomponente durch die Periodizität gegeben war, sind nicht zu motivieren. Als geschätzte Komponente wird dann die einfachste bzw. am wenigsten komplexe Kandidatenfunktion gewählt. Ein sinnvolles Kriterium ist dafür die totale Variation $tv(.)$. Diese ist für eine an diskreten Stellen $t = 1, \ldots, n$ gegebene Funktion $f(t)$ definiert durch

$$tv(f(t)) = \sum_{t=2}^{n} |f(t) - f(t-1)|. \tag{4.10}$$

Demnach resultiert als Schätzung für das Signal in (4.9) diejenige Funktion, die die Modellierung der Residuen durch einen einfachen ARMA(p, q)-Prozess erlaubt und die gleichzeitig eine geringe totale Variation besitzt. Indem Funktionswerte dann auch für Zeitpunkte $t \in T_1$ berechnet werden, ist die Schätzung $\hat{c}(t)$ für die zyklische Komponente im Dekompositionsmodell (3.11) gegeben für alle $t = 1, \ldots, 313$. Dies erfolgt je nach Methode durch Interpolation oder durch das im folgenden beschriebene iterative Verfahren.

Die an Zeitpunkten $t \in T_1$ liegenden Beobachtungen werden in (4.9) nicht berücksichtigt und damit implizit als fehlende Werte angesehen. Eine durch LOK oder AWS erhaltene Schätzung von $f(t)$ aus (4.9) kann durch lokale Inter- bzw. Extrapolation auch an nicht betrachteten Zeitpunkten bestimmt werden. SSA und STS können jedoch nicht in Situationen mit nicht äquidistanten Beobachtungen bzw. fehlenden Werten angewendet werden. Allerdings ist mit folgendem iterativen Verfahren eine sinnvolle Schätzung von $c(t)$ auch für Wochen $t \in T_1$ möglich: Zunächst werden in einem Initialisierungsschritt die trend- und saisonbereinigten Beobachtungen $y_t - \hat{m}(t) - \hat{s}(t)$ der betreffenden Wochen $t \in T_1$ durch lineare Interpolationen der angrenzenden Beobachtungen mit $t \in T_0$ ersetzt. Beobachtungen, die am Beginn und Ende des Untersuchungszeitraums liegen, werden durch Meldefallzahlen der nächstliegenden Wochen $t \in T_0$ ersetzt. Für die hier analysierten Datensätze werden also y_1, y_2, y_3 durch y_4 und y_{313} durch y_{312} ersetzt. Das Signal der um Trend- und Saisoneffekt bereinigten Zeitreihe wird durch

ein gewähltes nichtparametrisches Verfahren entsprechend (4.9) für $t \in T_0 \cup T_1$ geschätzt und das Ergebnis mit $\tilde{c}^{(1)}(t)$ bezeichnet. Für den folgenden Schritt werden erneut die Beobachtungen y_t aus Wochen mit Kalendereffekten $t \in T_1$ ersetzt, diesmal jedoch durch die soeben geschätzten Werte $\hat{m}(t)+\hat{s}(t)+\tilde{c}^{(1)}(t)$, $t \in T_1$. Für die so konstruierte Zeitreihe wird wiederum nach Bereinigung um die geschätzten Trend- und Saisoneffekte die zyklische Komponente $\tilde{c}^{(2)}(t)$ geschätzt. Diese Schritte werden nun solange wiederholt, bis die in zwei aufeinanderfolgenden Iterationen resultierenden Ergebnisse $\tilde{c}^{(k)}(t)$ und $\tilde{c}^{(k+1)}(t)$ eine zuvor festgelegte Schranke eines geeigneten Abstandsmaßes unterschreiten. Als Abstandsmaß ist beispielsweise die euklidische Distanz

$$d^{(k)} = \sqrt{\sum_{t=1}^{n} \left(\tilde{c}^{(k)}(t) - \tilde{c}^{(k+1)}(t)\right)^2}$$

geeignet, weil zu ihrer Minimierung alle Einzelabstände $\tilde{c}^{(k)}(t) - \tilde{c}^{(k+1)}(t)$, $t = 1, \ldots, n$, klein werden müssen. Für die im folgenden dargestellten Ergebnisse wurde die Iterationsprozedur abgebrochen, sobald $d^{(k)} < \epsilon$, $\epsilon = 0.01$ erreicht wurde. Wenn die Prozedur nicht konvergiert bzw. das Abbruchkriterium nicht erreicht wird, kann mit dem gewählten Verfahren keine Schätzung berechnet werden.

Abb. 4.9 zeigt verschiedene Kandidatenfunktionen für die geschätzte zyklische Komponente am Beispiel der transformierten trend- und saisonbereinigten Meldefälle von Campylobacter. Eine Kandidatenfunktion nach dem STS-Verfahren ist nicht abgebildet, da die zugehörige Iterationsprozedur in diesem Fall nicht konvergiert. Für die gezeigten Kandidaten gilt, dass nach darauffolgender Schätzung der Kalendereffekte die Reihe der Residuen als ARMA(p, q)-Prozess mit $p, q \leq 2$ angesehen werden kann, vgl. Abschnitt 4.7. Die zugehörigen Werte der totalen Variation sowie für weitere Parametereinstellungen sind in Tabelle 4.5 aufgeführt. Da mit der Einstellung $L = 142$ und dem ersten Eigentripel die geringste totale Variation erreicht wird, wird das zugehörige Ergebnis als die geschätzte zyklische Komponente verwendet. Sie ist charakterisiert durch einen sinusähnlichen Verlauf, dessen Periodenlänge über die Zeit abnimmt, dessen Amplitude aber ansteigt. Dies belegt eine zunehmende Differenz zwischen den tatsächlichen und den nur durch Trend und Saison geschätzten Meldefällen. Die Spannweite von $\hat{c}(t)$ beträgt mit 19.9 etwa ein Drittel der von $\hat{s}(t)$, die 56.9 beträgt. Damit wird deutlich, dass die zyklische Komponente einen geringeren Einfluss auf die Meldefälle besitzt als die Saisonkomponente.

Auch bei der Wahl anderer Parameter weisen mit SSA bestimmte Kandidaten geringere Variationen auf: Wenn wenige wie z. B. nur das erste Eigentripel

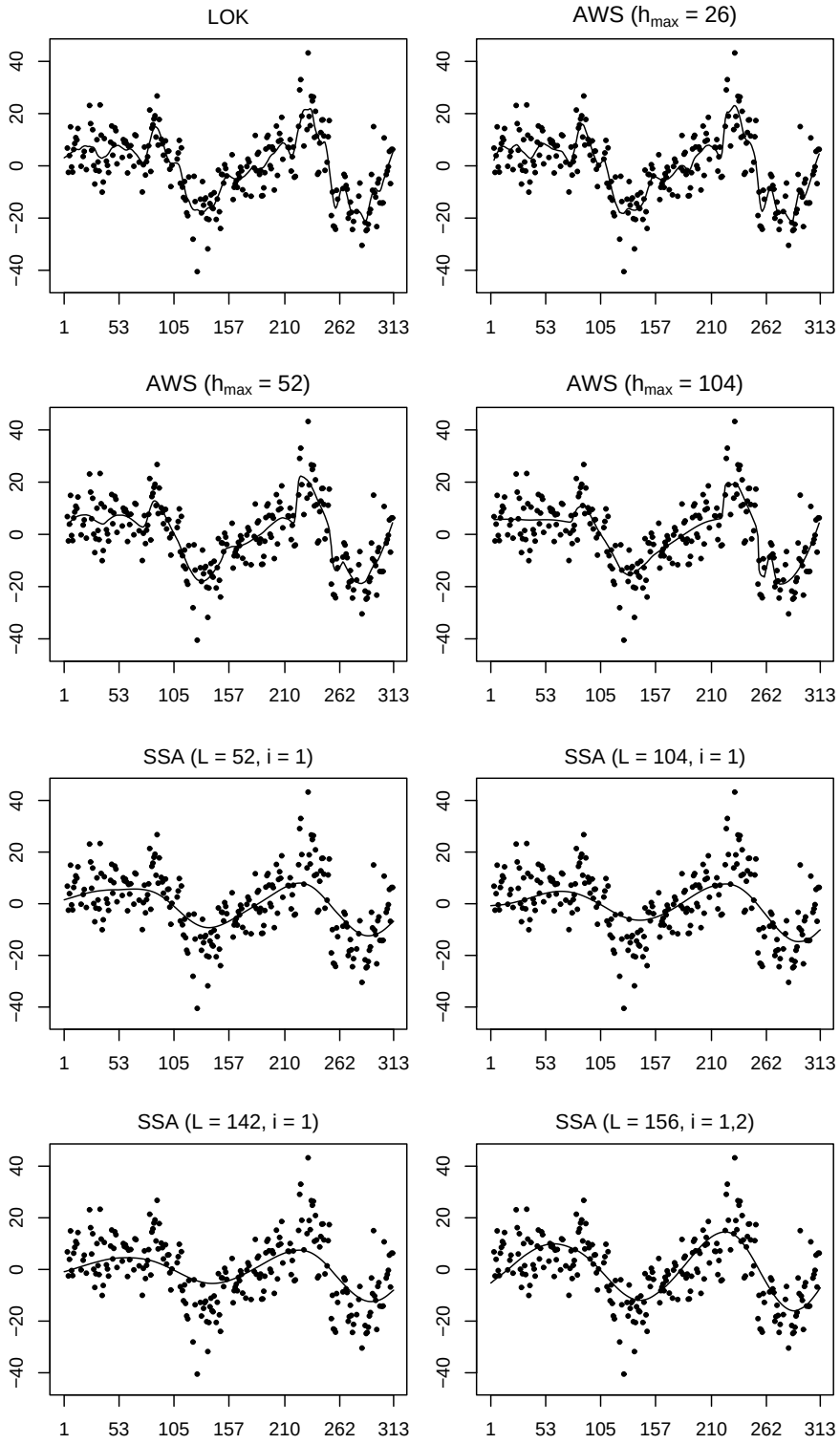

Abb. 4.9: Trend- und saisonbereinigte Meldefälle $y_t - \hat{m}(t) - \hat{s}(t)$, $t \in T_0$, und Kandidaten für die geschätzte zyklische Komponente $\hat{c}(t)$ unter Verwendung verschiedener Verfahren und Parameter für CAM.

Verfahren	Parameter	tv [CAM]	tv [RTV]
LOK		217.1	–
AWS	$h_{\max} = 52$	198.0	37.7
AWS	$h_{\max} = 78$	185.0	36.1
AWS	$h_{\max} = 104$	171.8	36.8
STS		–	–
SSA	L = 52 (i = 1)	62.5	6.4
SSA	L = 104 (i = 1)	57.7	5.6
SSA	L = 142 (i = 1)	52.7	5.4
SSA	L = 153 (i = 1)	55.7	5.4
SSA	L = 156 (i = 1)	56.6	5.4
SSA	L = 104 (i=1,2)	107.9	12.1
SSA	L = 156 (i = 1,2)	103.0	12.6

Tab. 4.5: Totale Variation tv von Kandidatenfunktionen für die geschätzte zyklische Komponente $\hat{c}(t)$ nach verschiedenen Verfahren für CAM und RTV.

bei mittleren bis großen Fensterbreiten verwendet werden, ist die Schätzung eher global, was auch an niedrigen Werten der totalen Variation zu erkennen ist. Wie den Abbildungen entnommen werden kann, sind demgegenüber die Ergebnisse durch LOK oder AWS zu geringe Glättungen mit einem wellenförmigen Verlauf und zahlreichen schwach ausgeprägten Extrema. (Als Ursache sind die bislang nicht modellierten zeitlichen Abhängigkeiten zwischen den Daten möglich, die die lokale Varianzschätzung der Verfahren verzerren, die unabhängige Fehler voraussetzen. Positiv korrelierte Beobachtungen z. B. führen dazu, dass die Varianz lokal unterschätzt wird, was dann zur Wahl schmaler lokaler Bandbreiten und damit zur Unterglättung führt.) STS führt zu keiner stabilen Schätzung in der Iterationsprozedur.

Bei Rotavirus weisen ebenfalls die Kandidatenfunktionen nach SSA mit dem ersten Eigentripel und bei Verwendung großer Fensterbreiten die geringste totale Variation auf. Die beste Einstellung wird für $L = 153$ erreicht. Abb. 4.10 zeigt die damit geschätzte Komponente $\hat{c}(t)$ mit den trend- und saisonbereinigten Meldefällen. Die Spannweite von 2.1 zeigt im Vergleich zu der der Saisonkomponente (15.0), dass die durch zyklische bzw. sonstige Effekte erklärbare Variation im Vergleich zur regelmäßigen periodischen sehr gering ist.

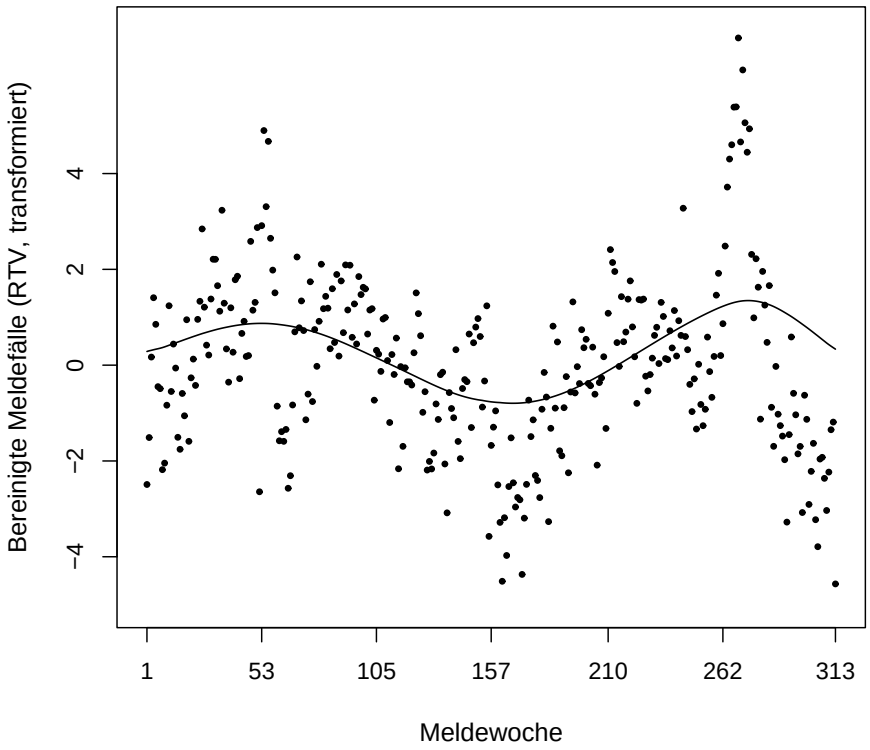

Abb. 4.10: Trend- und saisonbereinigte Meldefälle $y_t - \hat{m}(t) - \hat{s}(t)$, $t \in T_0$ und geschätzte zyklische Komponente $\hat{c}(t)$ nach SSA ($L = 153$, $i = 1$) für RTV.

4.6 Kalendereffekte

Die Komponente $k(t)$ modelliert für die Wochen mit Feiertagen bzw. Schulferien den Unterschied zwischen $\mathrm{E}\,Y_t$ und der Summe der Trend-, Saison- und zyklischer Komponente. Die grundlegende Annahme ist, dass in Wochen $t \in T_1$ kurzfristig wirkende Kalendereffekte auftreten. Weil sie zeitlich begrenzt auftreten, und darüber hinaus teilweise beweglich von Jahr zu Jahr sind, ist ihre Modellierung durch die bisher betrachteten Komponenten nicht möglich und aus Interpretationssicht auch nicht erwünscht. Die möglichen Ursachen für diese zusammenfassend als Kalendereffekte bezeichneten Einflüsse sind vielseitig: Beispielsweise ist denkbar, dass ein Patient wegen eines Feiertages keinen Arzt aufsuchen kann. Klingen die Symptome an den folgenden Tagen ab, wird der Arztbesuch möglicherweise nicht nachgeholt. In Wochen mit Schulferien, insbesondere denen der bedeutenden Feiertage Weihnachten und Ostern, sind nicht nur Arztpraxen häufig geschlossen, sondern auch Teile der Bevölkerung verreist, so dass Erkrankungen am Reiseort behandelt und den lokalen Gesundheitsbehörden nicht gemeldet werden.

Die Vermutung, dass Kalendereffekte auch noch in den jeweils nachfolgenden Wochen feststellbar sind, kann mit den Beobachtungen aus 4.2 nicht gestützt werden. Für alle „normalen“ Wochen (5 Werktage, keine Schulferien) $t \in T_0$ kann daher $k(t) = 0$ gesetzt werden. Da viele Feiertage beweglich sind, muss bei der Schätzung des Kalendereffekts die relative Position der zugehörigen Meldewoche im Jahr berücksichtigt werden. Wie bei der Schätzung der Saisonkomponente werden die Beobachtungen deshalb zeitlich neu, entsprechend ihres Abstandes $\delta(t)$ zum Jahresbeginn, durch (4.6) angeordnet. Die zeitliche Neuordnung führt wiederum zu nicht äquidistanten Beobachtungen, weswegen nur das LOK- und AWS-Verfahren eingesetzt werden können. Mit diesen kann die Kalenderkomponente als Signal der Modellgleichung

$$y_t - \hat{m}(t) - \hat{s}(t) - \hat{c}(t) \;=\; f(\delta(t)) + \varepsilon_t, \qquad t \in T_1, \tag{4.11}$$

geschätzt werden. Dies erfolgt abschnittsweise für die in Abschnitt 4.2 erstellten Gruppen. Die Menge T_1 wird dazu disjunkt in die Teilmengen T_W, T_O und T_R zerlegt, die die Wochen aus der Weihnachts-/Neujahrperiode (beginnend mit der Weihnachtswoche insgesamt vier Wochen), Ostern (die Woche vor und nach Ostern) bzw. die übrigen Wochen mit einzelnen Feiertagen (1. Mai, Christi Himmelfahrt, Pfingsten, Fronleichnam, Tag der Deutschen Einheit, Allerheiligen) enthalten. Die Bestimmung von $\hat{k}(t)$ erfolgt zunächst durch Schätzung von $f(\delta(t))$ aus (4.11) getrennt für diese Teilmengen.

Innerhalb der Weihnachts-/Neujahrperiode ist bei den um trend-, saison- und zyklische Komponente bereinigten Meldefällen von Campylobacter ein deutlicher Anstieg zu beobachten, der zu Beginn und am Ende des Zeitraums abflacht. Dieser wird bei Verwendung des LOK-Verfahrens durch einen weniger stark gekrümmten Verlauf als bei Verwendung von AWS geschätzt. Bei Einsatz dieses Verfahrens wird $h_{\max} = 0.09$ gesetzt, so dass mit der maximalen Bandbreite der gesamte zeitliche Bereich von T_W abgedeckt werden kann. Die resultierende Schätzung und die des LOK-Verfahren unterscheiden sich insgesamt nur geringfügig. Das Ergebnis nach AWS wird als Schätzung $\hat{k}(t)$, $t \in T_W$ verwendet, weil es mit $RSS = 830.2$ eine geringere summierte quadratische Abweichung von den bereinigten Beobachtungen als das nach LOK mit 871.9 aufweist. Bei den Rotavirus-Fällen ist der Anstieg weniger stark und gleichmäßiger ausgeprägt, so dass die Schätzungen einen annähernd linearen Verlauf besitzen. Als geschätzte Kalenderkomponente wird hier ebenfalls das Ergebnis von AWS verwendet, dessen Abstand zu den Beobachtungen mit 58.6 etwas geringer ist als das von LOK mit 59.1. Abb. 4.11 zeigt die bereinigten Meldefälle mit den verschiedenen Schätzungen für beide Krankheiten.

Der Osterzeitraum umfasst die zwei Schulferienwochen vor und nach dem Osterwochenende, die beide auch einen Feiertag (Karfreitag, Ostermontag) enthalten. Da es sich um bewegliche Feiertage handelt, kann neben dem eigentlichen Ostereffekt auch die unterschiedliche zeitliche Position beider Wochen relativ zum Jahresbeginn einen Einfluss ausüben. T_O enthält jedoch nur 10 Wochen; und deswegen sollte die Schätzung einfach sein, um eine Überanpassung zu vermeiden. Abb. 4.12 zeigt die Meldefälle des Osterzeitraums; die jeweils zur Karwoche gehörenden Beobachtungen sind durch schwarze, die zur Woche mit Ostermontag gehörenden Fälle durch weiße Punkte gekennzeichnet. Bei beiden Krankheiten ist nicht erkennbar, ob sich die Kalendereffekte von der ersten zur zweiten Osterwoche ändern oder ob sie von der Position im Jahr abhängen.

Dazu wird die Hypothese, dass die Beobachtungen der ersten und zweiten Osterwoche Realisationen derselben Verteilung sind, für die Campylobacter-Fälle durch den zweiseitigen Test von Mann und Whitney (1947) zum Niveau 5% untersucht. Da das Testergebnis mit einem p-Wert = 0.39 nicht zum Ablehnen der Hypothese führt, ist kein Indiz für eine Änderung des Kalendereffekts gegeben. Ebensowenig ist eine Änderung des Kalendereffekts abhängig von $\delta(t)$ nachweisbar: Im einfachen linearen Modell, in dem der Erwartungswert der Beobachtungen durch $\beta_0 + \beta_1\delta(t)$, $\beta_0, \beta_1 \in \mathbb{R}$, modelliert wird, lässt sich die Hypothese, $H_0 : \beta_1 = 0$, nicht ablehnen (p-Wert = 0.51). Wegen der geringen Zahl an Beobachtungen wird daher auf eine genauere Modellierung des Osterzeitraums verzichtet und

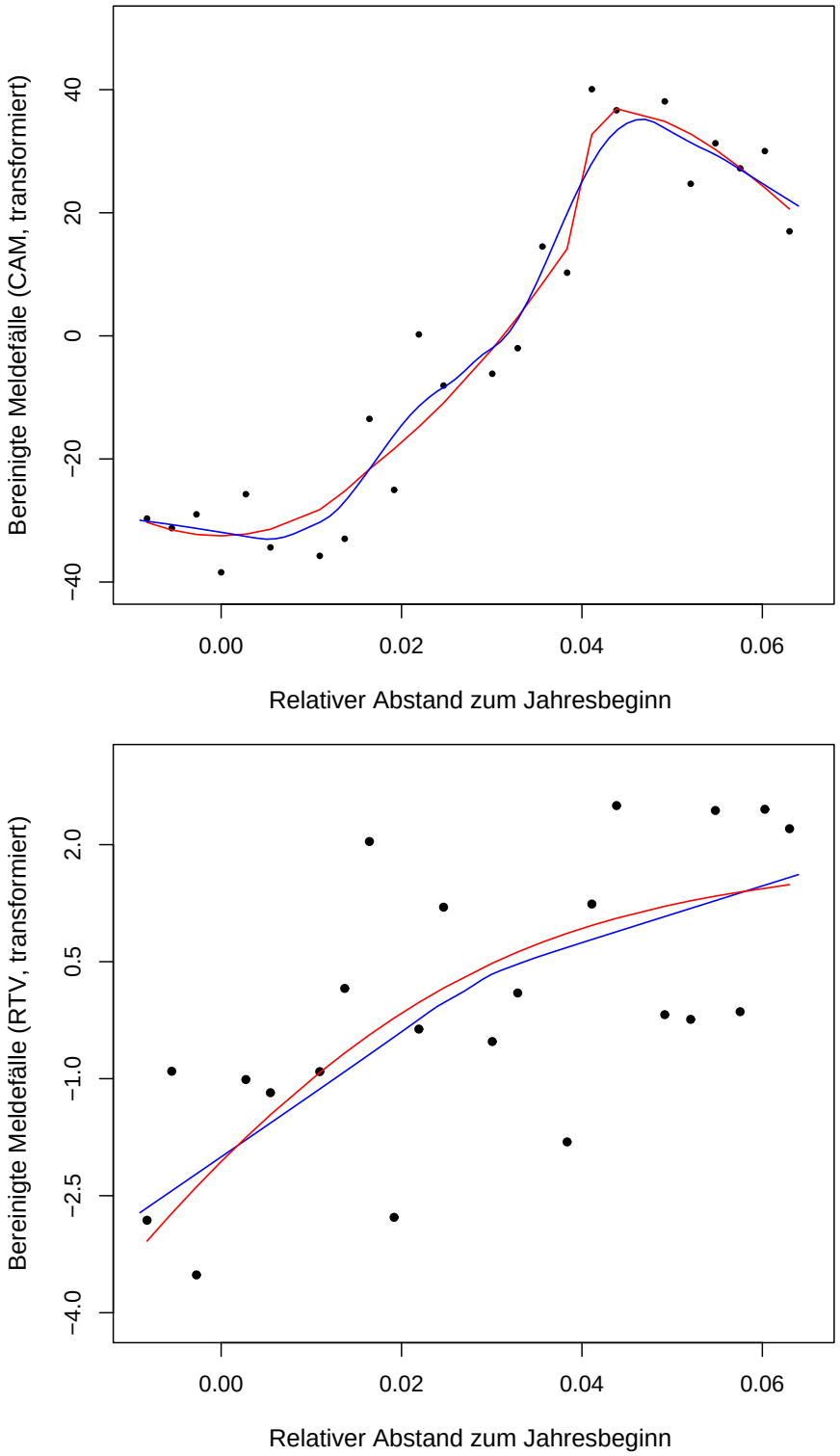

Abb. 4.11: Bereinigte Meldefälle $y_t - \hat{m}(t) - \hat{s}(t) - \hat{c}(t)$ (•) aus dem Zeitraum Weihnachten / Neujahr $t \in T_W$, entsprechend dem relativen Abstand der zugehörigen Meldewochen zum Jahresbeginn, und Kandidatenfunktionen für die geschätzte Kalenderkomponente nach Verwendung des AWS- ($h_{max} = 0.09$, rot) und LOK-(blau) Verfahrens für CAM (oben) und RTV (unten).

der Kalendereffekt für diese Meldewochen nur durch das arithmethische Mittel der $y_t - \hat{m}(t) - \hat{s}(t) - \hat{c}(t)$, $t \in T_O$, geschätzt.

