DISCUSSION PAPERS IN STATISTICS
AND ECONOMETRICS

SEMINAR OF ECONOMIC AND SOCIAL STATISTICS
UNIVERSITY OF COLOGNE

No. 5/95

Nichtparametrische Analyse parametrischer Wachstumsfunktionen – Eine Anwendung auf das Wachstum des globalen Netzwerks Internet

von

Klaus Brachmann*

März 1995

Zusammenfassung: Besonders in der betriebswirtschaftlich relevanten Marktanalyse besteht ein großer Bedarf an möglichst einfachen Prognoseverfahren z. B. in Form von endogenen, d. h. alleine von der Zeit abhängigen, Wachstumsfunktionen. Der hier vorgestellte Test dient dazu, diejenigen Funktionen auszuwählen, welche den vorliegenden Sachverhalt hinreichend gut zu beschreiben vermögen. Dabei zeigt sich in einer empirischen Anwendung, daß das Wachstum des globalen Netzwerks Internet tatsächlich durch Exponential– bzw. logistische Funktionen zu beschreiben ist.

*) Seminar für Wirtschafts– und Sozialstatistik, Universität zu Köln, Albertus–Magnus–Platz, 50923 Köln, Deutschland; Tel: +49/221/470 4129, Fax: +49/221/470 2816, e–mail: brach@wiso.uni-koeln.de

1 Einleitung

In der Regressionsanalyse gibt es verschiedene Maße, mit denen man die Angemessenheit eines parametrischen Ansatzes bei dem untersuchten Datensatz beurteilen kann. Meistens basieren sie auf einer Analyse der bei der Regression sich ergebenden Residuen.

In dieser Arbeit wird versucht, basierend auf einem Ansatz von Mammen (1992) eine Prüfung eines parametrischen Ansatzes mit Hilfe eines nichtparametrischen Schätzers durchzuführen. Die Idee dieses Verfahrens stammt bereits aus dem Bereich der Dichteschätzung und ist zunächst von Bickel und Rosenblatt (1973) vorgestellt worden (Power–Untersuchungen zu diesem Test bei Ghosh und Huang, 1991).

Der Test soll zunächst allerdings nur für den einfachen zweidimensionalen Fall vorgestellt werden, da die Verwendung nichtparametrischer Schätzungen für diesen Fall besonders gut untersucht ist. Ein entsprechendes parametrisches Analogon stellen die bekannten endogenen Wachstumsfunktionen dar, die mit der Variable Zeit ebenso nur eine exogene Größe unterstellen (Erweiterungen zu den grundlegenden Modellen bei Mertens und Falk (1994)).

Es liegt nun nahe, ein quadratisches Abstandsmaß zwischen der parametrischen und der nichtparametrischen Schätzung zu verwenden. In dieser Arbeit soll der integrierte quadratische Abstand verwendet werden. Dieser folgt, wie bei Mammen (1992) gezeigt wird, approximativ einer Normalverteilung. Dabei weist Mammen allerdings bereits selber auf eine sehr langsame Konvergenz hin. In dieser Arbeit wird diese theoretische Erkenntnis durch Simulationen gestützt. Danach ist die Konvergenz derart langsam, daß im Rahmen empirischer Untersuchungen mit entsprechend begrenzter Zahl von Beobachtungen zur Berechnung kritischer Werte nicht verläßlich auf die Normalverteilung zurückgegriffen werden kann. Es zeigt sich, daß eine bestimmte Form des Bootstrap–Verfahrens verläßlichere Aussagen über kritische Werte treffen kann.

Zur konkreten Anwendung dieses Verfahrens wird das Wachstum des globalen Kommunikationsnetzwerks Internet untersucht. Tatsächlich ist das durchgeführte Verfahren in der Lage, zwischen verschiedenen unterstellten Wachstumsfunktionen zu differenzieren.

Im nächsten Abschnitt wird zunächst die Ausgangssituation formal dargestellt und die Testgröße T_n hergeleitet. Der dritte Abschnitt enthält dann eine Untersuchung der Konvergenz durch Simulation. Der vierte Abschnitt dient der Darstellung der zwei überprüften

Bootstrap–Verfahren. Im fünften Abschnitt erfolgt eine punktweise Berechnung der Gütefunktion für ein konkretes simuliertes Beispiel. Im letzten Abschnitt werden die Ergebnisse des empirischen Beispiels dargestellt und interpretiert.

