Beispiel in R: ZR mit Trend

Regina Tüchler & Thomas Rusch

November 2, 2009

Beispiel: Holt Verfahren für die Hotel-Daten :
Die Daten `hotelts-annual.rda` werden geladen. Sie müssen sich dafür im working directory befinden.

```
> load("hotelts-annual.rda")
```

Die Daten sind jetzt im Arbeitsspeicher

```
> ls()
```

und sind unter dem Namen `hotann` verfügbar. Es handelt sich um eine ZR für die Jahre 1973 bis 2003 mit jährliche Nächtigungszahlen in Österreich in 4/5 Sterne Hotels.

1. Schritt: Zeitreihenplot wird in Fig. 1 gezeichnet.

```
> plot(hotann)
```

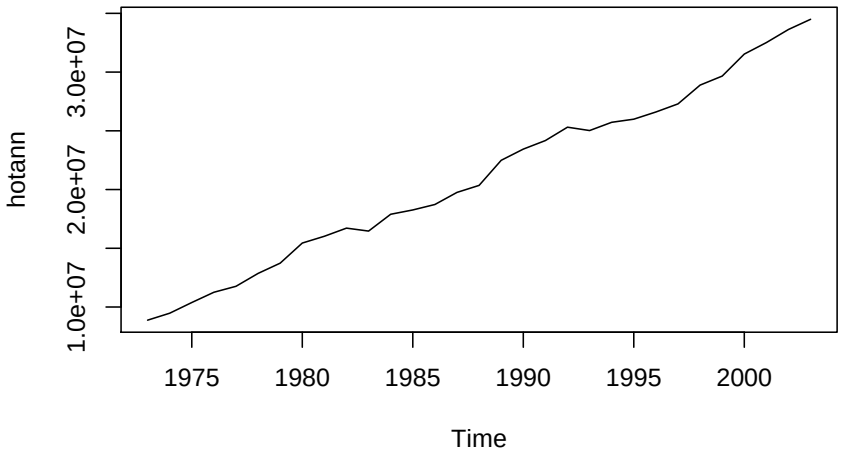

Figure 1: Zeitreihenplot Hotel-Daten.

2. Schritt: Analyse des Zeitreihenplots:
Welche Strukturen sind in Fig. 1 zu erkennen? Sollen die Daten transformiert werden? Gibt es Ausreißer? Sind Brüche erkennbar?

Der Trend in dieser ZR kann gut durch eine lineare Trendfunktion modelliert werden. Die Daten müssen nicht transformiert werden. Es sind keine Brüche oder Ausreißer erkennbar.

In den LV-Folien haben wir die Jahre 1973-2000 füer die Prognose von 2001-2005 verwendet. Die dafür notwendigen Befehle sind die folgenden:

3. Schritt: Die Daten der Jahre 1973-2000 sollen Input für das Holt Verfahren sein, um eine Prognose für die Jahre 2001-2005 zu erstellen. Die Daten werden in `past` und in `future` geteilt:

```
> past <- window(hotann, end = 2000)
> future <- window(hotann, start = 2001)
```

Für die beiden Glättungsparameter wird der Wert 0.6 bzw. 0.2 angenommen. Die Argumente in der R-Funktion `HoltWinters` sind die Daten der Jahre 1973-2000: `past`, die beiden Glättungsparameter `alpha=0.6`,`beta=0.2`. Um das Holt Verfahren auszuwählen, muss der Wert des Parameters `gamma` auf `FALSE` gesetzt werden! Der Output wird im Parameter `model` gespeichert.

```
> model <- HoltWinters(past, alpha = 0.6, beta = 0.2, gamma = FALSE)
```

Sämtliche Outputparameter, die im Objekt `model` gespeichert wurden, können angezeigt werden:

```
> str(model)
```