Bei den Rotavirus-Fällen führen entsprechende Tests zu ähnlichen Ergebnissen. Die Hypothese auf Gleichheit in der Verteilung der Beobachtungen der ersten und der zweiten Osterwoche kann keinesfalls abgelehnt werden (p-Wert = 1). Auch die Hypothese, dass der zeitliche Effekt konstant bleibt, kann im linearen Modell nicht widerlegt werden (p-Wert = 0.63). Der Kalendereffekt wird deswegen auch hier nur durch das arithmetische Mittel und damit als Konstante geschätzt. Die resultierenden Schätzer zeigt Abb. 4.12.

Wie aus Abb. 4.13 hervorgeht, ist für die übrigen Wochen mit Kalendereffekten, die nur einen einzelnen Feiertag in der Woche aufweisen, ein Lage-Unterschied im zeitlichen Verlauf, auch zwischen Wochen in Frühling und Herbst, ebenfalls nicht erkennbar. Deutliche Unterschiede zwischen den verschiedenen Feiertagswochen sind nicht vorhanden.

Bei Campylobacter liegen die Mittelwerte der bereinigten Beobachtungen aus den Meldewochen mit Feiertagen zwischen -14.5 (Christi Himmelfahrt/Fronleichnam) und -11 (1. Mai), und die Varianzen sind mit Werten zwischen 36.3 (Pfingsten) und 124.1 (Christi Himmelfahrt/Fronleichnam) zu groß, als dass ein Lageunterschied nachweisbar wäre. Entsprechend wird der Kalendereffekt für die zugehörigen Wochen $t \in T_R$ nur durch das arithmethische Mittel geschätzt. Damit ist die Sitation ähnlich zu der der Meldefälle durch Rotavirus: Auch hier sind keine Unterschiede zwischen den jeweiligen Feiertagswochen auszumachen. Die Mittelwerte liegen zwischen -0.53 (Tag der Einheit) und -1.17 (1. Mai), die Varianzen betragen zwischen 0.24 (Pfingsten) und 2.18 (1. Mai). Für beide Krankheiten zeigt Abb. 4.13 die bereinigten Beobachtungen sowie die resultierenden Schätzer.

Abb. 4.14 zeigt die Beobachtungen der (transformierten) Zeitreihen mit den geschätzten Erwartungswerten $\hat{\mathrm{E}}\, Y_t$, $t = 1, \ldots, 313$, wie sie aus der Summe der einzelnen geschätzten Komponenten bestimmt werden kann. Der Verlauf der Fälle wird bei beiden Krankheiten gut abgebildet. Nur die Meldefälle von Rotavirus werden im letzten Jahr an den Extremstellen weniger gut geschätzt. Zur Beurteilung der Anpassungsgüte kann auch das Bestimmtheitsmaß

$$R^2 = 1 - \frac{\sum_{t=1}^{n} (z_t - \hat{z}_t)^2}{\sum_{t=1}^{n} (z_t - \bar{z}_t)^2} \tag{4.12}$$

herangezogen werden, mit dem der Anteil der durch das Modell erklärten Varianz an der Gesamtvarianz der Beobachtungen berechnet wird. Für beide Krankheiten

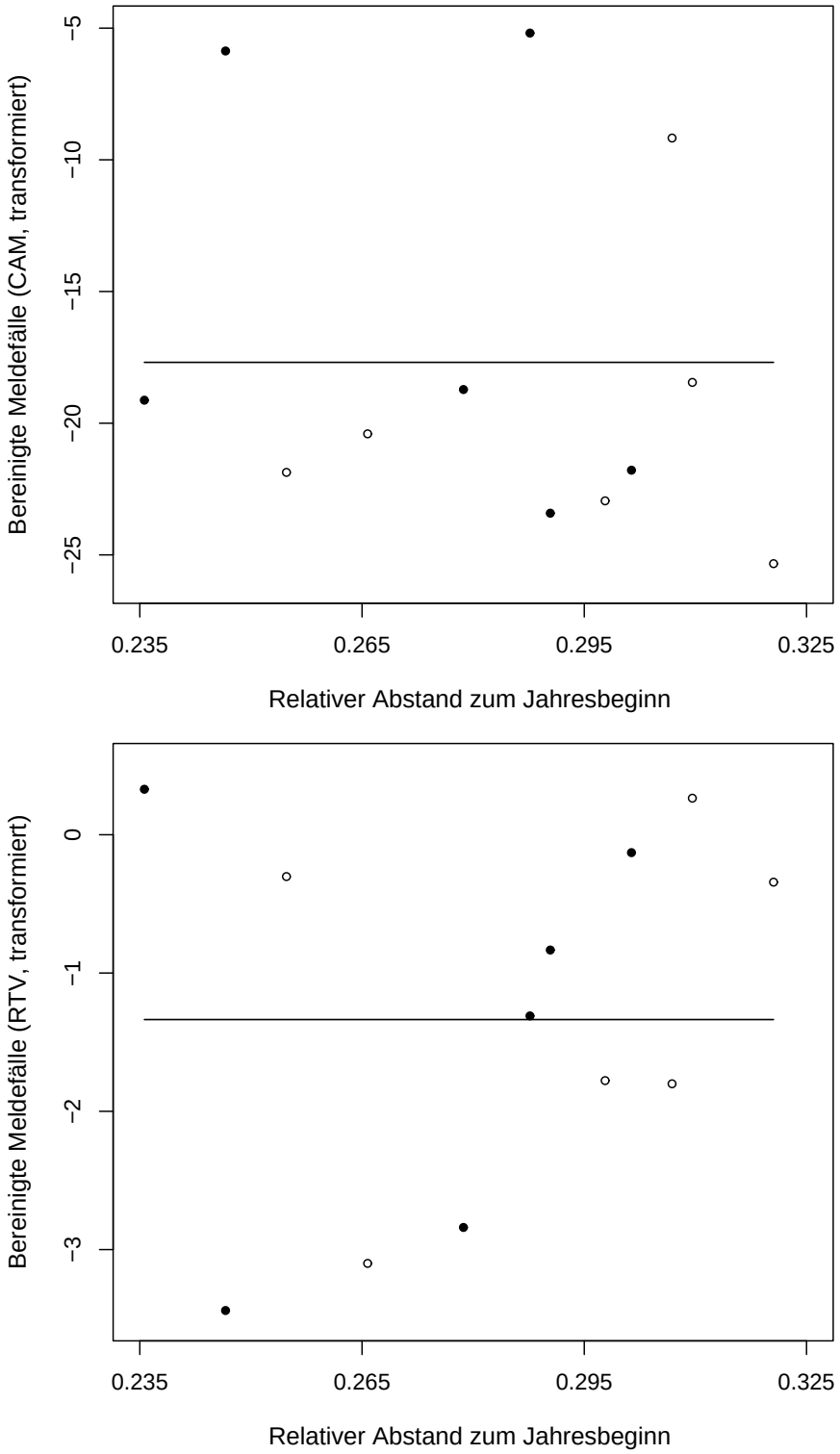

Abb. 4.12: Bereinigte Meldefälle $y_t - \hat{m}(t) - \hat{s}(t) - \hat{c}(t)$ in den Wochen um Ostern $t \in T_O$, entsprechend dem relativen Abstand ihrer zugehörigen Meldewochen zum Jahresbeginn für CAM (oben) und RTV (unten). Meldefälle aus den Wochen vor Ostern sind durch (●), die aus den darauffolgenden durch (○) gekennzeichnet. Die Linie markiert das arithmetische Mittel aller gezeigten Beobachtungen.

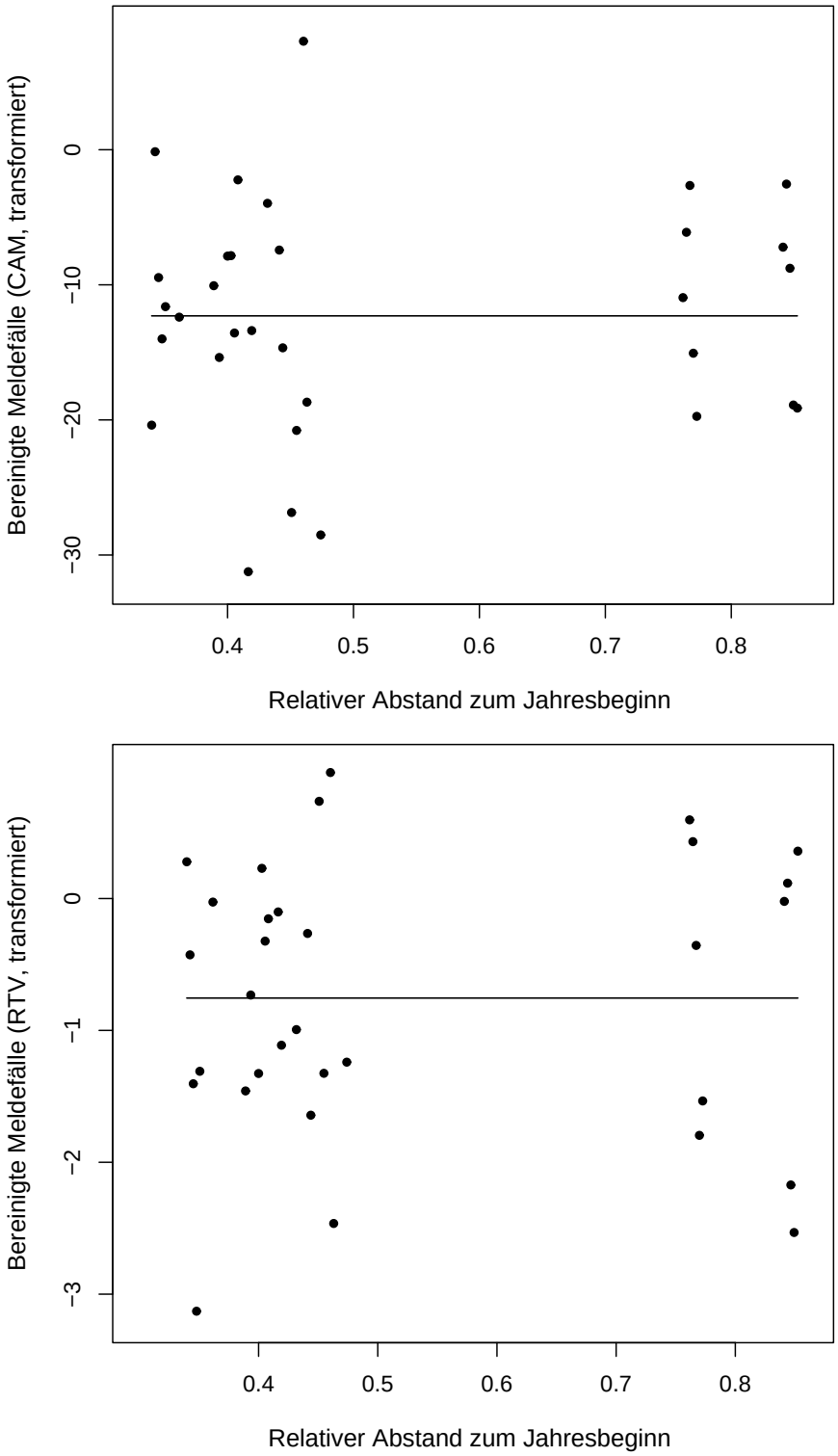

Abb. 4.13: Bereinigte Meldefälle $y_t - \hat{m}(t) - \hat{s}(t) - \hat{c}(t)$ in den Wochen mit einzelnen Feiertagen $t \in T_R$ entsprechend dem relativen Abstand ihrer zugehörigen Meldewoche zum Jahresbeginn für CAM (oben) und RTV (unten). Die Konstante markiert das arithmetische Mittel der gezeigten Beobachtungen.

werden mit $R^2_{\text{CAM}} = 0.880$ und $R^2_{\text{RTV}} = 0.921$ hohe Werte für die Anpassung erreicht. Um Aussagen über die Verteilung der Fehler treffen zu können, werden im folgenden die Residuen untersucht.

4.7 Residuen

Nach Transformation der Beobachtungen und anschließender Subtraktion der geschätzten Trend-, Saison-, zyklischen und Kalenderkomponente erhält man als Zeitreihe der geschätzten Fehler im erweiterten Dekompositionsmodell (3.11) die Residuen

$$r_t \;=\; y_t - \hat{m}(t) - \hat{s}(t) - \hat{c}(t) - \hat{k}(t), \qquad t = 1, \ldots, 313. \tag{4.13}$$

Wie aus Abb. 4.15 deutlich hervorgeht, können diese weder bei Campylobacter noch bei Rotavirus als Weißes Rauschen angesehen werden.

Die Zeitreihen können jedoch durch einfache ARMA(p,q)-Prozesse modelliert werden, d. h. durch solche von geringer Ordnung (p,q). Die Residuen r_t werden dann als Realisierungen von Zufallsvariablen R_t angesehen, für die entsprechend (3.10)

$$R_t - \phi_1 R_{t-1} - \ldots, -\phi_p R_{t-p} \;=\; X_t + \theta_1 X_{t-1} + \ldots, +\theta_q X_{t-q}, \quad \forall t \in \mathbb{Z}, \tag{4.14}$$

erfüllt ist mit $\{X_t\} \sim WN(0, \sigma^2)$.

Um ein sparsames Modell zu finden, werden verschiedene ARMA-Prozesse untersucht, wobei schrittweise Modelle der Ordnungen (1,0), (0,1), (1,1), (2,0), (2,1), (1,2), (2,2) mit zunehmender Komplexität angepasst werden. Das erste Modell, das als adäquat angesehen kann, wird dann als das am wenigsten komplexe auch als endgültig gewählt. Zur Beurteilung der Adäquatheit eines Modells werden die Autokorrelationsfunktion $\rho_{\hat{x}}(h)$ und die partielle Autokorrelationsfunktion $p\rho_{\hat{x}}(h)$ der geschätzten Innovationen

$$\hat{x}_t \;=\; r_t - \hat{\phi}_1 r_{t-1} - \ldots, -\hat{\phi}_p r_{t-p} - \hat{\theta}_1 \hat{x}_{t-1} - \ldots, -\hat{\theta}_q \hat{x}_{t-q}, \quad t = 1, \ldots, 313, \tag{4.15}$$

des ARMA(p, q)-Prozesses untersucht.

Zum Testen der Hypothese, dass $\hat{x}_1, \ldots, \hat{x}_n$ eine Realisation des Weißen Rauschens ist, kann der von Ljung und Box (1978) entwickelte Test eingesetzt werden. Dieser ist eine Weiterentwicklung des Portmanteau-Tests in seiner ursprünglichen Version von Box und Pierce (1970), der die Normalverteilung der

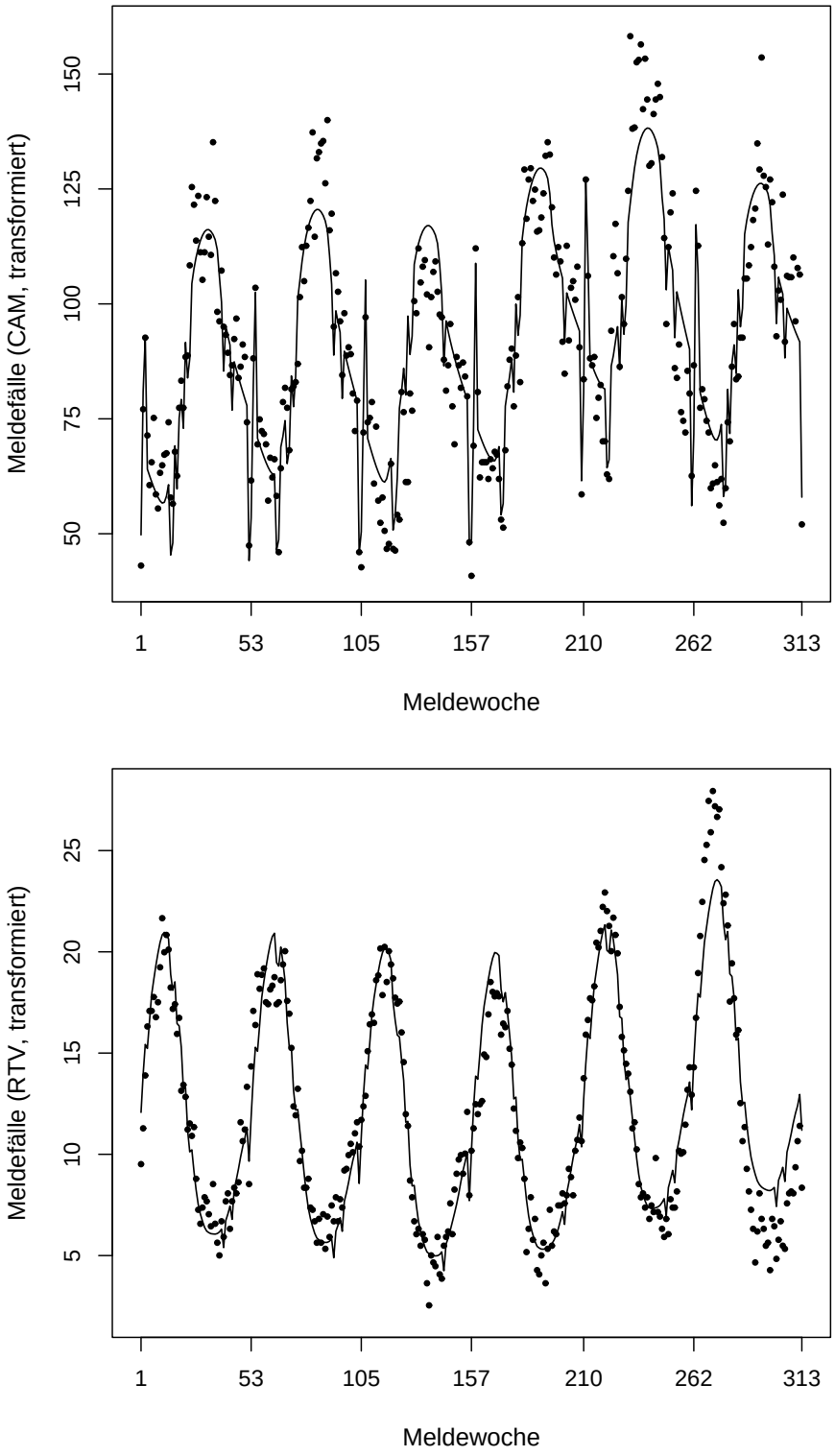

Abb. 4.14: Beobachtete transformierte Meldefälle (•) und die Summe der geschätzten Dekompositionskomponenten $\hat{m}(t) + \hat{s}(t) + \hat{c}(t) + \hat{k}(t)$ (Linie) für CAM (oben) und RTV (unten).

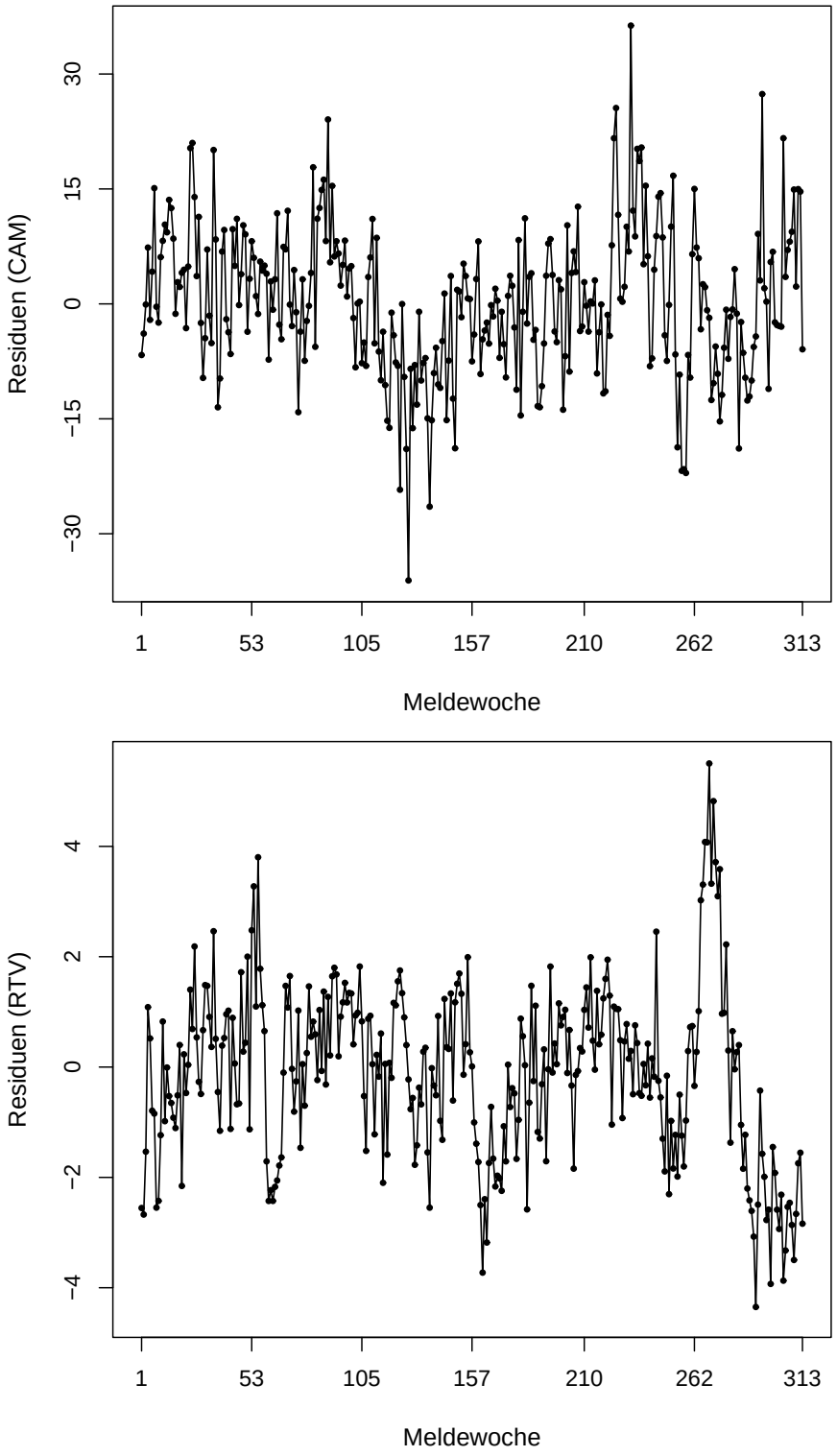

Abb. 4.15: Zeitreihe der Residuen $r_t = y_t - \hat{m}(t) + \hat{s}(t) + \hat{c}(t) + \hat{k}(t)$ für CAM (oben) und RTV (unten).