Der Zugriff auf die einzelnen Elemente des Objekts `model` erfolgt immer unter Angabe des Objekts. Wir verwenden hauptsächlich `model$fitted`, `model$coeff` und `model$SSE`: Mit

```
> model$fitted
```

```
Time Series:
Start = 1975
End = 2000
Frequency = 1
         xhat    level    trend
1975 10065820  9472823 592997.0
1976 10887177 10256120 631056.9
1977 11784560 11109116 675444.7
1978 12444512 11771649 672862.5
1979 13416493 12693777 722715.6
1980 14367476 13606716 760760.2
1981 15904408 15014286 890122.2
1982 16880883 15976369 904514.3
1983 17666123 16781488 884635.2
1984 17685199 16944823 740375.3
1985 18577436 17811750 765685.6
1986 19116810 18388843 727967.0
1987 19564575 18883309 681266.8
1988 20383879 19679606 704272.9
1989 21065516 20365016 700500.2
1990 22790964 21919639 871324.8
1991 24131318 23181821 949496.4
1992 25109195 24154969 954226.6
1993 26206593 25228504 978088.4
1994 26331458 25495573 835884.5
1995 26730211 25967182 763029.3
```

```
1996 26960530 26286286 674244.3
1997 27376660 26745435 631225.2
1998 27947110 27326014 621096.0
1999 29248494 28514016 734477.3
2000 30279040 29495217 783822.0
```

erhält man spaltenweise die ZR der geglätteten Werte $\hat{x}_{t-1}(1)$: `xhat`, die eine Summe aus der Niveaukomponente $\hat{a}_t$: `level` und der Trendkomponente $\hat{b}_t$: `trend` sind. Mit

```
> model$SSE
```

```
[1] 1.107619e+13
```

kann die Fehlerquadratsumme ausgegeben werden.

Von `HoltWinters` werden automatisch die Koeffizienten für zukünftige Prognosen ausgegeben: `model$coeff`. Für die Niveaukomponente $\hat{a}_{T+1}$: `a =` 31031763.2 und für die Trendkomponente $\hat{b}_{T+1}$: `b =` 934366.77.

Wie wir aus den Folien der LV wissen, erfolgt die h-Schritt-Prognose vom Jahr 2000 ausgehend nach

$$\hat{x}_{2000}(h) = \hat{a}_{2001} + \hat{b}_{2001}h, \quad h = 1, 2, \ldots, 5.$$

Die Werte von $\hat{a}_{2001}$ bzw. $\hat{b}_{2001}$ stehen in

```
> model$coeff
```

```
         a          b
31031763.2   934366.8
```

Die Prognose für das Jahr 2001 ergibt sich aus der Summe dieser beiden Koeffizienten: $x_{2000}(1) = \hat{a}_{2001} + \hat{b}_{2001} = 31966130$,

für 2002: $x_{2000}(2) = \hat{a}_{2001} + \hat{b}_{2001} \cdot 2 = 32900497$, usw. für 2003, 2004 und 2005.

Diese Werte köennen mit `R` über die Funktion `predict` automatisch berechnet werden. Die Vorhersage für 2001-2005:

```
> progn <- predict(model, n.ahead = 5, prediction.interval = TRUE)
```

```
Time Series:
Start = 2001
End = 2005
Frequency = 1
          fit      upr      lwr
2001 31966130 33252257 30680003
2002 32900497 34485306 31315688
2003 33834864 35752875 31916852
2004 34769230 37050284 32488176
2005 35703597 38374285 33032909
```

Im Object `progn` sind jetzt die prognostizierten Werte in der Spalte `fit` enthalten. Wenn man die Option `prediction.interval = TRUE` dazuschreibt, werden die obere und untere Schranke für die 95%-Prognoseintervalle unter `upr` bzw. `lpr` mitberechnet.

Mit dem Befehl `plot(model)` wird die ZR gemeinsam der geglätteten ZR gezeichnet. Wenn man auch die prognostizierte ZR für 2001-2005 einzeichnen möchte, gibt man im Plot-Befehl als Argument auch den Namen an, unter dem die Prognosewerte gespeichert wurden (bei uns: `progn`). In Fig. 2 wird die ZR gemeinsam mit der geglätteten ZR und der Prognose gezeichnet:

Wir sehen in Fig. 2, dass die ZR für die Jahre 1973-2000 mit dem Holt Verfahren gut geglättet werden konnte. Der Verlauf der ZR in den Jahre 2001-2003 konnte gut vorhergesagt werden. Für

```
> plot(model, progn)
> lines(hotann)
```