Werte der Autokorrelationsfunktion beim Weißen Rauschen nutzt, um eine Teststatistik mit χ^2-Verteilung abzuleiten. Die Teststatistik von Ljung und Box ist gegeben durch

$$Q_{\mathrm{LB}}(X,k) \;=\; n(n+2)\sum_{h=1}^{k}\frac{\rho_X^2(h)}{n-h}, \tag{4.16}$$

wobei $\rho_X(h)$ der Wert der Autokorrelationsfunktion einer Zeitreihe $\{X_t\}$ an der Stelle $h = 1, \ldots k < n$ ist. Falls das aktuell betrachtete Modell z. B. ein ARMA(1,0)-Prozess ist, bilden die geschätzten Innovationen $\hat{x}_t = r_t - \hat{\rho} r_{t-1}$ des geschätzten Modells die zu untersuchende Zeitreihe. Für die Autokorrelationsfunktion von Weißem Rauschen X_t, $t = 1, \ldots, n$, ist $\rho_X(h) \sim I\!N(0, 1/n)$, $h < n-1$. Damit gilt unter der Hypothese für ein fest gewähltes Lag $h \in \{1, \ldots, 313\}$

$$P\left(|\rho_{\hat{X}}(h)| \leq q_{1-\alpha/2}\sqrt{\frac{1}{n}}\right) \leq \alpha, \tag{4.17}$$

wobei $q_{1-\alpha/2}$ das $1-\alpha/2$-Quantil der Standardnormalverteilung ist. Folglich ist $Q_{\mathrm{LB}}(X,k)$ unter der Hypothese χ^2_k-verteilt. Sinnvoll ist die Untersuchung nur für Lags bis $k = 52$, da Abhängigkeiten zwischen Residuen, die zeitlich mehr als ein Jahr auseinanderliegen, nicht anzunehmen sind.

Der Test kann analog auch für die partielle Autokorrelationsfunktion eingesetzt werden, um dieselbe Hypothese zu überprüfen. In diesem Fall wird $\rho^2_{\hat{x}}(h)$ durch $p\rho_{\hat{x}}{}^2(h)$ ersetzt. Während mit der Autokorrelationsfunktion insbesondere unberücksichtigte MA-Anteile festgestellt werden können, weicht die partielle Autokorrelationsfunktion insbesondere bei vorhandenen AR-Anteilen von 0 ab.

Kann durch keinen der Tests zum Niveau $\alpha = 0.05$ die Hypothese abgelehnt werden, ist der jeweilige ARMA(p,q)-Prozess ein geeignetes Modell für die Residuen. Demzufolge wird die zugehörige geschätzte zyklische Komponente $\hat{c}(t)$ als adäquat angesehen (vgl. Abschnitt 3.2), und die Schätzung aller Bestandteile des Dekompositionsmodells ist abgeschlossen.

Für die Reihe der Residuen r_t, $t = 1, \ldots, 313$, der Campylobacter-Meldefälle kann ein ARMA(1,0) Prozess angepasst werden. Die geschätzten Parameterwerte sind $\hat{\phi} = 0.520$ und $\hat{\sigma}^2 = 68.65$, wobei σ^2 die Varianz der geschätzten Innovationen $\hat{x}_t$ sind. Diese sind zwar unimodal, aber leicht linksschief und somit insbesondere nicht normalverteilt. Eine graphische Darstellung der Zeitreihe der geschätzten Innovationen enthält Abb. A.3 im Anhang.

Abb. 4.16 (oben, 1. Zeile) zeigt die Autokorrelationsfunktion $\rho_{\hat{x}}(h)$ für $h = 0, \ldots, 52$ sowie Konfidenzbereiche für die jeweiligen Werte. Unter der Hypothese, dass es sich um Weißes Rauschen handelt, liegt der Wert der Autokorrelationsfunktion für ein beliebiges festes Lag h mit Wahrscheinlichkeit 0.05 außerhalb

der eingezeichneten Grenzen. Dies kann als grobe Richtlinie für eine visuelle Beurteilung genutzt werden. Werden 52 Lags betrachtet, ist insgesamt etwa eine einmalige Überschreitung der Grenzen zu erwarten. Die Autokorrelationsfunktion steht also im Einklang mit der Annahme, dass die Residuen Weißes Rauschen sind. Diese Hypothese kann auch durch den Ljung-Box-Test nicht widerlegt werden, der zugehörige p-Wert beträgt 0.315. Wird die Hypothese durch denselben Test mit den Werten der partiellen Autokorrelationsfunktion, s. Abb. 4.16 (oben, 2. Zeile), durchgeführt, resultiert als p-Wert 0.614. Damit ist das Kriterium für die Adäquatheit des ARMA(1,0) Prozesses erfüllt.

Für die Rotavirus-Fälle kann ebenfalls ein ARMA(1,0)-Prozesses an r_t, $t = 1, \ldots, 313$ angepasst werden. Die geschätzten Parameterwerte sind $\hat{\phi} = 0.713$, und $\hat{\sigma}^2 = 1.246$. Abb. 4.16 zeigt die Autokorrelationsfunktion mit ähnlichen Eigenschaften wie der bei Campylobacter. Die Ljung-Box-Statistik liefert für die Autokorrelationsfunktion den p-Wert 0.130 und für die partielle Autokorrelationsfunktion den p-Wert 0.092. Damit ist auch hier das Kriterium für die Adäquatheit erfüllt, und die Zeitreihe der geschätzten Innovationen kann als Realisation Weißen Rauschens angesehen werden, vgl. Abb. A.3 im Anhang.

Anmerkung: Da die Werte der Auto- wie auch der partiellen Autokorrelationsfunktion für $h = 1$ die Konfidenzbereiche unter der Nullhypothese überschreiten, könnte unter diesem Aspekt ein ARMA(1,0)-Prozess als Modell der Zeitreihe der Rotavirus-Residuen nicht als geeignet angesehen werden. Ein ARMA(1,1)-Prozess würde Korrelogramme der geschätzten Innovationen liefern, die besser zur Annahme Weißen Rauschens passen würden. Im Anhang A.2 werden die Ergebnisse dieser Schätzung dargestellt. Das Verständnis des gesamten Verlaufs der Zeitreihe würde damit jedoch nur unwesentlich verändert, weil die Schätzungen der deterministischen Komponenten nicht verändert würden. Aus diesem Grund kann auf eine genauere Modellierung verzichtet werden. Aus demselben Grund wurde zur Beurteilung der Adäquatheit lediglich die Ljung-Box-Statistik statt eines erweiterten, verfeinerten Kriteriums gewählt.

4.8 Schätzung des allgemeinen Verlaufs

Durch die Summe der in den Abschnitten 4.3 - 4.6 geschätzten Komponenten und Umkehrung der in 4.1 ermittelten Transformationsfunktion wird die Dekomposition rückgängig gemacht und der Erwartungswert der Meldefälle durch

$$\hat{z}_t \;=\; g^{-1}\left(\hat{m}(t) + \hat{s}(t) + \hat{c}(t) + \hat{k}(t)\right), \quad t = 1, \ldots, 313, \tag{4.18}$$

geschätzt. Die in 4.7 geschätzten Innovationen des Fehlers werden dabei nicht berücksichtigt. Da die geschätzten ARMA-Prozesse ohne weitere Bedingungen

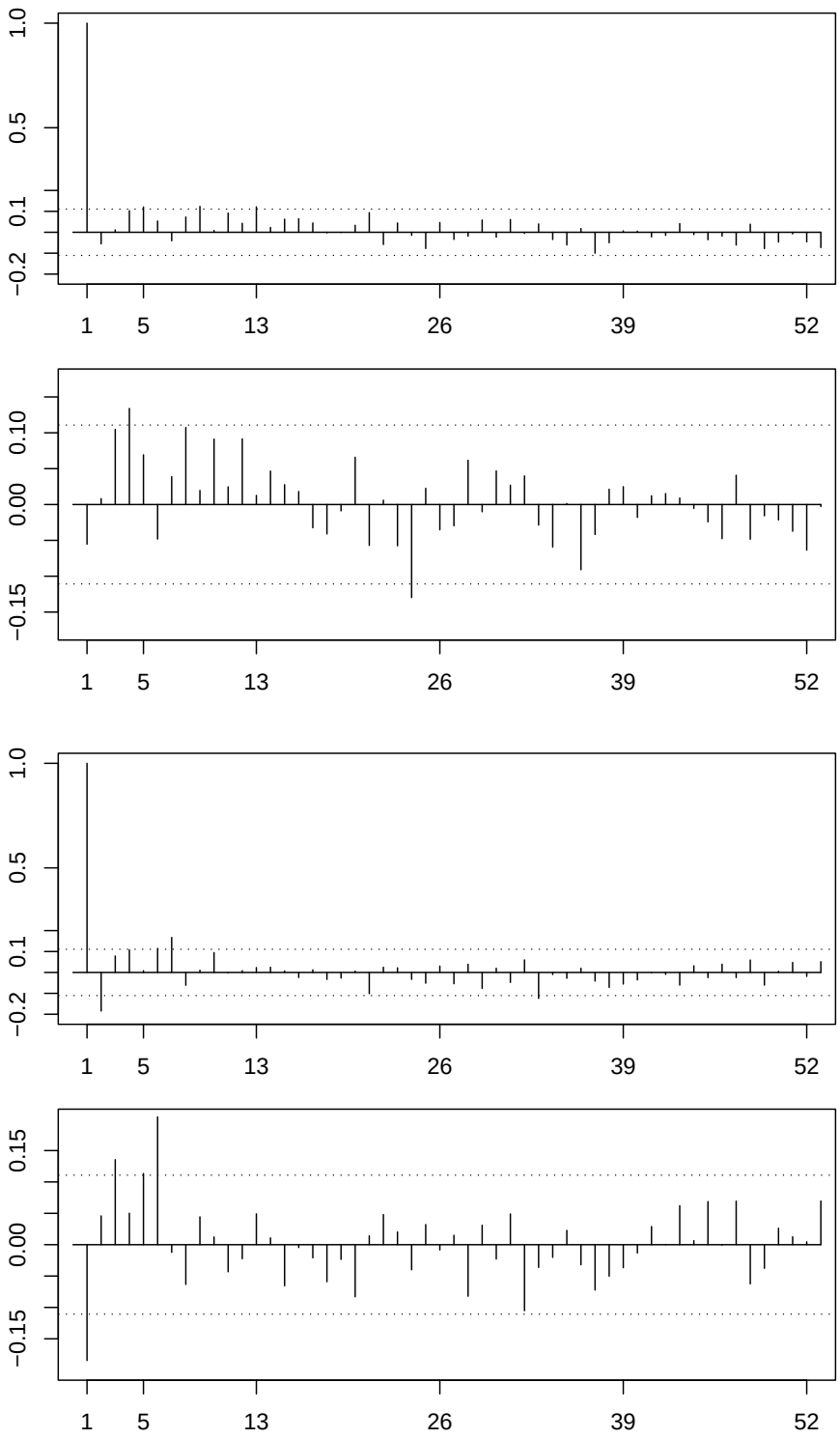

Abb. 4.16: Autokorrelationsfunktion (oben) und partielle Autokorrelationsfunktion (unten) mit 5% Konfidenzgrenzen der geschätzten Innovationen $\hat{x}(t)$ nach jeweiliger Anpassung eines ARMA(1,0)-Prozesses für die Residuen für CAM (Zeilen 1, 2) und RTV (Zeilen 3, 4).

in Form geschätzter Innovationen an jeder Stelle Erwartungswert 0 besitzen, genügen die deterministischen Komponenten zur Darstellung des globalen Verlaufs. Alternativ könnte der Erwartungswert für jede Stelle t geschätzt werden, indem zusätzlich auf die vorhergehenden geschätzten Innovationen bedingt wird. Die sich ergebenden Änderungen wären jedoch nicht wesentlich und würden die Deutung des ansonsten glatten Gesamtverlaufs erschweren ohne sie um einen epidemiologisch interpretierbaren Aspekt zu erweitern.

Für die beiden untersuchten Krankheiten werden mit $R^2 = 0.858$ (CAM) bzw. $R^2 = 0.882$ (RTV) hohe Werte erreicht, die eine grundsätzlich hohe und damit gute Anpassung des Modells an die Daten belegen. Diese ist auch in Abb. 4.17 erkennbar, in denen die beobachteten Meldehäufigkeiten mit den geschätzten Erwartungswerten gezeigt werden. Mit den 0.025- und 0.975-Quantilen der geschätzten Innovationen in den angepassten ARMA-Prozessen können punktweise Vertrauensbereiche zum Niveau 0.95 bestimmt werden. Unter der Annahme der Richtigkeit des Modells enthält ein solches Intervall mit Wahrscheinlichkeit 0.95 die Meldefallhäufigkeit einer Woche, wenn die Realisation des ARMA-Prozesses in der vorhergehenden Woche genau 0 betrug bzw. die Abhängigkeit der zufälligen Fehler generell vernachlässigt wird. Der Vertrauensbereich berücksichtigt damit nur die globalen Eigenschaften des konstanten Erwartungswerts und der konstanten Varianz und wird ohne Bedingung auf bereits bekannte zurückliegende Beobachtungen bestimmt.

4.9 Eine automatisierte Prozedur

Das Vorgehen der in den Abschnitten 4.1 - 4.7 am Beispiel der Fälle von Campylobacter und Rotavirus entwickelten Einzelschritte zeigt, dass das gesamte Verfahren auch durch eine automatisierte Prozedur formuliert werden kann. In diesem Abschnitt wird untersucht, ob eine solche Prozedur für die Zeitreihe der Salmonellen-Fälle zu einer sinnvollen Schätzung führt. Entsprechend der obigen Vorgehensweise werden die folgenden Schritte in dieser Reihenfolge durchlaufen:

- Vorläufige Schätzung des Signals $f(t)$ im nichtparametrischen Regressionsmodell $z_t = f(t) + \varepsilon_t$, $t = 1, \ldots, 313$.
- Wahl einer geeigneten Transformationsfunktion: Als zugehöriger Parameter wird derjenige Wert λ_0 gewählt, für den der Unterschied U^λ der Varianzen der Residuen $\hat{\varepsilon}_t = y_t - \hat{f}(t)$ minimal ist.
- Identifizierung von Meldewochen mit Kalendereffekten T_1: Die Weihnachts-/Neujahrperiode T_W wird in jedem Fall als Zeitraum mit besonderen ka-

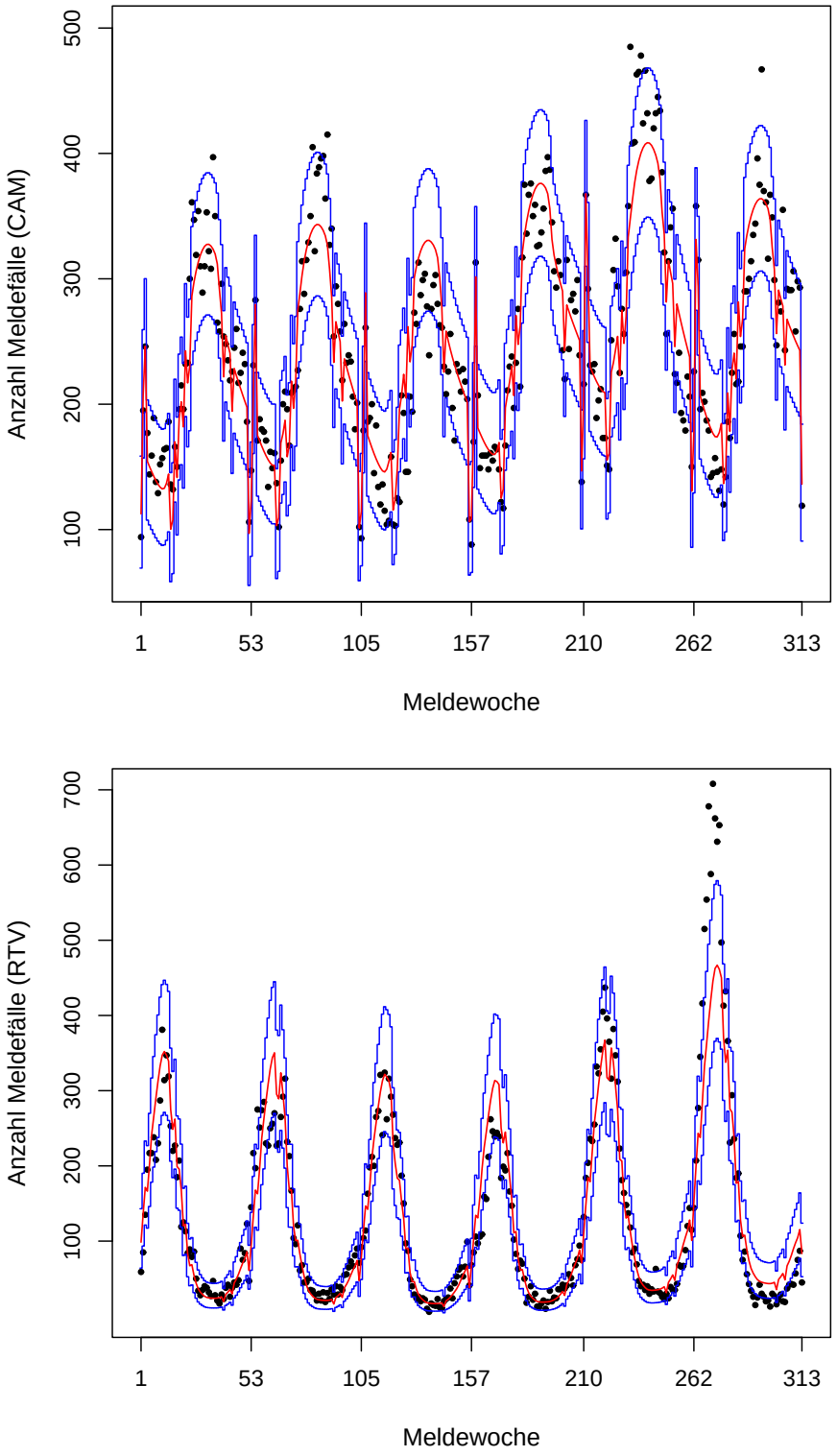

Abb. 4.17: Geschätzter Verlauf (rot) und 95%-Vertrauensbereiche (blau) für die Zahl der beobachteten (nicht transformierten) Meldefälle für CAM (oben) und RTV (unten).

lendarischen Effekten behandelt. Außerdem werden als solche die Wochen um Ostern T_O und die Menge der Wochen mit einem einzelnen Feiertag T_R behandelt, wenn die Residuen $z_t - \hat{f}(t)$ in diesen Wochen signifikant von 0 nach unten abweichen. Dies wird jeweils durch den einseitigen Wilcoxon-Vorzeichen-Rangtest zum Niveau 5% getestet.

- Transformation der Zeitreihe $y_t = g_{\lambda_0}(z_t)$.

- Schätzung der Trendkomponente $m(t)$ von y_t, $t = 1, \ldots, 313$, durch das erste Eigentripel einer SSA-Zerlegung. Die Fensterbreite L wird so gewählt, dass das Kriterium γ aus (3.2) minimiert wird und damit die resultierende Schätzung die geringste Krümmung aufweist.

- Schätzung der Saisonkomponente $s(t)$ aus den trendbereinigten Beobachtungen $y_t - \hat{m}(t)$, $t \in T_0$, ohne Berücksichtigung von Wochen mit Kalendereffekten T_1. Die geschätzte Komponente ist dasjenige Resultat unter den stetigen Ergebnissen durch AWS und LOK, das den kleinsten quadratischen Abstand zu den Beobachtungen aufweist.

- Schätzung der zyklischen Komponente $c(t)$ aus den trend- und saisonbereinigten Beobachtungen $y_t - \hat{m}(t) - \hat{s}(t)$, $t \in T_0$. Schätzungen durch AWS, LOK, STS und SSA werden als adäquat und damit als mögliche Kandidaten angesehen, wenn nach anschließender Schätzung der Kalendereffekte (s. u.) für die dann verbleibenden Residuen ein ARMA(p, q)- Prozess, $p, q \leq 2$, angepasst werden kann. Die geschätzte zyklische Komponente $\hat{c}(t)$ ist diejenige Funktion, die unter den Kandidaten die geringste totale Variation tv besitzt. Zur Berechnung von Kandidatenfunktionen wird das in Abschnitt 4.5 vorgestellte iterative Verfahren verwendet.

- Schätzung der Kalendereffekte $k(t)$ aus den trend-, saison- und zyklikbereinigten Beobachtungen $y_t - \hat{m}(t) - \hat{s}(t) - \hat{c}(t)$, $t \in T_1$. In der Weihnachts-/Neujahrperiode sind die Schätzungen nach AWS und LOK mögliche Kandidaten; das stetige Ergebnis mit der geringeren quadratischen Abweichung darunter ist die geschätzte Komponente.

  Sofern $T_O \subset T_1$ bzw. $T_R \subset T_1$, wird die Kalenderkomponente in der Osterzeit und den übrigen Wochen jeweils als Konstante durch das arithmetische Mittel geschätzt.

- Modellierung der um alle deterministischen Komponenten bereinigten Beobachtungen $y_t - \hat{m}(t) - \hat{s}(t) - \hat{c}(t) - \hat{k}(t)$, $t = 1, \ldots, 313$, durch einen ARMA(p, q)-Prozess. Derjenige Prozess mit der kleinsten Ordnung, für den

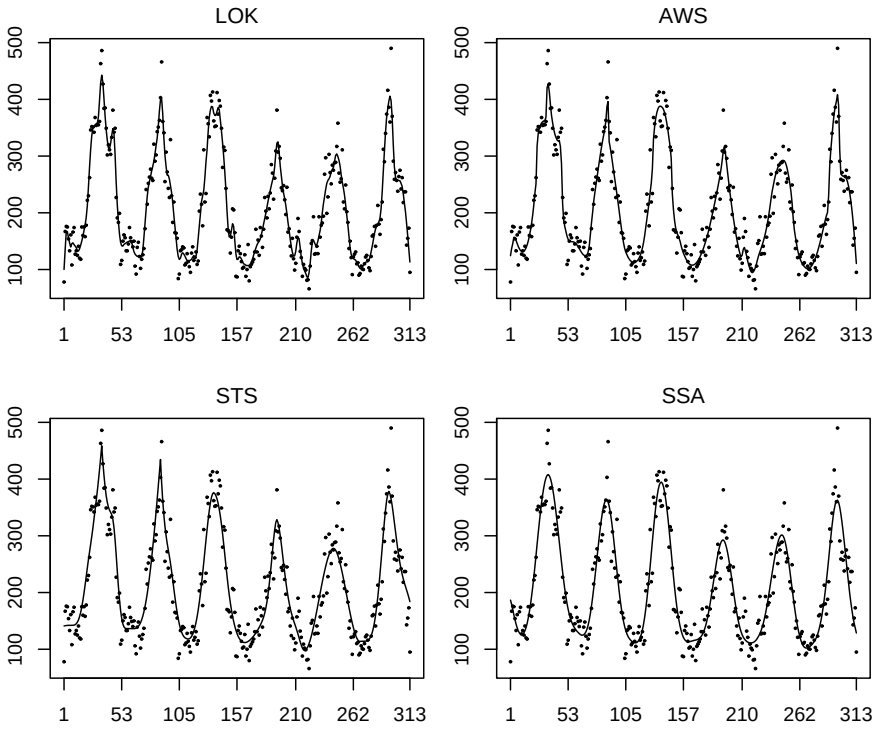

Abb. 4.18: Meldefälle von Salmonellose (SAL) mit geschätztem Signal durch LOK, AWS ($h_{max} = 26$), STS und SSA ($L = 52$, 5 Eigentripel).

die damit geschätzten Innovationen als Weißes Rauschen angesehen werden können, stellt das geschätzte Modell für die Residuen dar.

Nur im ersten Schritt sollte die Auswahl einer geeigneten Signalschätzung für die untransformierten Daten nicht automatisch mittels Kennzahlen sondern auch durch visuelle Beurteilung erfolgen. Ein allgemeingültiges, eindimensionales und damit eindeutiges Gütekriterium für die Signalschätzung kann nämlich nicht aufgestellt werden. Der einfache Grund ist, dass die verschiedenen Anforderungen wie Glattheit der Kurve, gute Anpassung in den Spitzen und Senken sowie möglichst geringe Abweichung von den Beobachtungen einerseits nicht allgemein definiert sind und andererseits ihre jeweilige Gewichtung innerhalb eines einzelnen Kriteriums subjektiv wäre. Daher ist es vertretbar, die Entscheidung für eine bestimmte Signalschätzung durch individuellen Vergleich der verschiedenen Ergebnisse zu treffen.