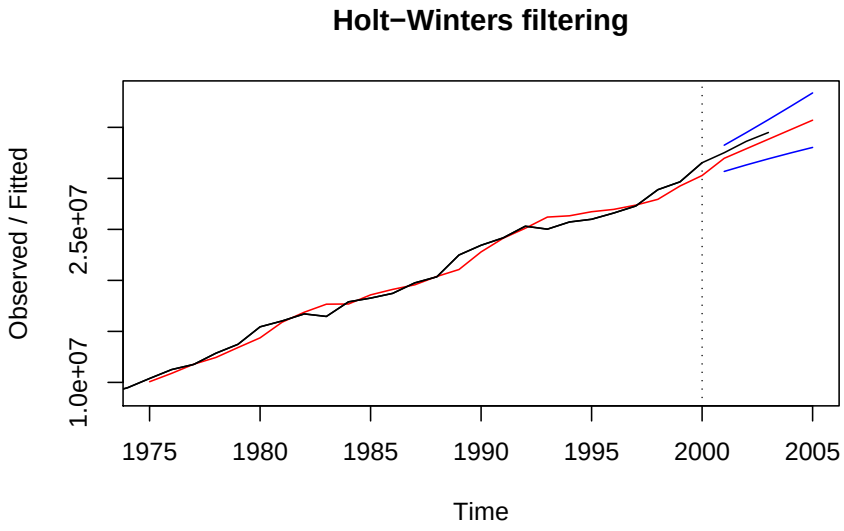

Figure 2: Holt Verfahren für die Hotel-Daten.

die Prognose ist `a = 31031763.2` der Koeffizient für die Niveaukomonente und `b = 934366.77` für die Trendkompnente. In Fig. 2 sieht man die entsprechende prognostizierte Gerade: $31031763.2 + h \cdot 934366.77$, $h = 1, \ldots, 5$, für die Jahre 2001-2005.

4. Schritt Modellwahl, Residuenanalyse:

Es stehen verschiedene Methoden zur Verfügung, um zwischen Modellen zu wählen. Einige davon sollen hier durchgeführt werden. Wir nehmen dazu an, dass das obige Modell mit einem Modell, das nur *Niveau* aber *keinen Trend* enthält, verglichen werden soll. Die Schätzung erfolgt mit *exponentiellem Glätten*:

```
> model.niveau <- HoltWinters(past, beta = FALSE, gamma = FALSE)
> progn.niveau <- predict(model.niveau, n.ahead = 5, prediction.interval = TRUE)
```

In der LV wurde die Fehlerquadratsumme im *Beobachtungszeitraum* 1975-2000 minimiert, um die optimalen Glättungsparameter zu wählen. Die Fehlerquadratsumme ist auch geeignet, um hier die beiden Modelle zu vergleichen. Hier haben wir für die Schätzung mit dem Holt Verfahren einen Wert von

```
> model$SSE

[1] 1.107619e+13
```

und für das exponentielle Glätten

```
> model.niveau$SSE

[1] 2.724447e+13
```

Wie zu erwarten war, ist das erste Modell klar vorzuziehen.

Eine zweite Möglichkeit für einen Vergleich ist, für den *Prognosezeitraum* 2001-2003 die mittlere Fehlerquadratsumme zwischen wahren Werten und geschätzten Werten zu berechnen (Da wir für die Jahre 2001-2003 in diesem Beispiel die wahren Werte kennen.) Die mittlere Fehlerquadratsumme ergibt sich aus: $MSE = \frac{1}{3}\sum_{t=2001}^{2003}(x_t - \hat{x}_t)^2$.

In der "EinfinR.pdf" haben wir von verschiedenen Datenstrukturen in `R` gehört, und dass verschiedene Operationen nur mit bestimmten Datenstrukturen funktionieren. Um die notwendigen Operationen zur Berechnung von MSE in `R` durchzuführen, werden die `R`-ZR `future` (wahren Werte für 2001-2003) und die `R`-ZR `progn` bzw. `progn.niveau` (die Schätzungen für 2001-2003) als `R`-Matrix gespeichert.

Dafür steht die Funktion `matrix` zur Verfügung. Die Argumente in `matrix` sind: die ZR, die Anzahl der Spalten (hier 1), die Anzahl der Zeilen (hier 3):

```
> true <- matrix(future, 3, 1)
> est <- matrix(progn[1:3, 1], 3, 1)
> est.niveau <- matrix(progn.niveau[1:3, 1], 3, 1)
```

Für die mittlere Fehlerquadratsumme muss man die Spaltensumme (`colSums`) durch 3 dividieren.

```
> mse <- colSums((est - true)^2)/3

[1] 4.21642e+11

> mse.niveau <- colSums((est.niveau - true)^2)/3

[1] 4.705774e+12
```

Auch aus diesen Berechnungen wird deutlich, dass die Trendkomponente einbezogen werden soll.

```
> plot(model.niveau, progn.niveau)
> lines(hotann)
```

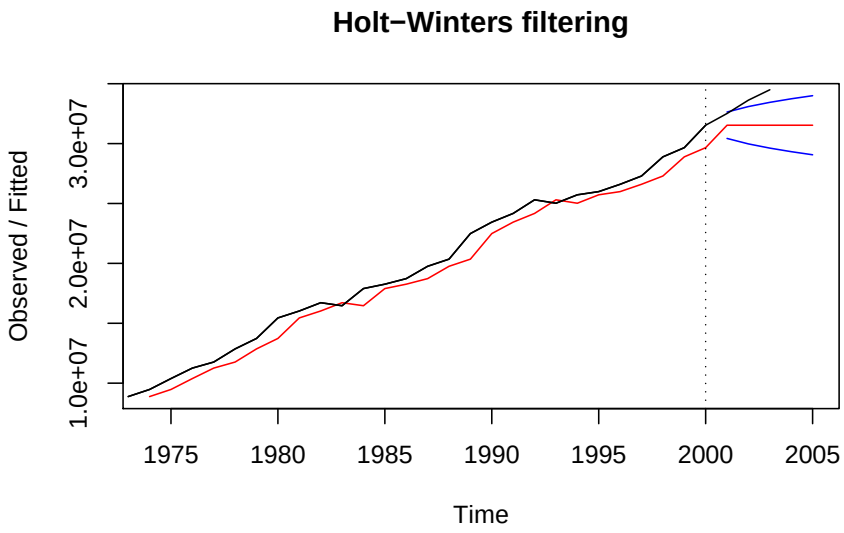

Figure 3: Exponentielles Glätten der Hotel-Daten.

Natürlich zeigt hier auch ein Vergleich der beiden Grafiken Fig. 2 und Fig. 3 deutlich, dass das Holt Verfahren klar besser für diese Daten geeignet ist.

Man kann die Residuen der Jahre 1973-2000: $x_t - \hat{x}_t$ auch als ZR auffassen und davon einen Plot machen:

```
> plot(past - model$fitted[, "xhat"])
> lines(past - model.niveau$fitted[, "xhat"], col = "red")
```

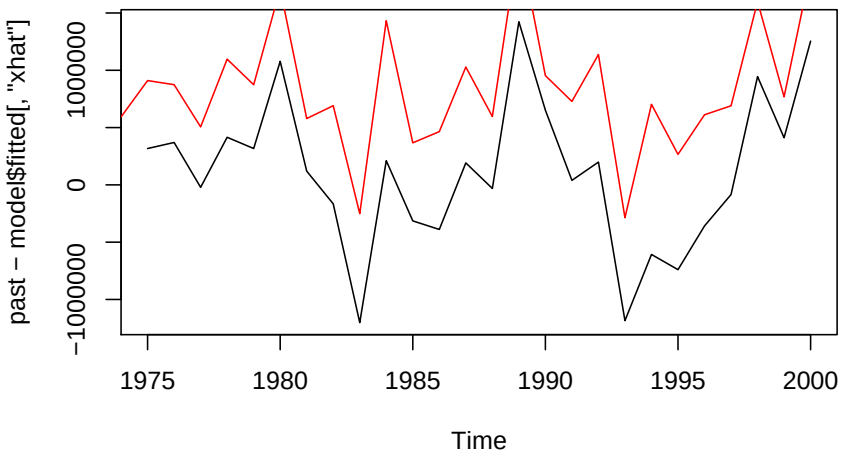

Figure 4: Vgl. d. Residuen.

In Fig. 4 liegt die rote Linie (- das ist jene, wo kein Trend im Modell einbezogen wurde) immer über der schwarzen Linie (- Niveau und Trend eingezogen). Also ist der Fehler zwischen wahren Werten und Schätzwerten für ein Modell ohne Trend für alle Beobachtungszeitpunkte größer.

Man beachte im Argument des Plotaufrufes, dass hier eine neue ZR `past-model$fitted[,"xhat"]` als Differenz von 2 ZR mit unterschiedlichem Beginnzeitpunkt gebildet wird. Die ZR `past` startet in 1973 und endet in 2000, während die ZR `model$fitted[,"xhat"]` erst im Jahr 1975 beginnt. `R` wählt hier automatisch den Überschneidungszeitraum 1975-2000 für die Differenzberechnung.