Am Beispiel der Meldefälle der Salmonellose wird das so vereinheitlichte Vorgehen illustriert. Zunächst werden Schätzungen des allgemeinen Signals betrachtet. Wie aus Abb. 4.18 hervorgeht, weist der geschätzte Verlauf durch LOK

an einigen Stellen zusätzliche Extrema und dadurch erhöhte Variation auf, die überangepasst erscheint. Das Ergebnis mit AWS zeigt einen insgesamt glatteren Verlauf, wenn auch an einigen Stellen noch plötzliche Steigungsänderungen auftreten. Die Resultate von STS und SSA (L=52, $i = 1, \ldots, 5$) sind hingegen sowohl genügend glatt wie auch gut bzgl. der Anpassung. Für die Ermittlung des geeigneten Transformationsparameters wird hier das Ergebnis der STS-Schätzung verwendet, da dieses die beste Anpassung in den Bereichen mit Extrema ermöglicht. Für die Menge der Transformationsparameter $\lambda \in [0, 1]$ gilt, dass der beste Wert U_λ gemäß (4.3) für $\lambda_0 = 0.44$ erreicht wird, vgl. Abb. A.4 (Anhang). Für eine Box-Cox-Transformation (3.12) wird also bei Wahl dieses Parameters der maximale Unterschied der Varianzen der Residuen, deren Menge disjunkt zerlegt wird, minimiert.

Anmerkung: Erfolgt die Schätzung von f nicht durch STS sondern durch SSA, resultiert in diesem Fall mit $\lambda_0 = 0.46$ ein ähnlicher Transformationsparameter.

Die Residuen $\hat{\varepsilon}_t = y_t - \hat{f}(t)$ sind in beiden Zeiträumen mit möglichen Kalendereffekten $t \in T_O$ und $t \in T_R$ signifikant von 0 verschieden. Wird die Hypothese der Gleichheit mit dem Wilcoxon-Vorzeichen-Rangtest getestet, resultiert in beiden Fällen der p-Wert 0. Die auffällige Lage dieser Beobachtungen kann auch Abb. A.5 (oben) im Anhang entnommen werden. Folglich werden die zugehörigen Beobachtungen bei der Schätzung von Saison- und zyklischer Komponente nicht berücksichtigt und später gesondert modelliert. Die auf diese Zeiträume direkt folgenden Residuen zeigen im Gegensatz hierzu kein auffälliges Verhalten, vgl. Abb. A.5 (unten). Die einseitige Hypothese, dass diese Residuen den Erwartungswert kleiner oder gleich 0 besitzen, kann nicht widerlegt werden (p-Wert = 0.46).

Die geschätzte Trendkomponente ist das Resultat einer SSA-Zerlegung, bei der nur aus dem ersten Eigentripel eine Zeitreihe rekonstruiert wird. Bei Wahl der Fensterbreite $L = 150$ ist die Krümmung geringer im Vergleich zu der anderer Kandidatenfunktionen, so dass das Ergebnis durch $L = 150$, $i = 1$ die geschätzte Trendkomponente ist. In Tabelle 4.6 ist die Glattheit bei Wahl verschiedener Fensterbreiten aufgeführt. Eine Abbildung, die die zugehörigen Ergebnisse der SSA-Zerlegung zeigt, enthält Abb. A.6 im Anhang. Die Rücktransformation von $\hat{m}(t)$ auf den ursprünglichen Maßstab zeigt einen langsamen Rückgang um etwa 58 Meldefälle in den ersten Jahren, der sich jedoch 2006 nicht fortsetzt und dort in einen schwachen Anstieg von etwa 5 Fällen in diesem Jahr übergeht.

L	104	130	150	156
$\gamma(\hat{m}(t))$	$1.9 \cdot 10^{-6}$	$7.5 \cdot 10^{-6}$	$1.4 \cdot 10^{-6}$	$2.0 \cdot 10^{-6}$

Tab. 4.6: Krümmung γ von Kandidaten für die geschätzte Trendkomponente durch SSA mit verschiedenen Fensterbreiten L unter Verwendung des jeweils ersten Eigentripels für SAL.

Für die trendbereinigten Fälle liefern sowohl AWS wie auch LOK stetige Kandidatenfunktionen für die geschätzte Saisonkomponente. Die bessere Anpassung an die Daten ermöglicht jedoch AWS, hier ist der summierte quadratische Abstand zu den Beobachtungen mit 847.5 etwas geringer als bei LOK (867.9). Für die Bandbreite $h_{\max}$ kann ein beliebiger Wert zwischen 0.5 und 1 gewählt werden, da in allen Fällen dasselbe Ergebnis resultiert. Abb. 4.19 ermöglicht einen grafischen Vergleich beider Resultate. Die geschätzte Saisonkomponente ist also das Ergebnis nach AWS und zeigt ähnlich wie bei den bereits betrachteten Krankheiten einen sinusähnlichen Verlauf, der jedoch vor und nach dem Maximum nur schwach gekrümmt ist. Die geschätzten Wochen mit den wenigsten und meisten zu erwartenden Fällen sind demnach die jeweils 9. bzw. 36. Woche eines Jahres. Die erwartete Differenz zwischen den jährlichen Maxima und Minima nimmt unter Berücksichtigung der geschätzten Trendkomponente von etwa 266 im Jahr 2001 auf etwa 228 im Jahr 2006 ab. Dies kann auch dem nur durch Trend- und Saisonkomponente geschätzten Verlauf nach Umkehrung der Transformation entnommen werden, wie er im Anhang, Abb. A.7, dargestellt ist.

Bei der Schätzung der zyklischen Komponente ist erneut festzustellen, dass die Kandidatenfunktionen mit der geringsten totalen Variation, die im Sinne des in Abschnitt 4.5 aufgestellten Kriteriums adäquat sind, mit SSA erreicht werden. Wie aus Tabelle 4.7 hervorgeht, liefern LOK und AWS deutlich variationsreichere Kandidaten; bei Einsatz der STS-Methode konvergiert das iterative Schätzverfahren nicht. Die geschätzte Komponente ist damit die Rekonstruktion aus dem ersten Eigentripel einer SSA-Zerlegung mit Fensterbreite $L = 91$. Der Verlauf ist im Gegensatz zu denen der zuvor betrachteten Krankheiten nicht regelmäßig und weist mehrere lokale Extrema auf, die aber nur schwach ausgeprägt sind. Die zugehörige Abbildung A.8 ist im Anhang enthalten.

Zur Schätzung des Kalendereffekts wird zunächst der Zeitraum T_W der Weihnachts-/Neujahrperiode untersucht. Hier weist die Kandidatenfunktion nach AWS-Schätzung einen größeren Abstand (RSS = 30.8) zu den Beobachtungen auf als die nach LOK (RSS = 23.9). Weil diese wegen der vielen Extrema und

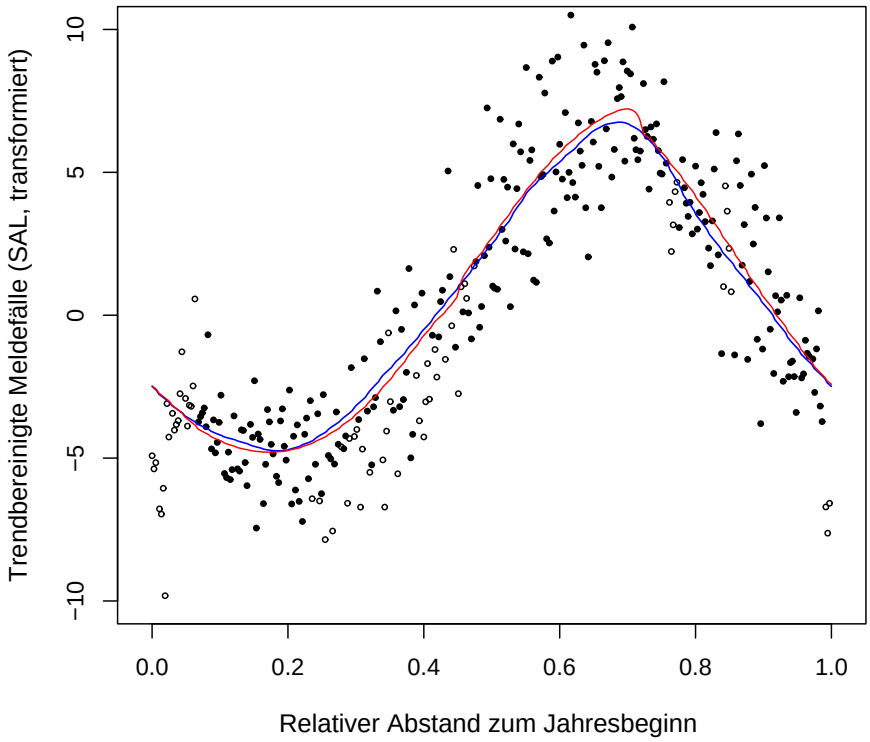

Abb. 4.19: Trendbereinigte Meldefälle $y_t - \hat{m}(t)$ (•), angeordnet entsprechend dem relativen Abstand der zugehörigen Meldewochen zum Jahresbeginn $\delta(t)$, $t = 1, \ldots, 313$, für SAL, Kandidatenfunktion nach LOK (blau) und geschätzte Saisonkomponente durch AWS ($h_{max} = 0.9$, rot). Meldefälle aus Wochen mit Kalendereffekten (○) sind nicht berücksichtigt.

Verfahren	*tv*	Verfahren	*tv*
LOK	75.0	SSA (L = 52, i = 1)	6.5
AWS ($h_{\max} = 52$)	41.4	SSA (L = 91, i = 1)	5.1
AWS ($h_{\max} = 104$)	31.9	SSA (L = 104, i = 1)	8.1
AWS ($h_{\max} = 156$)	32.3	SSA (L = 156, i = 1)	13.2
STS	–	SSA (L = 104, i = 1, 2)	10.7

Tab. 4.7: Totale Variation tv von Kandidaten für die geschätzte zyklische Komponente (SAL) nach verschiedenen Verfahren.

Wendepunkte überangepasst wirkt, vgl. Abb. A.9 im Anhang, wird jedoch das Ergebnis nach AWS als geschätzte Komponente $\hat{k}(t)$, $t \in T_W$, verwendet. Für den Osterzeitraum wird der zugehörige Effekt durch eine Konstante, das arithmetische Mittel der Beobachtungen geschätzt; die zugehörige Abb. A.10 ist im Anhang enthalten. Auf dieselbe Art wird der Kalendereffekt für alle übrigen Wochen $t \in T_R$ geschätzt. Hier zeigt sich allerdings, vgl. Abb. A.11 im Anhang, dass insbesondere die bereinigten Fälle aus den Wochen mit Allerheiligen offenbar keinen Kalendereffekt aufweisen.

Abb. A.12 (Anhang) zeigt die (transformierten) Beobachtungen und den durch die geschätzten Komponenten bestimmten Verlauf.

Nach Subtraktion aller geschätzten Komponenten von den Beobachtungen bleiben Residuen wie in Abb. A.13 (Anhang) dargestellt. Der einfachste ARMA-Prozess, der daran angepasst werden kann und für den die Adäquatheitsbedingung erfüllt ist, besitzt die Ordnung (2, 1). Die Hypothese, dass die durch diesen Prozess geschätzten Innovationen als Weißes Rauschen angesehen werden können, wird durch den Ljung-Box-Test zum Niveau $\alpha = 0.05$ weder unter Verwendung der Autokorrelationsfunktion (p-Wert $= 0.146$) noch unter der der partiellen Autokorrelationsfunktion (p-Wert $= 0.063$) abgelehnt. Darstellungen dieser Funktionen enthält Abb. A.14 im Anhang. Die geschätzten Koeffizienten des Prozesses sind $\hat{\phi}_1 = -0.296$, $\hat{\phi}_2 = 0.330$ und $\hat{\theta}_1 = 0.808$; die geschätzte Varianz beträgt $\hat{\sigma}^2 = 2.02$. Wie bei den anderen untersuchten Krankheiten ist die Verteilung der Innovationen zwar unimodal, aber leicht schief und kann insbesondere nicht als Normalverteilung angesehen werden.

Abb. 4.20 zeigt die tatsächlichen untransformierten Beobachtungen mit punktweisen 95% Vertrauensbereichen. Die offensichtlich gute Anpassung wird auch durch ein hohes Bestimmtheitsmaß $R^2 = 0.867$ verdeutlicht.

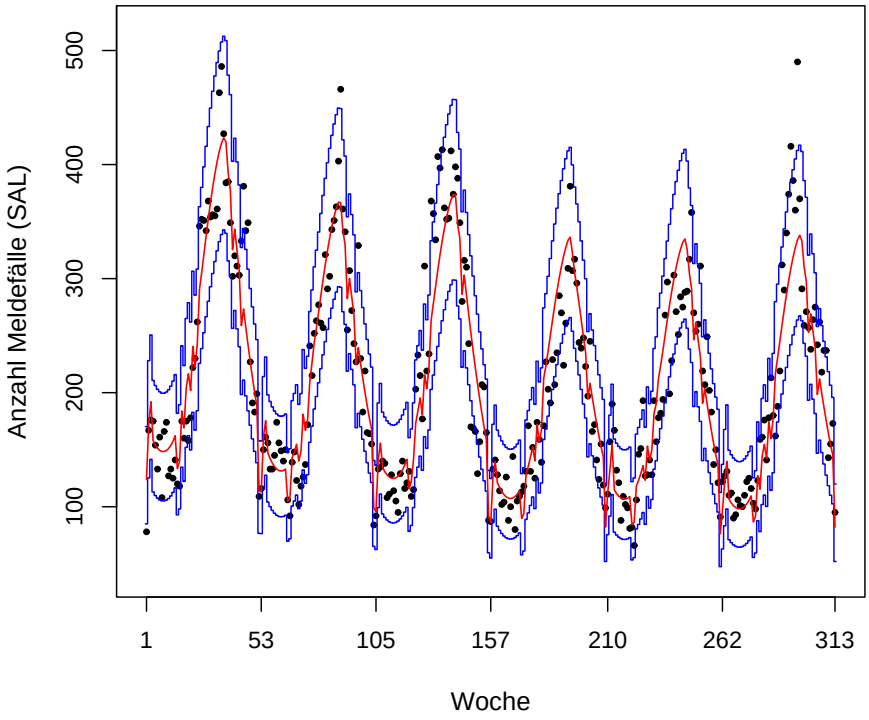

Abb. 4.20: Geschätzter Verlauf (rot) und 95%-Vertrauensbereiche (blau) für die Zahl der beobachteten (nicht transformierten) Meldefälle für SAL.

5 Parametrische Regression

5.1 Poissonregressionsmodell

Das Poissonregressionsmodell ist Teil der Klasse der generalisierten linearen Modelle, die erstmals von Nelder und Wedderburn (1972) beschrieben wurden und die Einsatzmöglichkeiten klassischer linearer Modelle bedeutend erweiterten. So besitzen generalisierte lineare Modelle den Vorteil, dass als Verteilung der Zielvariablen eine beliebige aus der Familie der Exponentialverteilungen vorausgesetzt werden kann. Die Art der Abhängigkeit des Erwartungswerts von den im Prädiktor enthaltenen Kovariablen kann darüber hinaus durch eine geeignete Linkfunktion flexibel modelliert werden. Gemäß den Eigenschaften der Poissonverteilung werden Poissonregressionsmodelle verwendet, wenn die abhängige Variable Y diskret ist und als Träger die Menge der natürlichen Zahlen besitzt. Als Linkfunktion $g(y)$ wird in der Regel $g(y) = log(y)$ verwendet. Die Linearkombination der unabhängigen Variablen $X_1, \ldots, X_p$ mit jeweiligen Parametern $\beta_1, \ldots, \beta_p$ bilden den Prädiktor η. Für Einstellungen $x_{1i}, \ldots, x_{pi}$ und zugehörige Zufallsvariablen Y_i, $i = 1, \ldots, n$, ist also

$$\log(\mathrm{E}\, Y_i) \;=\; \eta_i \;=\; \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}, \qquad i = 1, \ldots, n \tag{5.1}$$

bzw. $\log(\mathrm{E}\, \mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ (in Matrixschreibweise). Wie im klassischen Regressionsmodell sind die Ausprägungen der unabhängigen Variablen nicht stochastisch und die Variablen $Y_1, \ldots, Y_n$ unabhängig voneinander. Eine besondere Eigenschaft dieses Modells, die aus den Eigenschaften der Poissonverteilung resultiert, ist, dass $\mathrm{E}\, Y_t = \mathrm{Var}\, Y_t$, $t = 1, \ldots, 313$.

5.2 Modellgleichung und Variablenselektion

Meldefälle von Infektionskrankheiten können wegen ihrer Eigenschaft, natürliche Zahlen zu sein, als Realisationen Poisson-verteilter Zufallsvariablen angesehen werden. Diese Annahme wird zusätzlich durch die Beobachtung gestützt,

dass die Varianz der Anzahlen mit steigendem Erwartungswert zunimmt, was unter den hier betrachteten besonders eindeutig an den Meldehäufigkeiten der Rotavirus-Fälle erkennbar ist. Die Poissonverteilung wird auch in anderen Arbeiten zur Analyse von Meldefällen verwendet, vgl. Abschnitt 1.3. Zur Modellierung des zeitlichen Verlaufs durch ein Regressionsmodell müssen passende Kovariablen gefunden werden. Um es mit dem nichtparametrischen Ansatz vergleichbar zu machen, werden auch hier eigene Komponenten zur Modellierung von Trend, Saison, zyklischem Verlauf und Kalendereffekten gewählt. Durch ein schrittweises Verfahren werden dann unter den jeweils zur Verfügung stehenden Kovariablen, die im folgenden vorgestellt werden, mit Hilfe eines Gütekriteriums die geeignetsten ausgewählt.

Für die Modellierung des Trends werden als Kovariablen die Ausprägungen der Funktionen

$$x_1(t) = t, \qquad x_2(t) = t^2, \qquad x_3(t) = exp(t),$$

zugelassen. Damit ist die geschätzte Trendkomponente in jedem Fall glatt und weist weder lokale Extrema noch Wendepunkte auf.

Ähnlich wie in (4.6) wird zur Schätzung der Saisonkomponente zunächst die zeitliche Indexvariable t transformiert. Durch

$$\delta^*(t) = \begin{cases} \frac{7t}{365}, & t \leq 156 \\ \frac{7(t-156)-3}{366} + 3, & 156 < t \leq 209 \\ \frac{7(t-208)+2}{365} + 4, & 209 < t \end{cases}, \tag{5.2}$$

wird der Abstand einer Woche zum Beginn des Beobachtungszeitraums angegeben, wobei die Länge jedes Jahres auf 1 standardisiert wird. Damit nimmt die reskalierte Zeitreihe $y_{\delta^*(1)}, \ldots, y_{\delta^*(313)}$ Werte im Intervall $[0, 6]$ an, in dem die ganzen Zahlen $i = 1, \ldots, 6$ jeweils den Beginn des Jahres $2000 + i$ markieren. Ein Paar von Beobachtungen y_{t_1} und y_{t_2} mit $\delta^*(t_1) = k + \delta^*(t_2)$, $k \in \mathbb{Z}$ hat demnach einen zeitlichen Abstand von genau k Jahren.

Zur Modellierung der Saison sind Sinus-/Kosinusfunktionen wegen ihrer periodischen Eigenschaften und der Differenzierbarkeit geeignet. Die Funktionen

$$\begin{aligned} x_{s01} &= \sin\left(2\pi\delta^*(t)\right), & x_{c01} &= \cos\left(2\pi\delta^*(t)\right), \\ x_{s02} &= \sin\left(2\pi 2\delta^*(t)\right), & x_{c02} &= \cos\left(2\pi 2\delta^*(t)\right), \\ & \vdots & & \vdots \\ x_{s13} &= \sin\left(2\pi 13\delta^*(t)\right), & x_{c13} &= \cos\left(2\pi 13\delta^*(t)\right), \end{aligned}$$

weisen alle eine Periodizität von einem Jahr auf, d. h. $x_j(\delta^*(t)) = x_j(\delta^*(t)+1)$, $j \in \{s01, \ldots, s13, c01, \ldots, c13\}$. Funktionen mit Frequenzen kleiner als 1/13 werden nicht verwendet, da die Periodenlänge ansonsten kleiner als der Abstand von 4 aufeinanderfolgenden Wochen ist und Minima und Maxima damit an nicht beobachtbaren Positionen auftreten.

Zyklische Eigenschaften der Zeitreihe werden ebenfalls durch trigonometrische Funktionen dargestellt. Durch die periodischen Eigenschaften von

$$\begin{aligned}
x_{s2}(t) &= \sin\left(2\pi\frac{1}{2}\delta^*(t)\right) \quad , \quad x_{c2}(t) = \cos\left(2\pi\frac{1}{2}\delta^*(t)\right) \\
x_{s3}(t) &= \sin\left(2\pi\frac{1}{3}\delta^*(t)\right) \quad , \quad x_{c3}(t) = \cos\left(2\pi\frac{1}{3}\delta^*(t)\right) \\
&\vdots \qquad\qquad \vdots \\
x_{s6}(t) &= \sin\left(2\pi\frac{1}{6}\delta^*(t)\right) \quad , \quad x_{c6}(t) = \cos\left(2\pi\frac{1}{6}\delta^*(t)\right)
\end{aligned}$$

sind regelmäßige Effekte mit einer Periodenlänge von 2 bis 6 Jahren modellierbar. Funktionen mit niedrigeren Frequenzen können in dem insgesamt sechs Jahre umfassenden Beobachtungszeitraum nicht als wiederkehrende und damit zyklische Effekte beobachtbar sein.

Zur Modellierung der Kalendereffekte werden Indikatorfunktionen eingesetzt, deren Werte nur in den zugehörigen Meldewochen von 0 verschieden sind. Es werden drei Gruppen von Kovariablen gebildet entsprechend der Einteilung der Wochen mit Kalendereffekten in Weihnachts-/Neujahrperiode, Ostern und die übrigen Wochen mit einzelnen Feiertagen. Indem

$$\begin{aligned}
x_{\mathrm{R}}(t) &= \mathbb{1}_{T_{\mathrm{R}}}(t), \\
x_{\mathrm{O}j}(t) &= (\delta^*(t))^j \mathbb{1}_{T_{\mathrm{O}}}(t), \quad j = 0, \ldots, 3, \\
x_{\mathrm{W}j}(t) &= (\delta^*(t))^j \mathbb{1}_{T_{\mathrm{W}}}(t), \quad j = 0, \ldots, 3,
\end{aligned}$$

gesetzt werden, können die Effekte aus den Weihnachts- und Osterzeiträumen durch Polynome bis dritten Grades geschätzt werden. Für die Gruppe der einzelnen Feiertage wird lediglich die Schätzung durch Konstanten erlaubt, da die zugehörigen Wochen nicht direkt aufeinanderfolgen und damit keine zusammenhängende Einheit bilden. Es wäre alternativ möglich, für jede Woche mit einem bestimmten Feiertag einen jeweiligen eigenen Effekt zu schätzen, doch das erscheint zu speziell und läuft der Absicht entgegen, einen diesen Wochen gemeinsamen Effekt zu schätzen.

Zur Wahl eines geeigneten Modells wird ein schrittweises Verfahren verwendet. Ausgehend von einem Nullmodell, in dem der Prädiktor nur durch eine Konstante β_0 gebildet wird, werden in jedem Schritt alle jeweils um eine unabhängige Variable erweiterten Prädiktoren betrachtet. Sinus- und Kosinusfunktionen mit gleicher Frequenz werden dabei stets gleichzeitig und paarweise behandelt, weil durch ihre Summe auch Funktionen gebildet werden können, die bei gleicher Frequenz nicht durch den Ursprung (0,0) gehen. Für jede entstehende Modellgleichung werden die beteiligten Parameter durch die Maximum-Likelihood-Methode geschätzt und das zugehörige Informationskriterium nach Akaike (AIC) bestimmt, s. Akaike (1974). Indem es zur (negativen) maximierten Log-Likelihood des Modells die Anzahl der Parameter addiert, setzt das Kriterium eine gute Anpassung an die Beobachtungen mit der Komplexität des Modells in Beziehung. Als Ergebnis des jeweiligen Schritts wird dann dasjenige Modell mit dem kleinsten AIC angesehen.

In folgenden Iterationen wird der erweiterte Prädiktor entsprechend schrittweise erweitert. Die endgültige Modellgleichung ist gefunden, wenn keine weitere Kovariable aufgenommen werden kann, so dass der Wert des AIC dadurch erneut verringert würde. Die Vorwärtsselektion begünstigt damit sparsame Modelle, deren Prädiktor wenige Kovariablen enthält.

5.3 Ergebnisse

Das Ergebnis dieser Modellselektion wird zunächst am Beispiel der Meldefälle von Campylobacter vorgestellt. Die resultierende Modellgleichung ist durch den Prädiktor

$$\begin{aligned}\eta(t) &= \beta_0 + \beta_1 x_1(t) + \beta_3 x_3(t) \\ &\quad + \sum_{j=01}^{13} \left(\beta_{sj} x_{sj}(t) + \beta_{cj} x_{cj}(t)\right) + \sum_{j=1}^{6} \left(\beta_{sj} x_{sj}(t) + \beta_{cj} x_{cj}(t)\right) \\ &\quad + \sum_{j=0}^{3} \beta_{\mathrm{W}j} x_{\mathrm{W}j}(t) + \beta_{\mathrm{O0}} x_{\mathrm{O0}}(t) + \beta_{\mathrm{E}} x_{\mathrm{E}}(t)\end{aligned}$$

gegeben, mit dem der minimale Wert des AIC von 3321 erreicht wird. Auffällig ist hieran, dass die Prädiktorgleichung sämtliche zur Verfügung stehenden Kovariablen zur Modellierung von Saison und Zyklik enthält. Nur die für Trend und Kalenderkomponente erlaubten Faktoren x_2 und $x_{\mathrm{O1}}, \ldots, x_{\mathrm{O3}}$ werden nicht für das endgültige Modell ausgewählt. Abb. 5.1 zeigt den damit geschätzten Erwartungswert sowie punktweise Vertrauensintervalle zum Niveau 95%. Ein Vertrau-

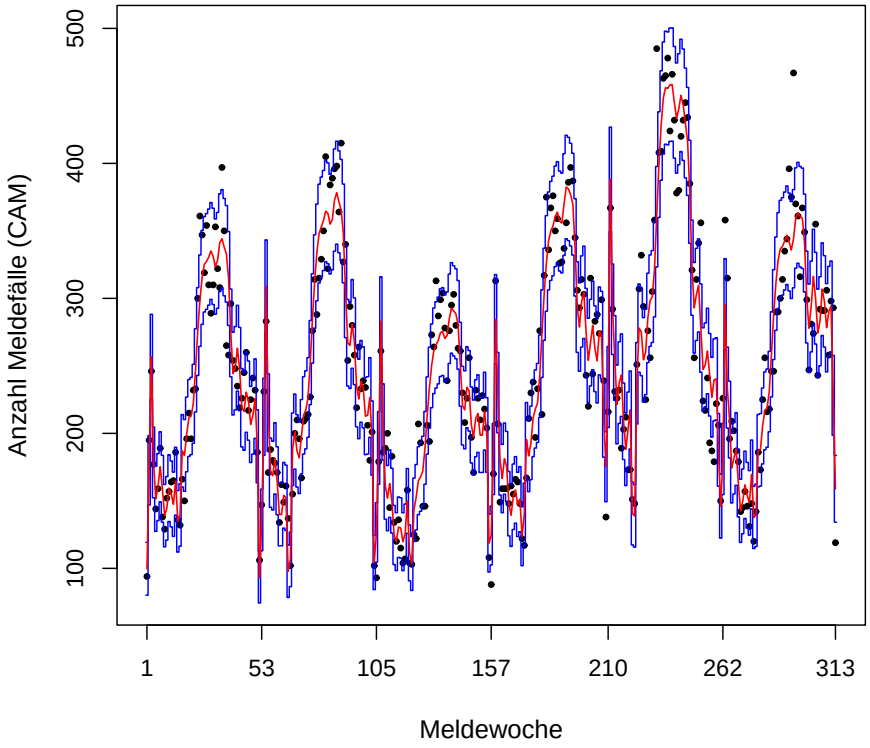

Abb. 5.1: Beobachtete Meldefälle (•) und geschätzter Erwartungswert (rot) sowie Vertrauensintervalle zum Niveau 95% (blau) nach Poissonregression (CAM).

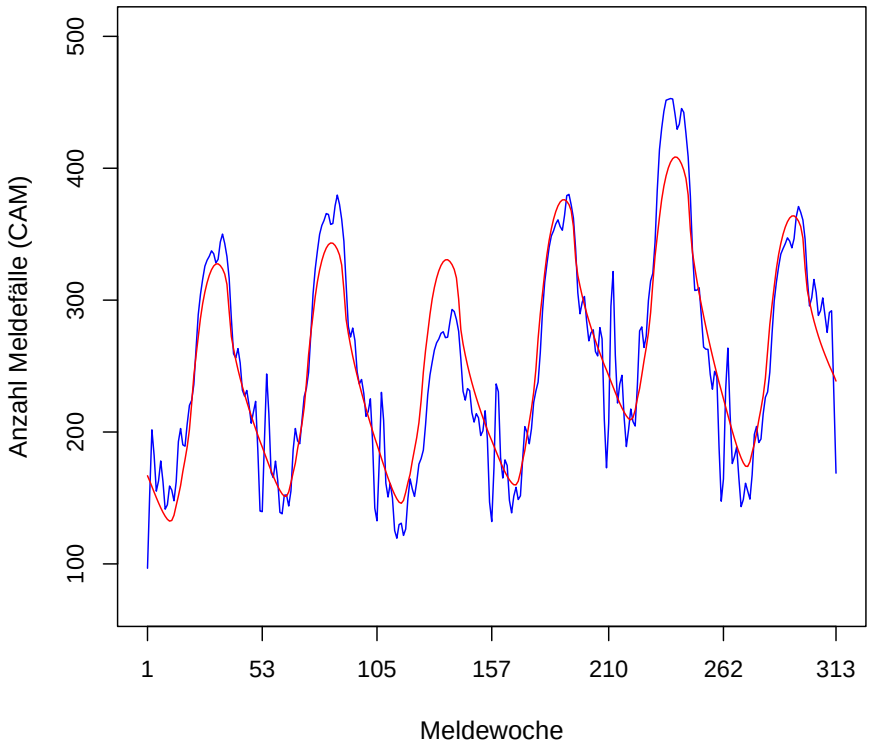

Abb. 5.2: Summe der geschätzten Trend-, Saison- und zyklischen Komponenten im nichtparametrischen (rot) und parametrischen (blau) Modell (CAM).

ensintervall zum Niveau $1-\alpha$ ist für eine Woche $t = 1, \ldots, 13$ gegeben durch $[\mathrm{E}\hat{Y}_t + u_{\alpha/2}\sqrt{\mathrm{E}\hat{Y}_t},\ \mathrm{E}\hat{Y}_t + u_{1-\alpha/2}\sqrt{\mathrm{E}\hat{Y}_t}]$, wobei u_α das α-Quantil der Standardnormalverteilung bezeichnet. Es kennzeichnet damit den Bereich, in dem ein Anteil von $1-\alpha$ aller Realisierungen einer $\mathcal{P}oi(\mathrm{E}\hat{Y}_t)$-verteilten Zufallsvariablen liegen.

Die Anpassung des geschätzten Erwartungswertes an die Beobachtungen ist mit $R^2 = 0.899$ sehr gut, jedoch deutet bereits der sehr variationsreiche Verlauf auf eine Überanpassung hin. Dieser Eindruck wird verstärkt, wenn das Resultat mit der Schätzung aus Abschnitt 4.8 verglichen wird. Abb. 5.2 zeigt dazu die Summe der geschätzten Trend-, Saison- und zyklischen Komponenten, die für den nichtparametrischen und den parametrischen Ansatz berechnet wurden. Insbesondere in den Zeiträumen zwischen den Jahreswechseln, wo die Senken der Kurve liegen, weist das Poissonregessionsmodell viele lokale Extrema für den geschätzten Erwartungswert auf. Durch das schrittweise Verfahren mit AIC werden zu viele Regressoren ausgewählt, so dass ein intendierter glatter Verlauf nicht erreicht wird und das Modell überangepasst wirkt.

Ähnliche Beobachtungen werden für das Ergebnis der Modellierung für die Salmonellen-Daten gemacht. Auch hier ist eine erhöhte Anzahl lokaler Extrema festzustellen sowie ein „wackeliger“ Verlauf, der insbesondere in den Zeiträumen mit wenigen Fällen bemerkbar ist, vgl. Abb. A.15 im Anhang. Diese als wenig glatt bewertbare Schätzung resultiert ebenfalls aus der Überanpassung des Modells, die einen leicht höheren Wert von R^2 als das Ergebnis der nichtparametrischen Schätzung besitzt. Dieses liefert $R^2 = 0.867$, während das geschätzte Modell nach Poissonregression mit $R^2 = 0.885$ knapp darüber liegt. Die bessere Anpassung durch den parametrischen Ansatz ist jedoch so gering, dass das zugehörige Modell wegen seiner größeren Komplexität nicht als geeigneter angesehen werden kann. Im Vergleich der nur durch Trend-, Saison- und Zyklikkomponenten geschätzten Verläufe werden erneut die strukturellen Unterschiede beider Ansätze deutlich, vgl. Abb. A.16 im Anhang.

Bei der Modellierung der Rotavirus-Fälle werden zwar alle Kovariablen selektiert, die zur Erklärung des zyklischen Verlaufs zur Verfügung stehen, aber die saisonale Komponente wird nur durch 7 anstatt wie bei den anderen Daten 13 Kovariablen erklärt. Das resultierende Modell ist damit zwar etwas sparsamer als das der anderen Krankheiten, dennoch werden die Beobachtungen sehr gut angepasst, vgl. Abb. A.15 im Anhang. Diesbezüglich ist das Modell mit $R^2 = 0.964$ dem Ergebnis durch das nichtparametrische Verfahren ($R^2 = 0.882$) überlegen, was vor allem durch die bessere Anpassung des Ausbruchszeitraums im letzten Jahr erreicht wird, vgl. dazu auch Abb. A.16 im Anhang. Eine mögliche Überanpassung ist wegen der größeren Spannweite der Beboachtungen nicht eindeutig zu erkennen. Die Kalendereffekte der Weihnachts-/Neujahrperiode werden durch eine lineare Funktion und die der Osterwochen durch eine Konstante geschätzt. Für die Wochen mit einzelnen Feiertagen wird kein Effekt modelliert.

6 Vorhersage

6.1 Ziele einer komponentenweisen Vorhersage

Auch wenn wie in dieser Arbeit die Analyse einer Zeitreihe hauptsächlich mit dem Ziel der Modellierung des beobachteten Verlaufs erfolgt, sind Möglichkeiten der Vorhersage oder Prognose mit dem resultierenden Modell ebenso von Interesse. In diesem Kapitel werden daher kurz Verfahren zur Prognose und ihre Anwendung auf die in Kapitel 4 gewonnenen Ergebnisse vorgestellt.

Entsprechend der Zerlegung der Zeitreihe in Komponenten kann die Prognose ebenfalls getrennt für diese Komponenten erfolgen. Wegen ihrer unterschiedlichen Interpretationen kann eine komponentenweise Vorhersage von größerer inhaltlicher Relevanz als die direkte Vorhersage der Summe aller Komponenten sein. Die möglichen Verfahren zur Vorhersage werden deswegen getrennt für die jeweiligen Komponenten vorgestellt und angewandt. Zur Beurteilung der Güte kann jedoch anhand der tatsächlichen Beobachtungen nur die Summe $\hat{z}_t$ der Komponenten, also die geschätzte Anzahl zukünftiger Meldefälle zu einem Zeitpunkt $t > n$, betrachtet werden, da die einzelnen Komponenten nicht beobachtbar sind.

Aus inhaltlichen Gründen ist die Prognose für ein mittelgroßes zukünftiges Zeitintervall von etwa 3 bis 12 Monaten interessant: Damit kann beispielsweise eine Abschätzung des allgemeinen Verlaufs erfolgen, der geschätzte Zeitpunkt mit den meisten Fällen im Jahr bestimmt und die Frage beantwortet werden, ob über einen längeren Zeitraum mehr oder weniger Fälle als im Vergleichszeitraum eines vergangenen Jahres erwartet werden. Demgegenüber ist der Nutzen kurzfristiger Vorhersagen für die betrachteten Krankheiten gering: Einem zu erwartenden kurzzeitigen Anstieg der Fälle könnte z. B. nicht durch spontane Impfprogramme entgegengewirkt werden. Außerdem ist wegen der hohen Variabilität insbesondere der Campylobacter- und Salmonellen-Fälle eine Vorhersage in Zeiträumen mit geringen Änderungen im Erwartungswert wenig nützlich.

Zur Vorhersage der einzelnen Komponenten sind verschiedene Verfahren notwendig. Die Prognose der Saisonkomponente wird durch ihre Eigenschaft $s_{\delta(t)} = s_{\delta(t)+1}$ direkt festgelegt. Ebenso werden wegen ihrer periodischen Ei-

genschaften die geschätzten Kalendereffekte für ein kommendes Jahr mit denen der vergangenen Jahre gleichgesetzt. Zur Prognose der Trend- und zyklischen Komponente sind hingegen spezielle Verfahren anzuwenden. Ein rein datenapproximativer Ansatz erlaubt ohne weitere Annahmen keine Extrapolation eines Signals außerhalb des beobachteten Zeitraums. In diesen Fällen muss also ein statistisches, parametrisches Modell formuliert werden, mit dessen Eigenschaften eine modellgestützte Vorhersage ermöglicht wird. Da die Schätzung der genannten Komponenten jeweils durch SSA erfolgte, ist es aussichtsreich, ein darauf basierendes Verfahren auch zur Prognose einzusetzen. Dies wird im folgenden vorgestellt.

6.2 SSA-basierte Prognose

Die grundlegende Modellannahme für eine Vorhersage mit SSA-Verfahren ist, dass eine Zeitreihe $y_1, \ldots, y_n$ durch eine linear rekursive Gleichung (*linear recurrent formula*, LRF) dargestellt werden kann, mit der eine Beobachtung y_t, $t \in \mathbb{Z}$, rekursiv durch gewichtete Summation von d vorhergehenden Beobachtungen als

$$y_t = a_1 y_{t-1} + \ldots + a_d y_{t-d} \tag{6.1}$$

berechnet werden kann. Dies ist äquivalent zu der Eigenschaft, dass jedes Element durch

$$y_t = \sum_{k=1}^{q} a_k(t) e^{\mu_k t} \sin(2\pi\omega_k t + \varphi_k) \tag{6.2}$$

als Summe von Produkten von Exponential-, Polynom- und Sinusfunktionen bestimmbar ist, wobei $a_k(t)$ Polynomfunktionen und $\mu_k, \omega_k, \varphi_k \in \mathbb{R}$ jeweilige Parameter sind. Für die Anzahl aller Terme in (6.2) gilt $q \leq d$. Unter diesen Bedingungen besitzt die Matrix $\mathbf{U}$ der linken Eigenvektoren der Singulärwertzerlegung (vgl. Kap. 2.5) für eine beliebige Fensterbreite höchstens den Rang d. Im Umkehrschluss bedeutet dies, dass für jede Zerlegung eine linear rekursive Gleichung gefunden werden kann, die über die linken Eigenvektoren definiert ist. Aus einer durch SSA-Zerlegung gewählten Gruppierung von Eigenvektoren kann damit eine lineare rekursive Gleichung ermittelt werden, mit der der Verlauf der Zeitreihe entsprechend dieser Eigenvektoren modelliert werden kann. Wenn die Werte für zukünftige Zeitpunkte $t > n$ bestimmt werden, wird die Vorhersage rekursiv berechnet.

Die für eine gewählte Fensterbreite L durch die Koeffizienten $a_1, \ldots, a_L$ definierte rekursive Gleichung kann für eine gegebene Gruppe von Eigenvektoren

$U_{i_1}, \ldots, U_{i_r}$ nach dem folgenden Verfahren bestimmt werden, vgl. Golyandina et al. (2001). Wie Danilov (1997) zeigt, muss als einzige technische Voraussetzung $e_L \notin \text{span}\{U_{i_1}, \ldots, U_{i_r}\}$ erfüllt sein, wobei $e_L = (0, \ldots, 0, 1)^\top \in I\!R^L$ ist.

Zunächst sei mit $\mathbf{Y}(L)$ die Übergangsmatrix einer Zeitreihe bei Wahl der Fensterbreite L bezeichnet, vgl. Kap. 2.5. Die Eigenvektoren $U_{i_1}, \ldots, U_{i_r}$ aus der Singulärwertzerlegung von $\mathbf{Y}\mathbf{Y}^\top$ bilden dann eine Orthonormalbasis für einen Unterraum $\mathcal{L}_r \subset I\!R^L$. Die Projektion der Übergangsmatrix in diesen Unterraum ist $\mathbf{Y}^* = \sum_{j=1}^r U_j U_j^\top \mathbf{Y}$. Durch diagonales Mitteln wird die zugehörige Matrix in Hankel-Form $\tilde{\mathbf{Y}} = \mathcal{H}\mathbf{Y}^*$ berechnet, die wiederum als Übergangsmatrix einer Zeitreihe $\tilde{y}_1, \ldots, \tilde{y}_n$ interpretiert werden kann.

Bezeichne weiter $\nu^2 = u_1^2 + \ldots + u_r^2$, wobei u_i der jeweils letzte Eintrag des Vektors U_i, $i = 1, \ldots, r$, ist. Dann gilt, dass für jeden Vektor $Z \in \mathcal{L}_r$ das letzte Element z_L als Linearkombination der übrigen Elemente dargestellt werden kann als $z_L = a_1 z_{L-1} + \ldots a_{L-1} z_1$. Der Vektor $A = (a_1, \ldots, a_r)'$ der Koeffizienten wird berechnet durch

$$A = \frac{1}{1-\nu^2} \sum_{i=1}^{r} u_i U_i^*,$$

wobei U_i^* der um das letzte Element u_i verkürzte Vektor der Länge $L-1$ ist. Als Basis kann statt $U_1, \ldots, U_r$ auch jede andere Orthonormalbasis gewählt werden, die denselben Raum $\mathcal{L}_r$ aufspannt.

Durch den Vektor A ist damit eine Schätzung für eine lineare rekursive Gleichung erfolgt. Zukünftige Beobachtungen $y_{n+1}, \ldots, y_{n+m}$ können durch rekursive Extrapolation als

$$\hat{y}_{n+k} = \sum_{i=1}^{L-1} a_i y_{n+k-i}, \qquad k = 1, \ldots, m, \tag{6.3}$$

geschätzt werden. Dabei werden in der Zukunft liegende Beobachtungen y_{n+k-i} mit $k - i \geq 1$ durch bereits prognostizierte Werte $\hat{y}_{n+k-i}$ gemäß (6.3) ersetzt.

Die Vorhersage der Fehlerkomponente $\{\varepsilon_t\}$ erfolgt, indem die geschätzten Innovationen als gegebene Realisierung des jeweils geschätzten ARMA-Prozesses angesehen wird. Der nächste zukünftige Fehler ε_{314} wird durch den bedingten Erwartungswert $E(\varepsilon_{314}|\hat{x}_1, \ldots, \hat{x}_{313})$ des Prozesses vorhergesagt. Dadurch dass diese Schätzung erneut als gegeben behandelt wird, lassen sich so schrittweise auch weitere zukünftige Fehler vorhersagen. Je größer der zeitliche Abstand zwischen der letzten und der vorherzusagenden Beobachtung ist, umso geringer wird der Betrag des geschätzten Fehlers sein, da langfristig allein die Eigenschaft des konstanten

Erwartungswertes 0 der Innovationen die Prognose bestimmt. Die Berechnung erfolgt hier mit dem R-Programm `predict.Arima`, mit dem der ARMA-Prozess in die Form eines Zustandsraummodells gebracht und mit Hilfe des Kalman-Filters vorhergesagt wird, vgl. dazu Durbin und Koopmann (2001).

6.3 Vorhersage für Januar - September 2007

Die Vorhersage der Meldefälle wird hier am Beispiel der Campylobacter-Meldefälle demonstriert. Ergebnisse für die Zeitreihen von Rotavirus und Salmonellen werden zusätzlich im Anhang skizziert, vgl. A.17. Die Prognose erfolgt für die Meldewochen 314 - 352 (Januar - September 2007) und enthält damit insbesondere den relevanten Ausbruchszeitraum in der Mitte des Jahres. Für die deterministischen Komponenten werden Vorhersagen für den Trend und die zyklische Komponente entsprechend des in 6.2 erläuterten Verfahrens erstellt. Für die Prognose des Trends wird mit $L = 156$ eine hohe Fensterbreite gewählt, da der Trend als langfristige Veränderung interpretiert wird und daher ein Einfluss auch weit zurückliegender Beobachtungen in der Rekursionsgleichung (6.3) gerechtfertigt ist. Bei der zyklischen Komponente wird hingegen eine kleinere Fensterbreite verwendet. Da diese Komponente Veränderungen beschreibt, die sich über mehr als die Länge eines Jahres erstrecken, sollte der Wert größer als 52 sein. Durch $L = 104$ wird er in der Mitte zwischen den sich so ergebenden Grenzen 156 und 52 gesetzt. Als Parameter wird $r = 3$ bzw. $r = 5$ gewählt.

Die Ergebnisse der Vorhersage, vgl. Abb. 6.1, zeigen einen weiteren, durch die Trendkomponente begründeten allgemeinen Anstieg der Fälle. Dieser wird noch verstärkt durch eine gleichzeitige weitere Zunahme des zyklischen Effekts, der etwa ab etwa Woche 335 (Juni 2007) leicht abflacht.

Der geschätzte Verlauf der Fehlerkomponente ist für die weitere Interpretation unbedeutend. Der Erwartungswert wird für die erste Woche 314 auf etwa -3.1 vorhergesagt und nimmt dann betragsmäßig für die folgenden Wochen mit geometrischer Rate ab. Bereits ab Woche 319 ist der Effekt mit etwa -0.1 nahezu verschwunden.

Die Summe der geschätzten Komponenten nach Umkehrung der Transformation und damit die prognostizierten Meldefälle zeigt Abb. 6.2 im Vergleich zu den tatsächlichen Beobachtungen des untersuchten Zeitraums. Grundsätzlich wird der beobachtete Verlauf durch die Vorhersage gut angenähert. Zwar werden die Beobachtungen im ersten Quartal (Wochen 314 - 326) fast immer leicht unterschätzt, aber im darauffolgenden Zeitraum passt die Vorhersage die Beobachtungen besser an. Im zweiten und dritten Quartal liegen nur die letzten

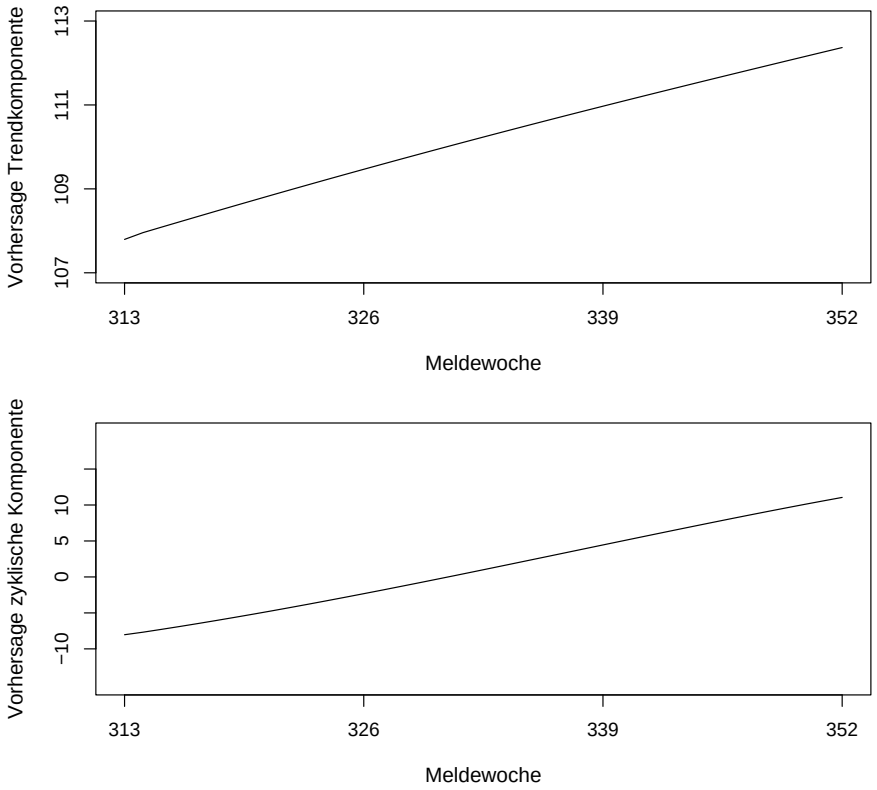

Abb. 6.1: Prognose der Trend- (oben) und zyklischen Komponente (unten) für die Wochen 314 - 352.

Beobachtungen deutlich außerhalb des zugehörigen punktweisen Vertrauensbereichs zum Niveau 95%. Die Position des vorhergesagten Maximums verfehlt die des tatsächlichen nur um 1 Woche.

Anmerkung: Die Vorhersage, wie sie in diesem Fall durch das Poissonregressionsmodell getroffen wird, ist im Vergleich deutlich schlechter. Abb. A.18 im Anhang zeigt, dass die tatsächlichen Fälle in den ersten beiden Quartalen fast ausnahmslos unterschätzt werden. Auch eine gute Anpassung in den Wochen 348 bis 352 kann diesen Nachteil nicht aufwiegen.

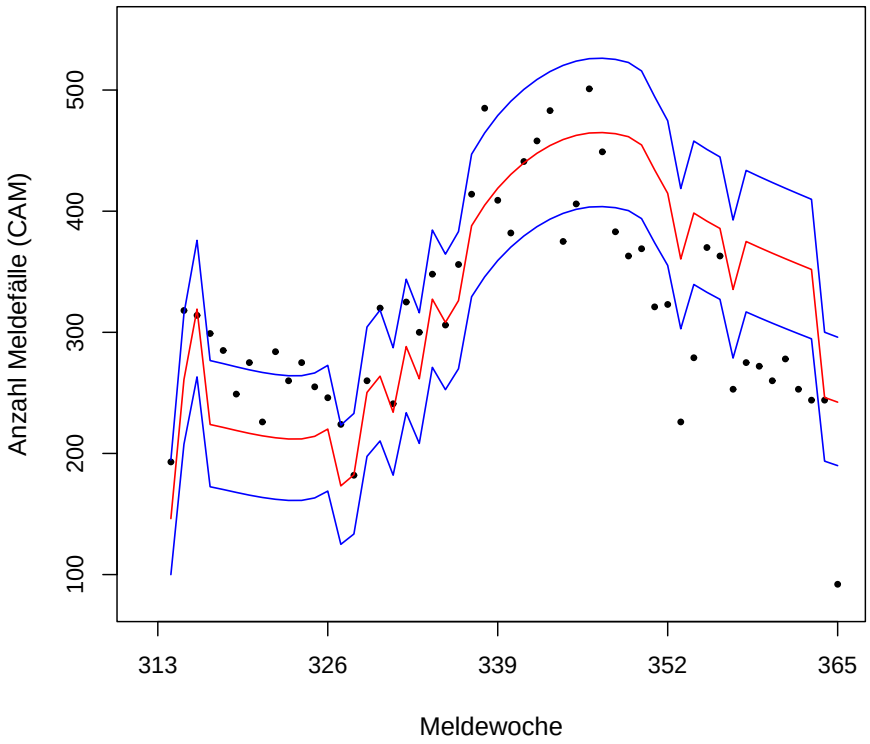

Abb. 6.2: Beobachtete Meldefälle (•) und geschätzte Vorhersage (rot) sowie Vertrauensintervalle zum Niveau 95% (blau) für die Wochen 314 - 352 (Januar - September 2007, CAM).

7 Zusammenfassung und Diskussion

In dieser Arbeit wurden nichtparametrische Verfahren zur Schätzung der Komponenten eines Dekompositionsmodells für wöchentliche Zeitreihen von Meldefällen von Infektionskrankheiten untersucht. Dabei wurden für die jeweiligen Komponenten individuelle Kriterien zur Auswahl einer geeigneten Schätzung festgelegt. Die Ergebnisse ermöglichen inhaltliche Aussagen über den allgemeinen Trendverlauf, die Art der saisonalen periodischen Struktur sowie über weitere aperiodische Veränderungen. Kalenderbedingte Abweichungen werden durch eine weitere Komponente erklärt. Für die verbleibende, zufällige Variation konnte gezeigt werden, dass diese als Realisation einfacher ARMA-Prozesse angesehen werden kann und damit der Erwartungswert der Zeitreihe durch die deterministischen Komponenten vollständig erklärt wird. Wie die Summe der geschätzten Komponenten im Vergleich zum tatsächlichen Verlauf der Beobachtungen zeigt, ermöglicht das resultierende Modell eine gute Anpassung an die Daten.

Ein parametrisches Poissonregressionsmodell, dessen Regressoren ähnlich komponentenweise zerlegbar sind, ermöglicht eine insgesamt bessere Anpassung an die Daten. Andererseits zeigt der Vergleich mit den deterministischen Komponenten des nichtparametrischen Modells, dass das parametrische Modell überangepasst ist. Die Variablenselektion durch AIC führt hier offensichtlich zu einem zu komplexen Modell. Statt der hier gewählten könnten auch andere Regressoren verwendet werden, da ihre Auswahl letztlich subjektiv erfolgt. Demgegenüber besitzt der nichtparametrische Ansatz den Vorteil, dass die grundsätzlichen Modellannahmen festgelegt und für jedes Verfahren dieselben sind.

Weiterhin wurde gezeigt, dass die Vorhersage der Zeitreihe auf Grundlage des geschätzten Modells ebenfalls auch komponentenweise möglich ist. Dies kann jedoch nicht ohne zusätzliche Struktur- bzw. Verteilungsannahmen erfolgen, da das nichtparametrische Regressionsmodell im Allgemeinen nicht an Stellen außerhalb des beobachteten Bereichs definiert ist.

Das vorgeschlagene Verfahren zur schrittweisen Komponentenschätzung kann an verschiedenen Stellen erweitert oder modifiziert werden. Grundsätzlich ist es möglich, statt der betrachteten noch weitere Methoden zur Bestimmung von Kandidatenfunktionen zu berücksichtigen. Da bereits mit den hier verwendeten eine adäquate Modellierung möglich ist, scheinen durch zusätzliche Methoden nur kleinere Verbesserungen hinsichtlich der Komplexität des Modells wahrscheinlich. Für die Schätzung der Kalenderkomponente für den Zeitraum Weihnachten/Neujahr kann es sinnvoll sein, statt der hier verwendeten Verfahren mit lokaler Anpassung ggf. auch ein globaleres oder robustes zu verwenden. Die Schätzung dieser Komponente ist nämlich besonders bei kurzen Zeitreihen anfällig gegenüber einzelnen Ausreißern.

Es ist grundsätzlich möglich, die in dieser Arbeit verwendeten Optimalitätskriterien zur Auswahl der Schätzungen individuell zu verändern und damit die Komponenten inhaltlich anders zu deuten. Das Modell kann damit flexibel entsprechend jeweiliger Anforderungen an die Interpretation modifiziert werden. Beispielsweise würden andere Ergebnisse resultieren, wenn das Kriterium verändert würde, mit dem die Anpassung der um die deterministischen Komponenten bereinigten Beobachtungen durch ARMA-Prozesse verlangt wird. Zur Überprüfung dieses Kriteriums, die hier mit Hilfe des Ljung-Box-Tests erfolgt, könnten darüber hinaus auch andere Tests oder Prozeduren eingesetzt werden. Möglich wäre eine Kombination aus einem Test, der zunächst die Stationarität der Residuen überprüft, und weiteren Tests, mit denen relevante Eigenschaften der geschätzten Innovationen bewertbar sind. Die Anwendung multipler Tests würde dann eine Anpassung der jeweiligen Niveaus zur Einhaltung des Gesamtniveaus erforderlich machen.

Das hier vorgestellte Verfahren kann auch eingesetzt werden, wenn unter Berücksichtigung weiterer Kovariablen wie z. B. Alter, Geschlecht oder Meldeort entsprechend multivariate Zeitreihen zu untersuchen sind. Für eine angemessene Vergleichbarkeit untereinander ist es dann sinnvoll, statt der eigentlichen Meldefälle Inzidenzen zu betrachten, um einen Bezug zur jeweiligen Bevölkerungsgröße herzustellen. Bei der Interpretation der Ergebnisse ist dann zu berücksichtigen, dass in Deutschland Bevölkerungszahlen für Kovariablen wie z. B. Altersgruppen, Geschlecht und (Land-)Kreiszugehörigkeit nur auf Schätzungen beruhen.

Ist in einem solchen Fall die räumliche Variation von größerem Interesse, könnte in Analogie zu Fanshawe et al. (2007) eine zweistufige Analyse erfolgen. Dabei würden zunächst für räumlich aggregierte Zeitreihen die hier vorgestellten Komponenten geschätzt werden. In einem zweiten Schritt könnte dann die

räumliche Verteilung der Residuen untersucht werden. Ein derartiges Vorgehen besitzt den Vorteil, dass für die zeitliche und räumliche Modellierung unterschiedliche Modellierungsverfahren eingesetzt werden können. Dies hat andererseits den Nachteil, dass die räumliche Variation nur bedingt auf eine bereits erfolgte zeitliche Modellierung erfolgen kann.

Sind zeitliche und räumliche Effekte gleichermaßen von Interesse, kann ihre gemeinsame Verteilung nur mit einem entsprechenden räumlich-zeitlichen Modell angemessen untersucht und interpretiert werden. Insbesondere Wechselwirkungen dieser beiden Größen sind dann zu berücksichtigen. Ein Modell, das in Analogie zu dem hier vorgestellten neben zeitlichen auch räumliche Komponenten beinhaltet und das mit nichtparametrischen Methoden geschätzt wird, müsste dazu noch entwickelt werden. Beispielsweise ließe sich dafür der Einsatz von Verfahren aus der Bildverarbeitung untersuchen. In jedem Fall könnten mit derartigen Modellen weitere Erkenntnisse zur Epidemiologie der Krankheiten gewonnen werden.

Literaturverzeichnis

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Allen, M. R. und Smith, L. A. (1996). Monte Carlo SSA: Detecting irregular oscillations in the presence of coloured noise. *Journal of Climate*, 9(12):3373–3404.

Anderson, T. W. (1971). *The statistical analysis of time series*. John Wiley & Sons, New York.

Box, G. E. und Jenkins, G. M. (1970). *Time series analysis : forecasting and control*. Holden-Day, San Francisco.

Box, G. E., Jenkins, G. M., und Reinsel, G. C. (1994). *Time series analysis : forecasting and control*. Prentice-Hall, Englewoods Cliffs (NJ).

Box, G. E. P. und Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26:211–252.

Box, G. E. P. und Pierce, D. A. (1970). Distribution of the autocorrelations in autoregressive moving average time series models. *Journal of the American Statistical Association*, 65:1509–1526.

Brillinger, D. R. (1975). *Time Series: Data Analysis and Theory*. Holt, Rinehart and Winston, New York.

Brockmann, M., Gasser, T., und Herrmann, E. (1993). Locally adaptive bandwidth choice for kernel regression estimators. *Journal of the American Statistical Association*, 88:1302–1309.

Brockwell, P. J. und Davis, R. A. (1991). *Time Series: Theory and Methods, 2nd Edition*. Springer-Verlag, New York.

Brockwell, P. J. und Davis, R. A. (2002). *Introduction to Time Series and Forecasting, 2nd Edition*. Springer, New York.

Brookmeyer, R. (2004). Temporal factors in epidemics: The role of the incubation period. In Brookmeyer, R. und Stroup, D. F., editors, *Monitoring the Health of Populations*, pages 127–146. Oxford University Press, New York.

Broomhead, D. S., Jones, R., King, G. P., und Pike, E. R. (1987). Singular system analysis with applications to dynamical systems. In Pike, E. R. und Lugaito, L. A., editors, *Chaos, Noise and Fractals*, pages 15–27. IOP Publishing, Bristol.

Broomhead, D. S. und King, G. P. (1986). Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena*, 20:217–236.

Chiogna, M. und Gaetan, C. (2005). Mining epidemiological time series: an approach based on dynamic regression. *Statistical Modelling*, 5:309–325.

Cleveland, W. P. und Scott, S. (2007). Seasonal adjustment of weekly time series with application to unemployment insurance claims and steel production. *Journal of Official Statistics*, 23(2):209–221.

Danilov, D. L. (1997). Principal components in time series forecast. *Journal of Computational and Graphical Statistics*, 6(1):112–121.

Davies, P. L. (1995). Data features. *Statistica Neerlandica*, 49:185–245.

Davies, P. L. (2003). Approximating data and statistical procedures - I. Approximating data. Technical Report 7/2003, Sonderforschungsbereich 475, University of Dortmund, Dortmund, Germany.

Davies, P. L., Gather, U., und Weinert, H. (2008). Nonparametric regression as an example of model choice. *Communications in Statistics - Simulation and Computation*, 37(2):274–289.

Davies, P. L. und Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *The Annals of Statistics*, 29:1–65.

Devine, O. (2004). Exploring temporal and spatial patterns in public health surveillance data. In Brookmeyer, R. und Stroup, D. F., editors, *Monitoring the Health of Populations*, pages 71–98. Oxford University Press, New York.

Diggle, P., Rowlingson, B., und Su, T. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, 16:423–434.

Diggle, P. J. (1990). *Time Series - A Biostatistical Introduction*. Oxford University Press, Oxford.

Durbin, J. und Koopmann, S. J. (2001). *Time Series Analysis by State Space Models*. Oxford University Press, Oxford.

Ethelberg, S., Simonsen, J., Gerner-Smidt, P., Olsen, K. E. P., und Mølbak, K. (2005). Spatial distribution and registry-based case-control analysis of campylobacter infections in Denmark, 1991-2001. *American Journal of Epidemiology*, 162(10):1008–1015.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. Marcel Dekker, Inc., New York.

Exner, M. (1997). Grundlagen der Infektionssurveillance. *Gesundheitswesen*, 59:686–695.

Fan, J. und Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.

Fan, J. und Yao, Q. (2003). *Nonlinear Time Series - Nonparametric and Parametric Methods*. Springer, New York.

Fanshawe, T. R., Diggle, P. J., Rushton, S., Sanderson, R., Lurz, P. W. W., Glinianaia, S. V., Pearce, M. S., Parker, L., Charlton, M., und Pless-Mulloli, T. (2007). Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach. *erscheint in: Environmetrics*.

Farrington, C. P., Andrews, N. J., Beale, A. D., und Catchpole, M. A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society, Series A*, 159(3):547–563.

Gasser, T. und Müller, H. (1979). *Smoothing Techniques for Curve Estimation*. Springer, New York.

Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F., und Yiou, P. (2002). Advanced spectral methods for climatic time series. *Reviews of Geophysics*, 40(1):1003–1044.

Gilks, W. R., Richardson, S., und Spiegelhalter, D. J. (1994). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, Boca Raton.

Golyandina, N., Nekrutkin, V., und Zhigljavsky, A. (2001). *Analysis of Time Series Structure – SSA and Related Techniques*. Chapman & Hall/CRC, Boca Raton.

Guerrero, V. M. (1993). Time-series analysis supported by power transformations. *Journal of Forecasting*, 12:37–48.

Györfi, L., Härdle, W., Sarda, P., und Vieu, P. (1989). *Nonparametric curve estimation from time series*. Lecture Notes in Statistics 60. Springer, Berlin.

Haber, M. (1997). Estimation of the population effectiveness of vaccination. *Statistics in Medicine*, 16:601–610.

Härdle, W., Lütkepohl, H., und Chen, R. (1997). A review of nonparametric time series. *International Statistical Review*, 65(1):49–72.

Hart, J. D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society, Series B*, 53(1):173–187.

Harvey, A., Koopman, S. J., und Riani, M. (1997). The modeling and seasonal adjustment of weekly observations. *Journal of Business & Economic Statistics*, 15(3):354–368.

Hashimoto, S., Murakami, Y., Taniguchi, K., und Nagai, M. (2000). Detection of epidemics in their early stage through infectious disease surveillance. *International Journal of Epidemiology*, 29:905–910.

Heiler, S. und Feng, Y. (2000). Data-driven decomposition of seasonal time series. *Journal of Statistical Planning and Inference*, 91:351–363.

Heisterkamp, S. H., Dekkers, A. L. M., und Heijne, J. C. M. (2006). Automated detection of infectious disease outbreaks: hierarchical time series models. *Statistics in Medicine*, 25(24):4179–4196.

Held, L., Hofmann, M., Höhle, M., und Schmid, V. (2006). A two-component model for counts of infectious diseases. *Biostatistics*, 7(3):422–437.

Held, L., Höhle, M., und Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, 5:187–199.

Herrmann, E. (1997). Local bandwidth choice in kernel regression estimation. *Journal of Computational and Graphical Statistics*, 6:35–54.

Herrmann, E., Gasser, T., und Kneip, A. (1992). Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika*, 79(4):783–795.

Jenkins, G. M. (1979). *Practical Experiences with Modelling and Forecasting Time Series*. Gwilym Jenkins & Partners (Overseas), St. Helier (Jersey).

Kendall, S. M. und Ord, J. K. (1990). *Time series, 3rd edition*. Edward Arnold, Sevenoaks.

Kermack, W. O. und McKendrick, A. G. (1927). A contribution on the mathematical theory of epidemics. *Proceedings of the Royal Society*, 115A:700–721.

Knorr-Held, L. und Richardson, S. (2003). A hierarchical model for space-time surveillance data on meningococcal disease incidence. *Applied Statistics*, 52(2):169–183.

Lindbäck, J. und Svenson, Å. (2001). Campylobacter infections in Sweden - a statistical analysis of temporal and spatial distributions of notified sporadic campylobacter infections. Technical Report 2001/4, Mathematical Statistics, Dept. of Mathmatics, Stockholm University, Stockholm.

Ljung, G. M. und Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

MacNab, Y. C. und Dean, C. B. (2001). Autoregressive spatial smoothing and temporal spline smoothing for mapping rates. *Biometrics*, 57:949–956.

Majidi, A. (2003). *Glatte nichtparametrische Regression unter formerhaltenden Bedingungen*. Dissertation, Universität Duisburg-Essen.

Mann, H. B. und Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.

Mugglin, A. S., Cressie, N., und Gemmell, I. (2002). Hierarchical statistical modelling of influenza epidemic dynamics in space and time. *Statistics in Medicine*, 21(18):2703–2721.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142.

Nelder, J. A. und Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135(3):370–384.

Opsomer, J., Wang, Y., und Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, 16(2):134–153.

Polzehl, J. und Spokoiny, V. (2003). Varying coefficient regression modeling by adaptive weights smoothing. Technical report, Weierstraß-Institut, Berlin, Germany.

Polzehl, J. und Spokoiny, V. G. (2000). Adaptive weights smoothing with applications to image restoration. *Journal of the Royal Statistical Society B*, 62:335–354.

Ranta, J., Mäkelä, P. H., und Arjas, E. (2004). Predicting meningococcal disease outbreaks in structured populations. *Statistics in Medicine*, 23:927–945.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer, New York.

Vautard, R., Yiou, P., und Ghil, M. (1992). Singular-spectrum analysis - a toolkit for short, noisy chaotic signals. *Physica D*, 58:95–126.

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia.

Waller, L. A. (2004). Detecting disease clustering in time or space. In Brookmeyer, R. und Stroup, D. F., editors, *Monitoring the Health of Populations*, pages 167–202. Oxford University Press, New York.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya, Series A*, 26:359–372.

Anhang

A.1 Medizinische und epidemiologische Informationen zu den betrachteten Krankheiten

Die Zusammenstellung der hier aufgeführten Informationen zu den Krankheiten erfolgt auf Grundlage der *Merkblätter für Ärzte* des RKI, die im Epidemiologischen Bulletin und im Internet (`www.rki.de`) veröffentlicht werden.

Campylobacteriose

Infektionen durch Campylobacter-Bakterien, von denen verschiedene Arten bekannt sind, werden weltweit, in Europa besonders während der warmen Jahreszeit beobachtet. Betroffen sind alle Altersgruppen, am häufigsten jedoch Kinder (0 - 10 Jahre) und junge Erwachsene (20 - 29 Jahre). Der Erreger kommt in der Natur bei vielen Vögeln und Säugetieren sowie bei Nutz- und Haustieren vor und wird zumeist durch die Einnahme kontaminierter Fleisch- und Milchprodukte übertragen, wenn diese unzureichend erhitzt worden sind. Daneben ist auch der direkte Kontakt von Mensch zu Mensch ein möglicher Infektionsweg, der insbesondere bei Kindern auftritt.

Nach einer Inkubationszeit von 2-5 Tagen treten häufig typische Symptome einer Durchfallerkrankung auf, daneben können auch Kopfschmerzen, Fieber und allgemeine Müdigkeit beobachtet werden. Viele Infektionen verlaufen auch ohne die genannten Symptome. Die Erkrankung dauert etwa 1 Woche an, und heilt in der Regel von selbst aus. Nur bei seltenen, schweren Verläufen ist eine antibiotische Behandlung notwendig.

Rotavirus-Infektion

Die Infektion erfolgt durch eine Virusinfektion einer der 7 bekannten, weltweit verbreiteten Serogruppen. Wegen ihres noch schwach ausgebildeten Immunsystems sind vor allem Kleinkinder und Säuglinge (0 - 5 Jahre), seltener auch Erwachsene, hier aber vermehrt die Gruppe der über 65-Jährigen betroffen. Der

Erreger wird meist fäkal-oral von Mensch zu Mensch durch kontaminiertes Wasser oder Lebensmittel übertragen.

Die Krankheit kann nach einer Inkubationszeit von 1-3 Tagen mit typischen Durchfall-Symptomen, in vielen Fällen (etwa 30 - 40 %) jedoch auch mit schweren Symptomen verlaufen, die einen Krankenhausaufenthalt erforderlich machen. Unbehandelt kann die Erkrankung in seltenen Fällen tödlich enden. Ansonsten schwächen sich die Symptome nach etwa 2-6 Tagen bei ausreichender Aufnahme von Flüssigkeit von selbst ab. Eine antivirale Therapie oder sonstige den Erreger bekämpfende Mittel sind nicht bekannt.

Salmonellose

Salmonellese wird durch Ansteckung mit einem Serovar der weltweit vorkommenden Salmonella-Gruppe ausgelöst. Der Kontakt erfolgt in der Regel durch den Verzehr kontaminierter Lebensmittel, hauptsächlich Geflügel- und Eiprodukte. Durch hygienewidrige Lagerung oder Aufbereitung, wie unzureichende Erhitzung oder zu lange Lagerung, kann sich der Erreger darin stark vermehren und dadurch eine zur Infektion ausreichende Keimzahl entwickeln. Darüber hinaus können Salmonellen auch direkt von Mensch zu Mensch übertragen werden.

Häufige Verlaufsformen der Erkrankung sind nach einer Inkubationszeit von 6-72 Stunden Durchfälle, Übelkeit oder Fieber, die nach spätestens 2-3 Tagen wieder abklingen. Daneben treten in vielen Fällen auch keine oder nur sehr schwache Symptome auf. Nur bei schweren Verlaufsformen oder für Säuglinge ist eine besondere Behandlung durch antibakterielle Chemotherapie notwendig.

A.2 Alternative ARMA-Prozesse

Die Residuen des Rotavirus-Modells in Abschnitt 4.7 können auch durch einen ARMA(1,1)-Prozess angepasst werden. Dabei resultieren die Schätzwerte $\hat{\theta} = 0.895$ und $\hat{\phi} = -0.396$. Die geschätzte Varianz der Innovationen beträgt $\hat{\sigma}^2 = 1.148$. Die Hypothese des Ljung-Box-Test kann weder mittels der Autokorrelations- noch der partiellen Autokorrelationsfunktion abgelehnt werden. Wie die folgende Abbildung A.1 zeigt, zeigen beide Funktion das unter der Hypothese Weißen Rauschens zu erwartende Verhalten, dass die Funktionswerte zufällig um 0 streuen.

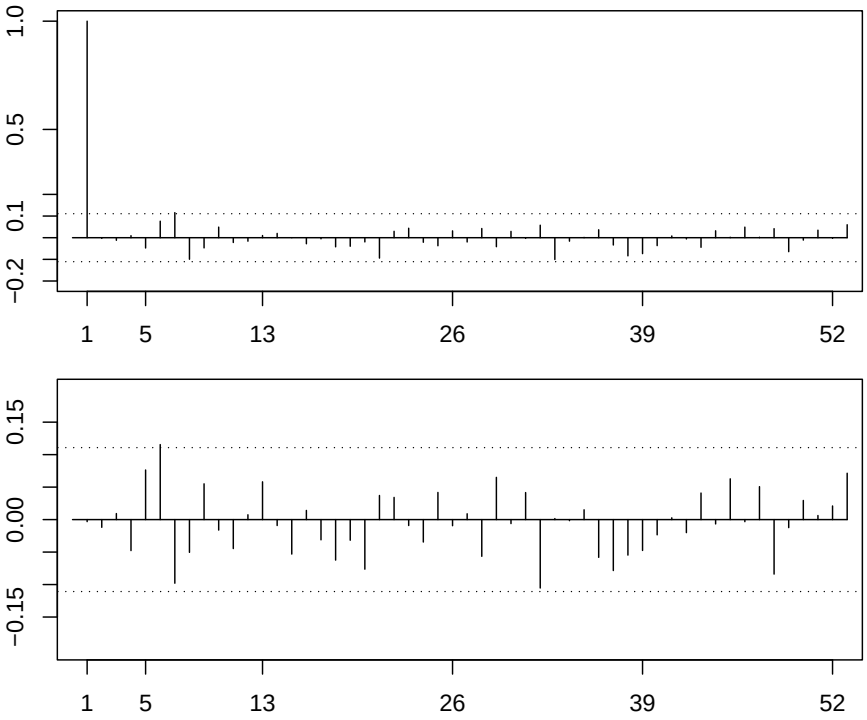

Abb. A.1: Autokorrelationsfunktion (oben) und partielle Autokorrelationsfunktion (unten) mit 5% Konfidenzgrenzen der geschätzten Innovationen $\hat{x}(t)$ nach Anpassung eines ARMA(1,1)-Prozesses an die Residuen (RTV).

A.3 Weitere Abbildungen

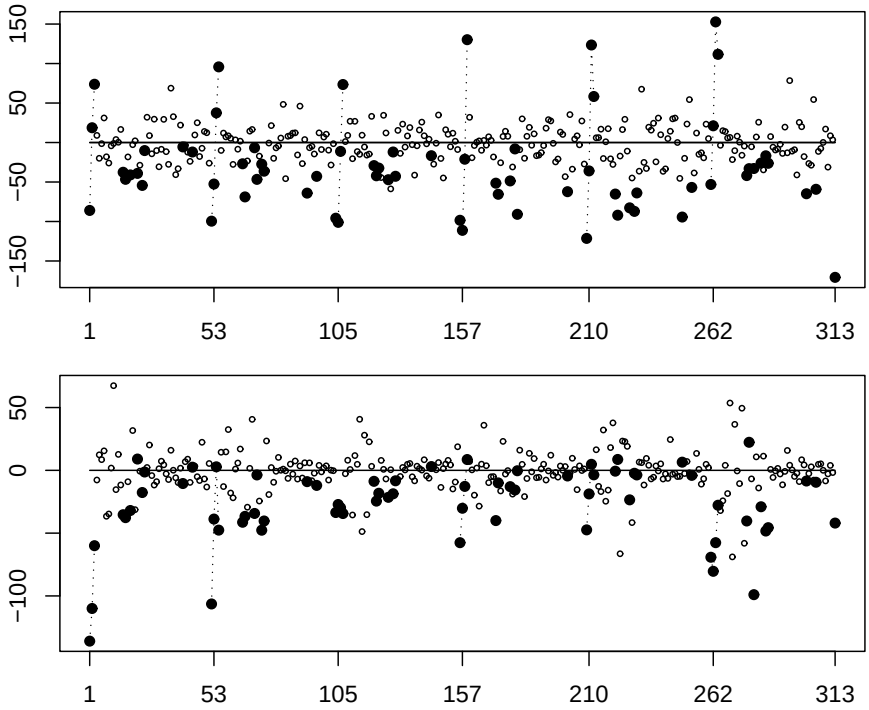

Abb. A.2: Residuen nach Schätzung der Signalkomponente $f(t)$ durch LOK für CAM (oben) und Residuen nach Schätzung der Signalkomponente $f(t)$ durch AWS für RTV (unten). Beobachtungen aus Wochen mit Kalendereffekten wurden bei der Schätzung durch interpolierte Werte ersetzt. Die Residuen für diese Wochen sind gesondert markiert (•), zeitlich aufeinanderfolgende Residuen in der Weihnachts-/Neujahrperiode sind zusätzlich durch Linien verbunden.

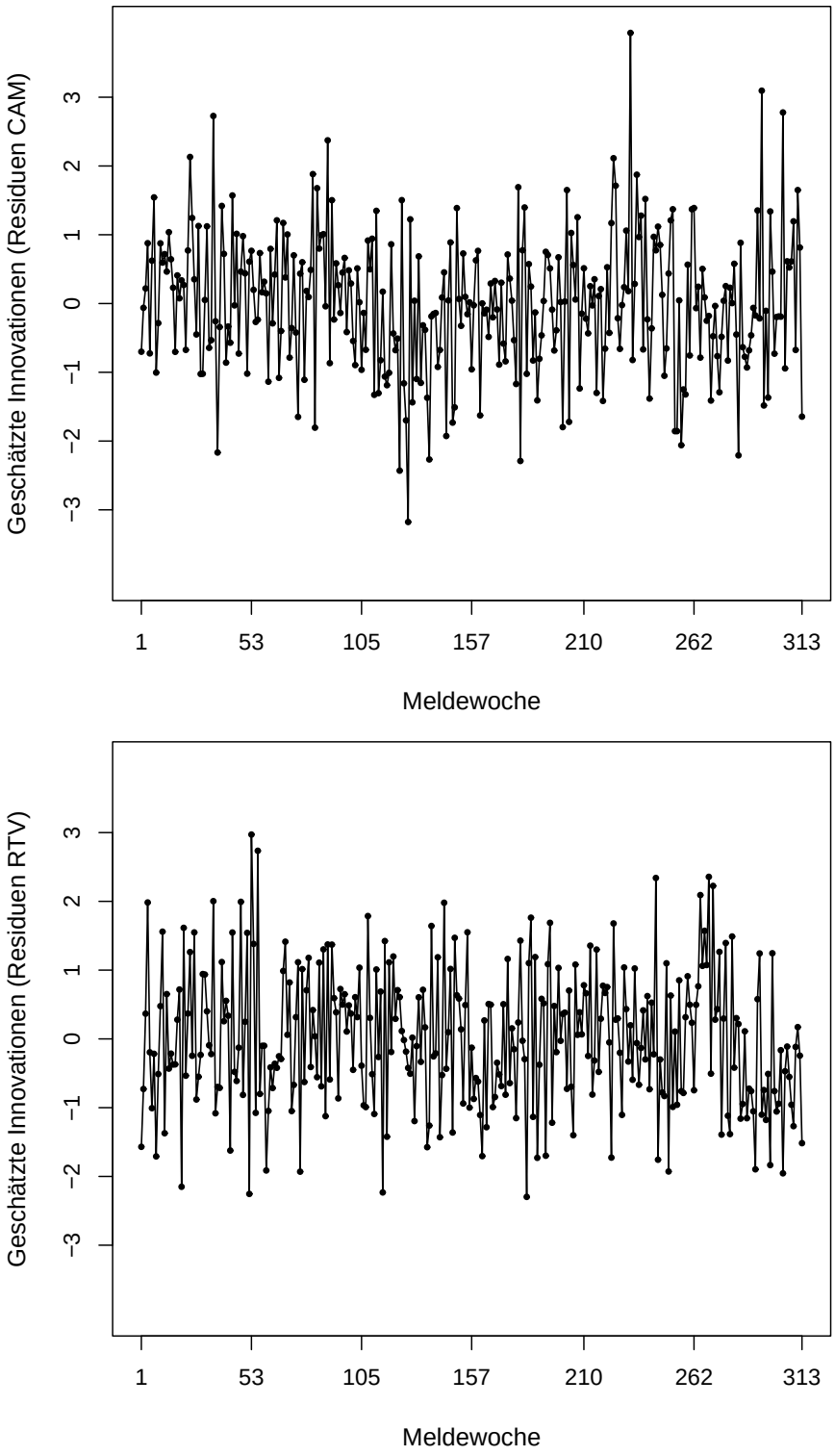

Abb. A.3: Zeitreihe der standardisierten geschätzten Innovationen nach Schätzung eines AR(1)-Prozesses für die Residuen für CAM (oben) und RTV (unten).

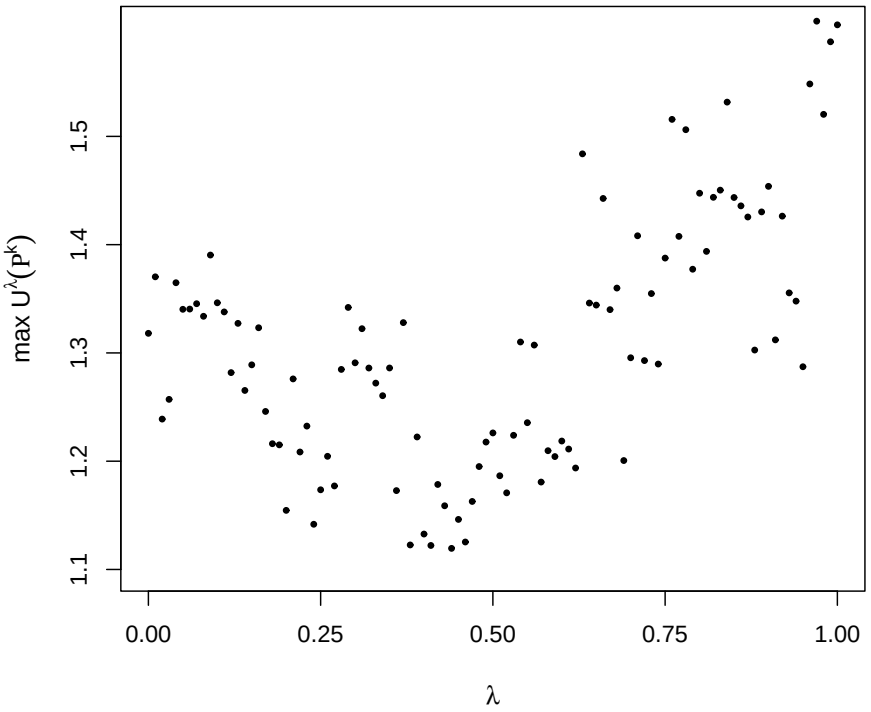

Abb. A.4: SAL: Maximum des Verhältnisses von MAD von Residuen bei disjunkter Zerlegung des Beobachtungszeitraums in 2 gleichgroße Teilbereiche mit jeweils 6 Intervallen gleicher Länge für verschiedene Werte von λ.

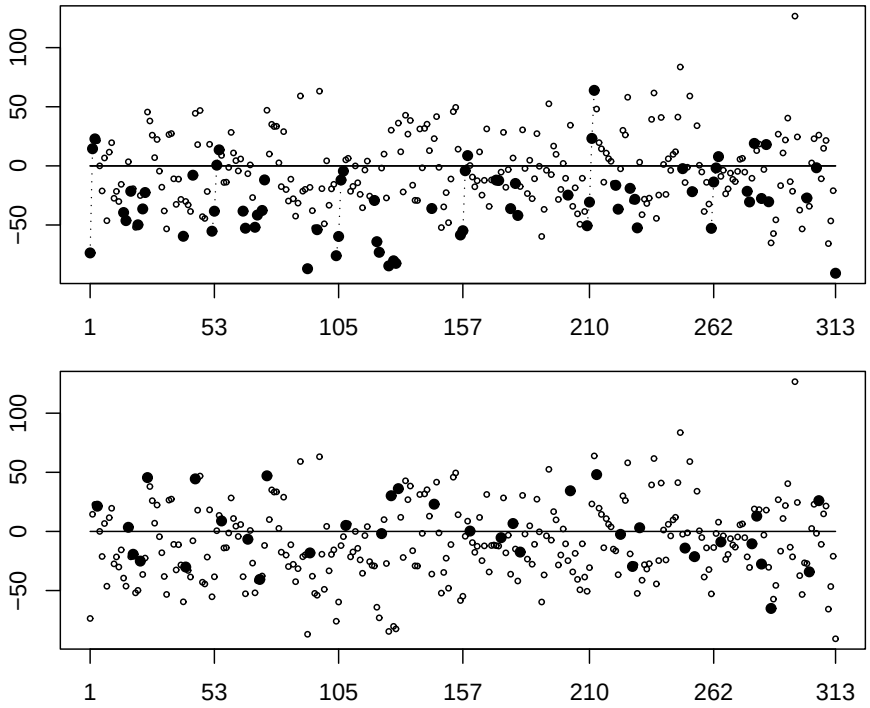

Abb. A.5: Oben: Residuen (SAL) nach Schätzung der Signalkomponente $f(t)$ durch STS. Beobachtungen aus Wochen mit Kalendereffekten wurden bei der Schätzung durch interpolierte Werte ersetzt. Die Residuen für diese Wochen sind gesondert markiert (•), zeitlich aufeinanderfolgende Residuen in der Weihnachts-/Neujahrperiode sind zusätzlich durch Linien verbunden. Unten: Residuen (SAL) nach Schätzung der Signalkomponente $f(t)$ durch STS. Beobachtungen aus Wochen mit Kalendereffekten wurden bei der Schätzung durch interpolierte Werte ersetzt. Die Residuen in Wochen ohne Feiertage, die auf Wochen mit Feiertagen folgen, sind gesondert markiert (•).

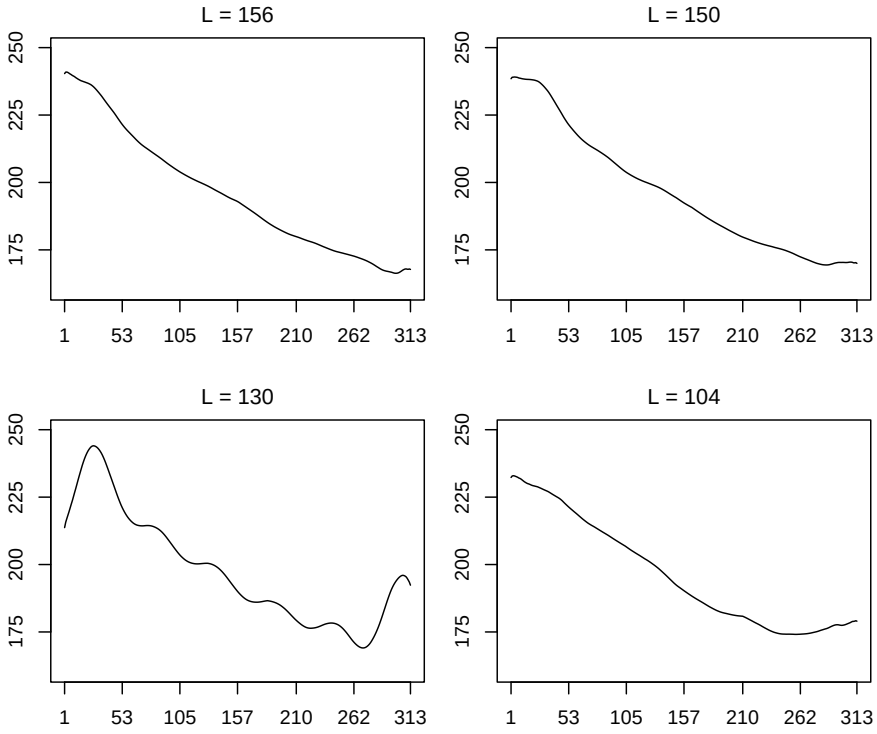

Abb. A.6: Komponeten der ersten Eigentriple nach SSA-Zerlegung der transformierten Meldefälle bei Wahl verschiedener Fensterbreiten für SAL als Kandidaten für die geschätzte Trendkomponente.

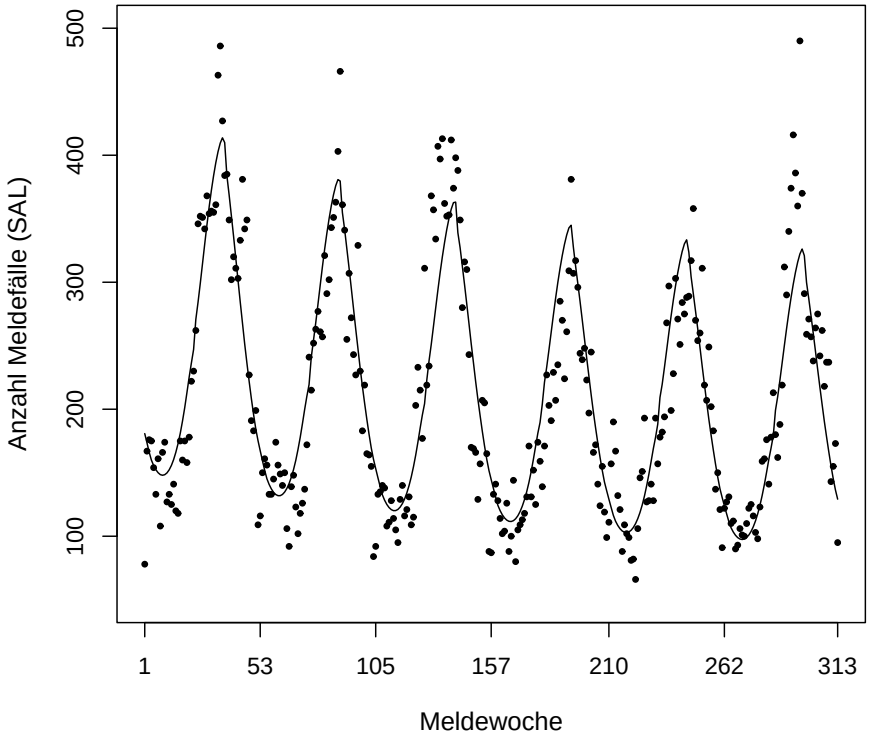

Abb. A.7: Meldefälle z_t (SAL) und ihr nur durch Trend- und Saisonkomponente geschätzter Verlauf $\hat{m}(t) + \hat{s}(t)$ nach Umkehrung der Transformation.

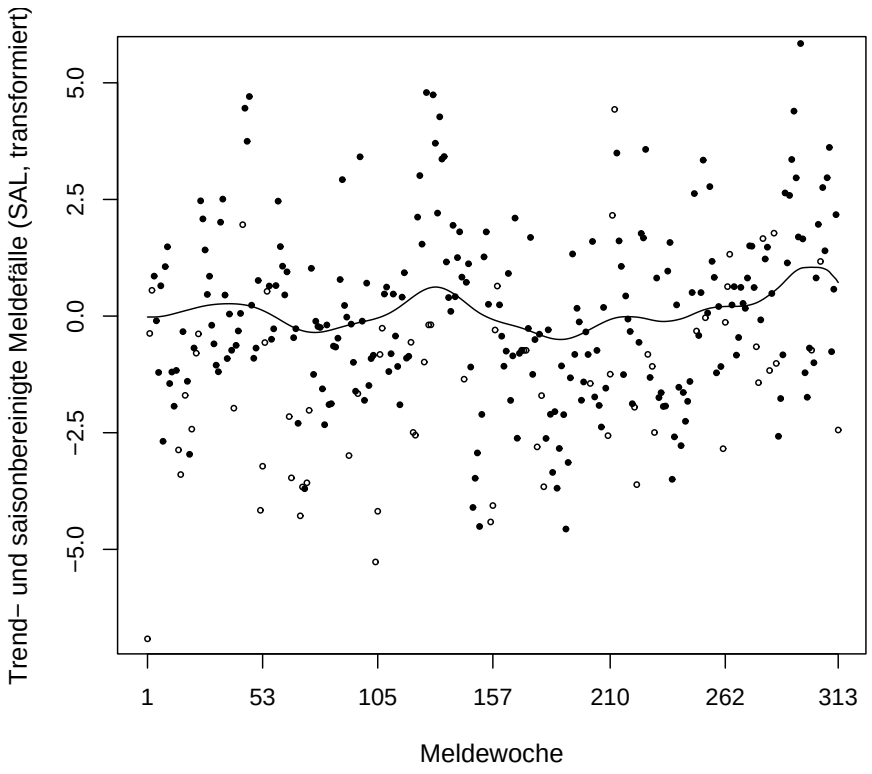

Abb. A.8: Transformierte, trend- und saisonbereinigte Meldefälle $y_t - \hat{m}(t) - \hat{s}(t)$ (•) und die geschätzte zyklische Komponente $\hat{c}(t)$ (Linie) durch SSA ($L = 156$, $i = 1$) für SAL. Beobachtungen aus Wochen mit Kalendereffekten (∘) sind dabei nicht berücksichtigt.

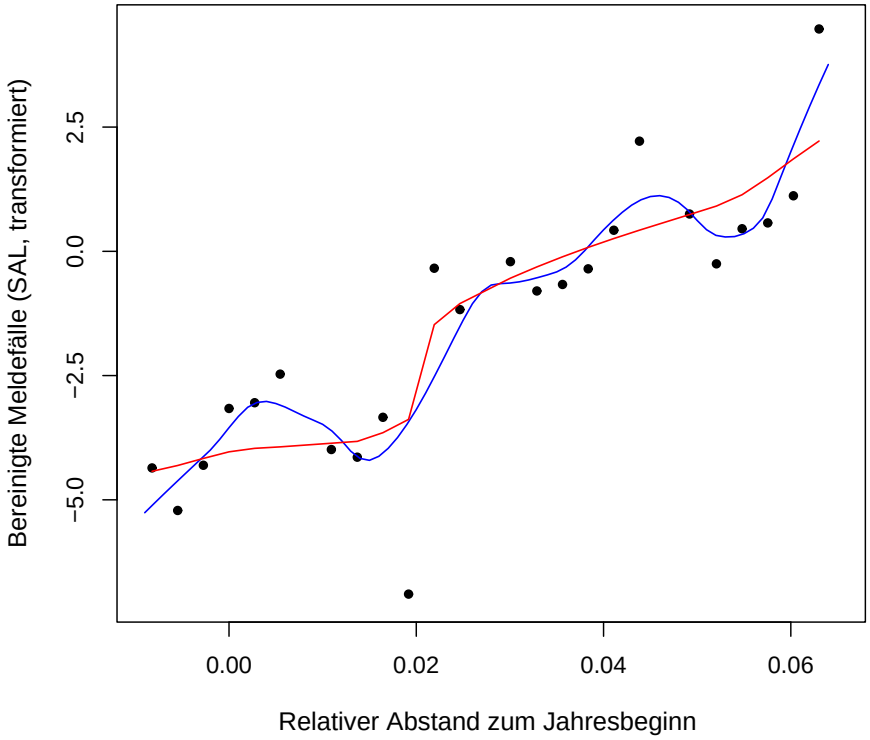

Abb. A.9: Bereinigte Meldefälle $y_t - \hat{m}(t) - \hat{s}(t) - \hat{c}(t)$ (•) aus dem Zeitraum Weihnachten / Neujahr $t \in T_W$, entsprechend dem relativen Abstand der zugehörigen Meldewochen zum Jahresbeginn, und Kandidatenfunktionen für die geschätzte Kalenderkomponente nach Verwendung des AWS- ($h_{max} = 0.09$, rot) und LOK-(blau) Verfahrens für SAL).

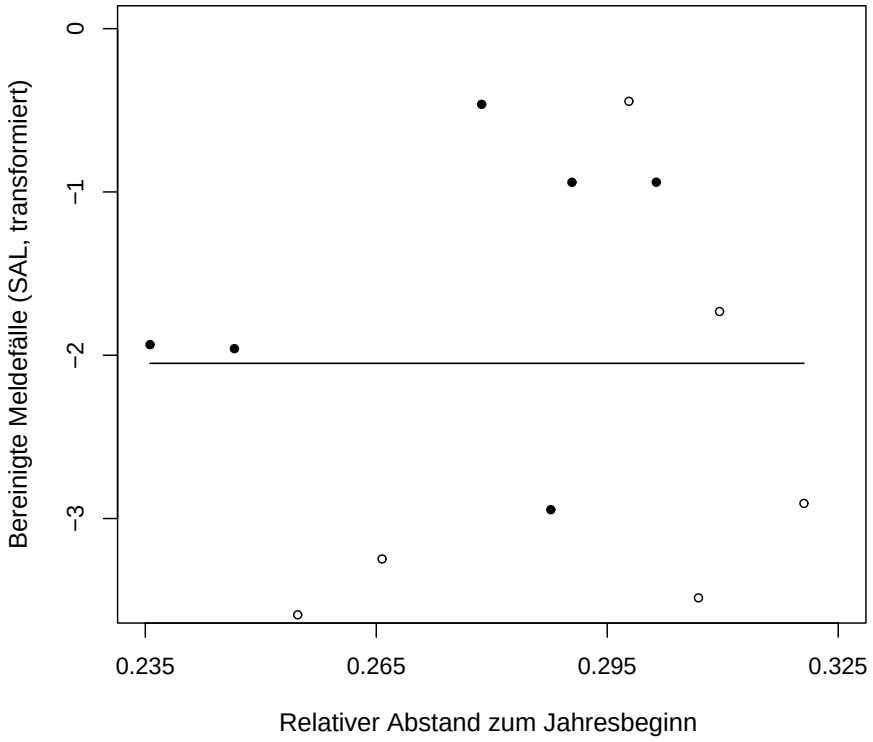

Abb. A.10: Transformierte bereinigte Meldefälle $y_t - \hat{m}(t) - \hat{s}(t) - \hat{c}(t)$ in den Wochen um Ostern $t \in T_O$ entsprechend dem relativen Abstand ihrer zugehörigen Meldewochen zum Jahresbeginn für SAL. Meldefälle aus den Wochen vor Ostern sind durch (•), die aus den darauffolgenden durch (∘) gekennzeichnet. Die Linie markiert das arithmetische Mittel aller gezeigten Beobachtungen.

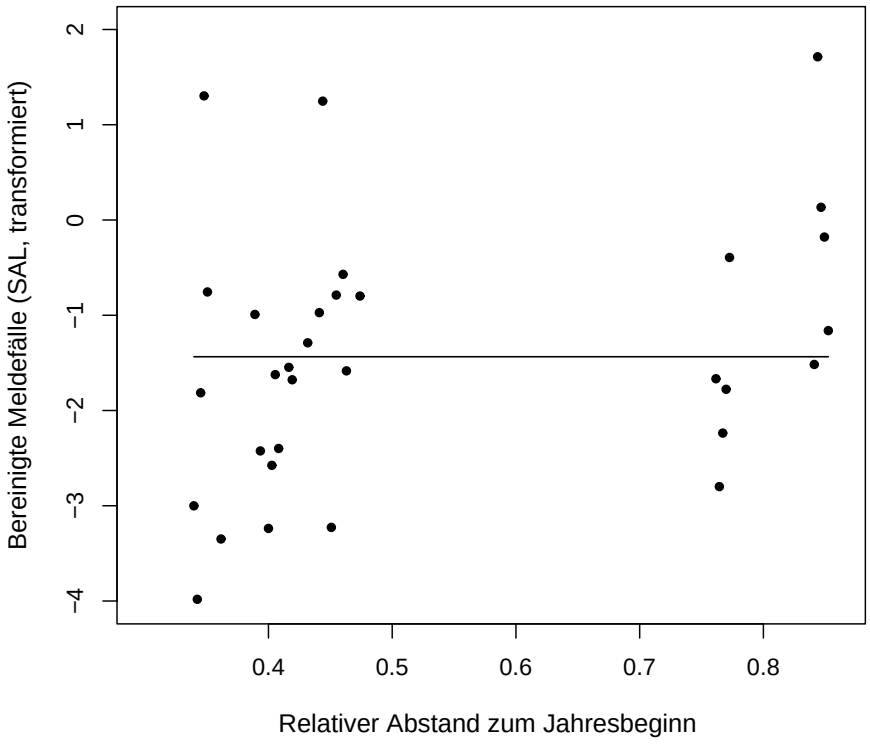

Abb. A.11: Transformierte bereinigte Meldefälle $y_t - \hat{m}(t) - \hat{s}(t) - \hat{c}(t)$ in den Wochen mit einzelnen Feiertagen $t \in T_R$ entsprechend dem relativen Abstand ihrer zugehörigen Meldewochen zum Jahresbeginn für SAL. Die Konstante markiert das arithmetische Mittel der gezeigten Beobachtungen.

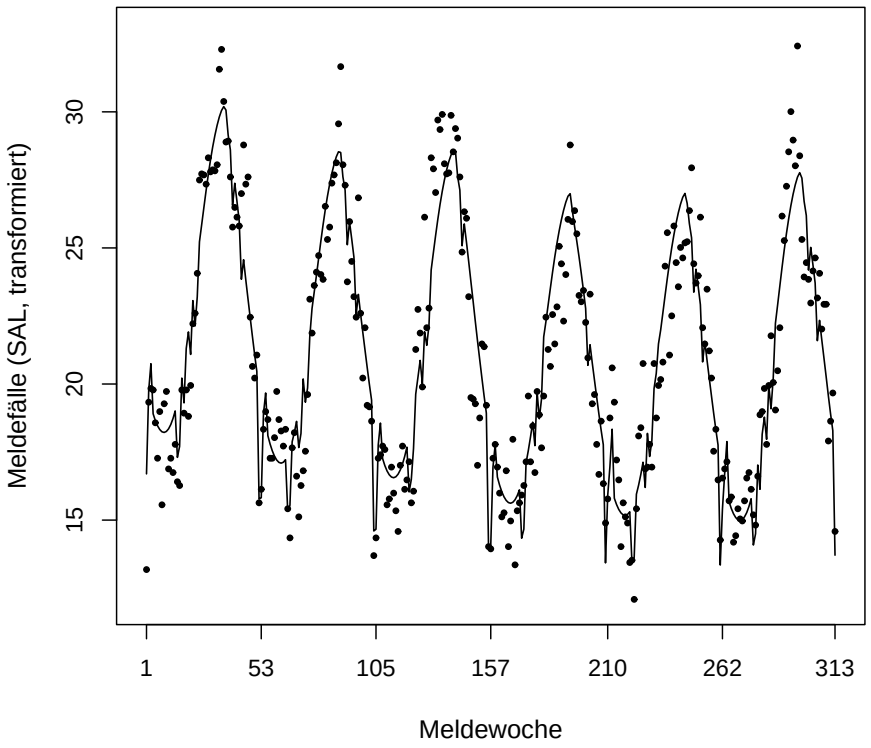

Abb. A.12: Beobachtete (transformierte) Meldefälle (•) und ihre Schätzung durch die Summe der geschätzten Dekompositionskomponenten (Linie) für SAL.

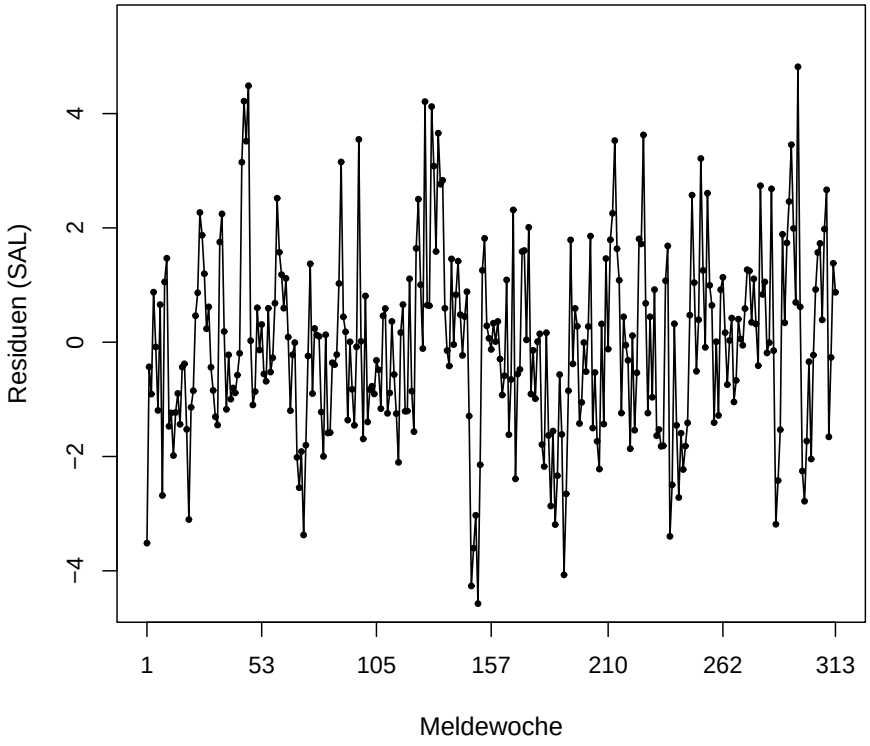

Abb. A.13: Zeitreihe der Residuen $r_t = y_t - \hat{m}(t) + \hat{s}(t) + \hat{c}(t) + \hat{k}(t)$ *für SAL.*

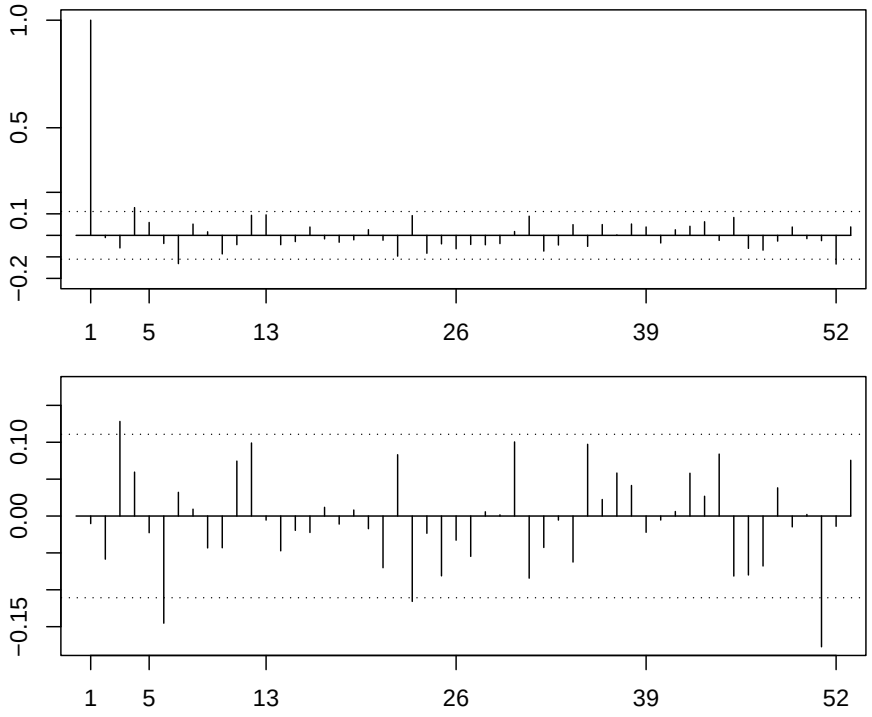

Abb. A.14: Autokorrelationsfunktion (oben) und partielle Autokorrelationsfunktion (unten) mit 5% Konfidenzgrenzen der geschätzten Innovationen $\hat{x}(t)$ nach Anpassung eines ARMA(2,1)-Prozesses für die Residuen für SAL.

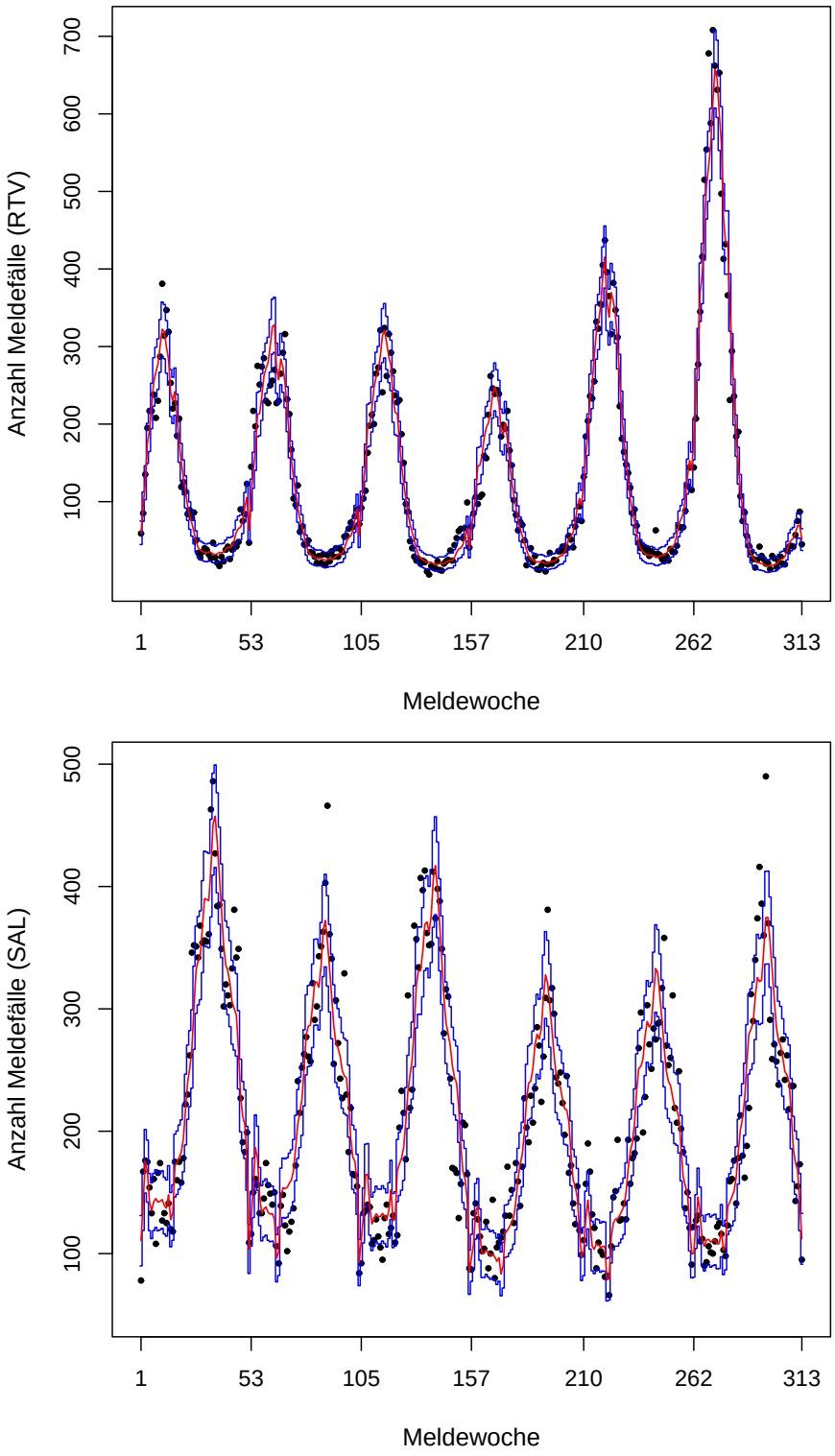

Abb. A.15: Geschätzter Verlauf (rot) nach Poissonregression und 95%-Vertrauensbereiche (blau) für die Zahl der Meldefälle bei RTV (oben) und SAL (unten).

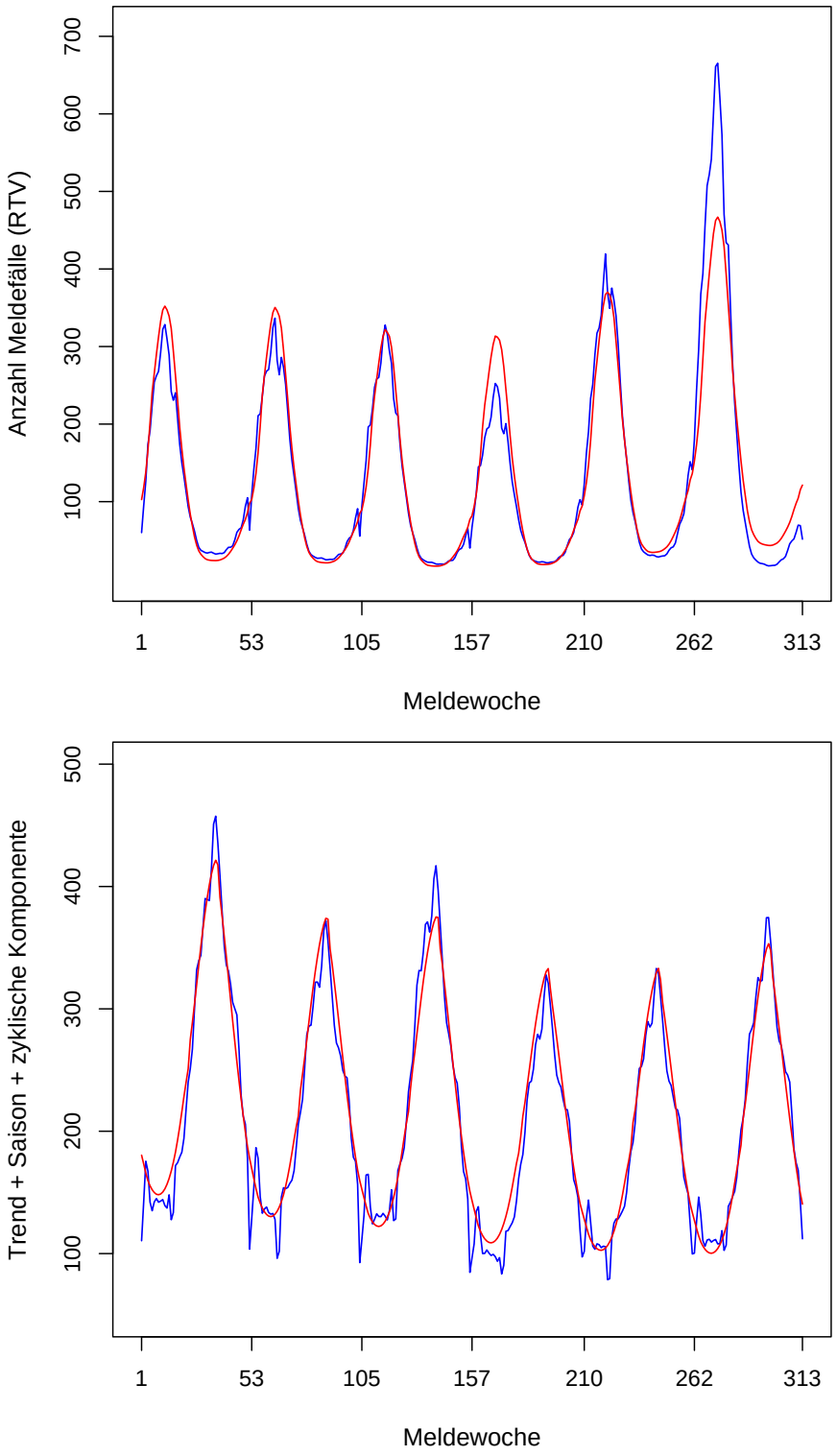

Abb. A.16: Summe der geschätzen Trend-, Saison- und zyklischen Komponenten im nichtparametrischen (rot) und parametrischen (blau) Modell für RTV (oben) und SAL (unten).

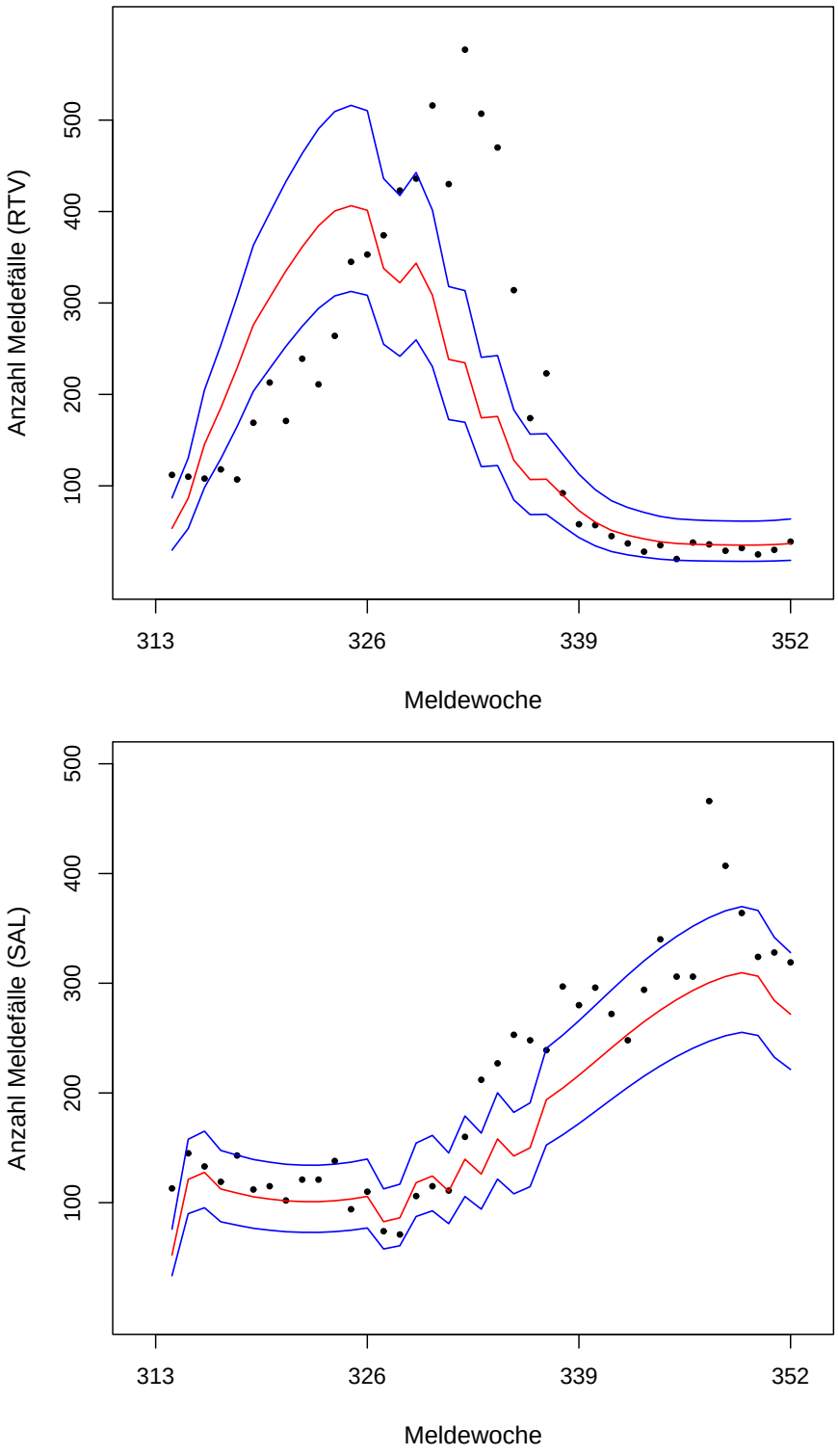

Abb. A.17: Beobachtete Meldefälle (•) und geschätzte Vorhersage (rot) sowie Vertrauensintervalle zum Niveau 95% (blau) auf Grundlage des nichtparametrischen Verfahrens für die Wochen 314 - 352 (Januar - September 2007) für RTV (oben) und SAL (unten).

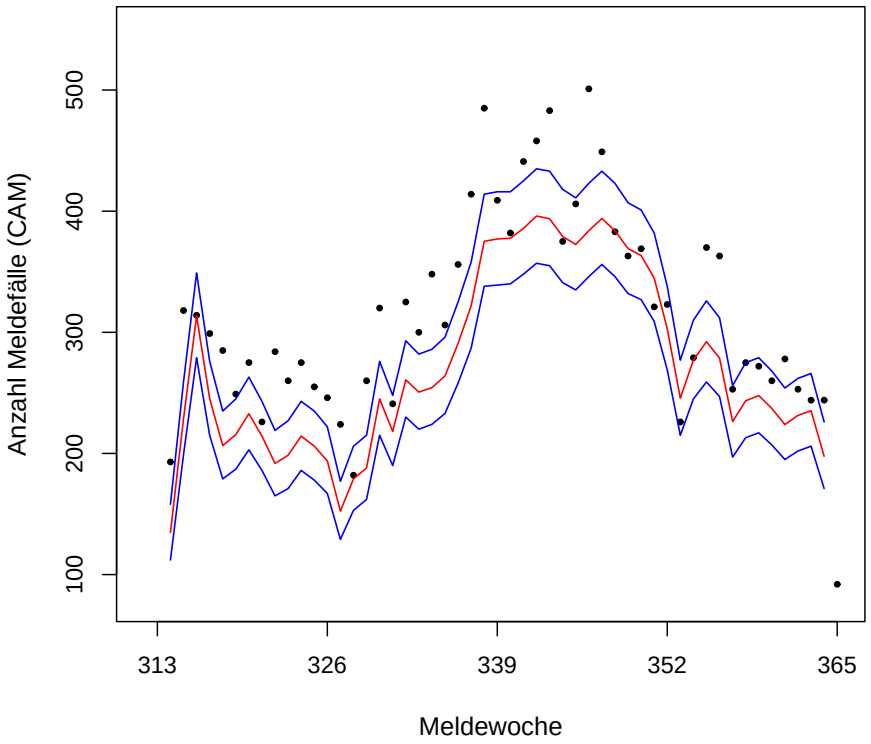

Abb. A.18: Beobachtete Meldefälle (•) und geschätzte Vorhersage (rot) sowie Vertrauensintervalle zum Niveau 95% (blau) auf Grundlage des Poissonregressionsmodells für die Wochen 314 - 352 (Januar - September 2007, CAM) .