2 Ausgangssituation

Gegeben seien n Paare von Beobachtungen $(X_i, Y_i)_{i=1}^n$, wobei $X, Y \in I\!R$. Dem liege eine unbekannte Regressionsfunktion $m(x) = E(Y|X = x)$ zugrunde. Ziel ist es nun, eine möglichst gute Schätzung $\hat{m}(x)$ für die wahre aber unbekannte Regressionsfunktion zu erhalten. Dies setzt im parametrischen Fall die richtige Wahl eines entsprechenden parametrischen Modells $\{m_\theta : \theta \in \Theta\}$ voraus. Ob die Auswahl korrekt ist, kann der hier verwendete Test analysieren. Getestet wird also:

$$H_0 : \ m \in \{m_\theta : \theta \in \Theta\}$$

Die Alternative ist, daß es sich bei $m(x)$ zunächst nur um eine beliebige glatte Funktion handelt.

Die parametrische Schätzung erfolge mit Hilfe des KQ–Schätzers, wobei der Parametervektor $\hat{\theta}$ durch folgende Minimierung zu bestimmen ist:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (Y_i - m_\theta(X_i))^2$$

Die "Referenzfunktion" wird durch einen nichtparametrischen Regressionsschätzer bereitgestellt. Hier soll der Schätzer von Nadaraya–Watson (1964) verwendet werden:

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n K_h(x - X_i) Y_i}{\sum_{i=1}^n K_h(x - X_i)}$$

Dabei wurde aus Gründen der Notation K_h als transformierte Kernfunktion eingesetzt:

$$K_h(x - X_i) = \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

Aufgabe des dargestellten Tests wird es nun sein, festzustellen, ob ein bestimmter Unterschied zwischen $m_{\hat{\theta}}$ und $\hat{m}_h$ auf zufällige Schwankungen zurückzuführen ist, oder ob tatsächlich das unterstellte parametrische Modell den vorgegebenen Sachverhalt nicht zutreffend beschreibt. Ein intuitiv eingänglicher Weg ist die Verwendung des L_2–Abstandes zwischen $m_{\hat{\theta}}$ und $\hat{m}_h$:

$$T_n^* = \int \left(\hat{m}_h(x) - m_{\hat{\theta}}(x)\right)^2 dx$$

Aufgrund des Bias des nichtparametrischen Schätzers ist T_n^* auch bei Gültigkeit der Nullhypothese nicht Null. Deshalb sei eine bereits analog bei Bickel und Rosenblatt (für den Fall der Dichteschätzung) untersuchte Größe verwendet. Danach betrachtet man den L_2–Abstand zwischen $\hat{m}_h$ und $E_0(\hat{m}_h)$. $E_0(\hat{m}_h)$ ist dabei der Erwartungswert des nichtparametrischen Schätzers unter der Nullhypothese. Folgender Glättungsoperator $\mathcal{S}_{h,n}$ sei im folgenden eingeführt:

$$\mathcal{S}_{h,n}t(x) = \frac{\sum_{i=1}^{n} K_h(t - X_i)t(X_i)}{\sum_{i=1}^{n} K_h(t - X_i)}$$

Da offensichtlich $E\left(\hat{m}_h(x)|(X_1, X_2, \ldots, X_n)\right) = \mathcal{S}_{h,n}m(x)$ gilt, ergibt sich folgende Testgröße:

$$T_n = n\sqrt{h} \int \left(\hat{m}_h(x) - \mathcal{S}_{h,n}m_{\hat{\theta}}(x)\right)^2 dx$$

3 Asymptotik von T_n

Unter bestimmten Voraussetzungen (siehe hierzu Mammen (1992)) konvergiert die Verteilung von T_n unter der Nullhypothese gegen eine Normalverteilung $N(b_h, V)$. Dabei gilt für die Parameter:

$$b_h = \frac{1}{\sqrt{h}} \int K^2(t)dt \int \frac{\sigma^2(x)}{f(x)}dx$$

$$V = 2 \int K^4(t)dt \int \frac{\sigma^4(x)}{f^2(x)}dx$$

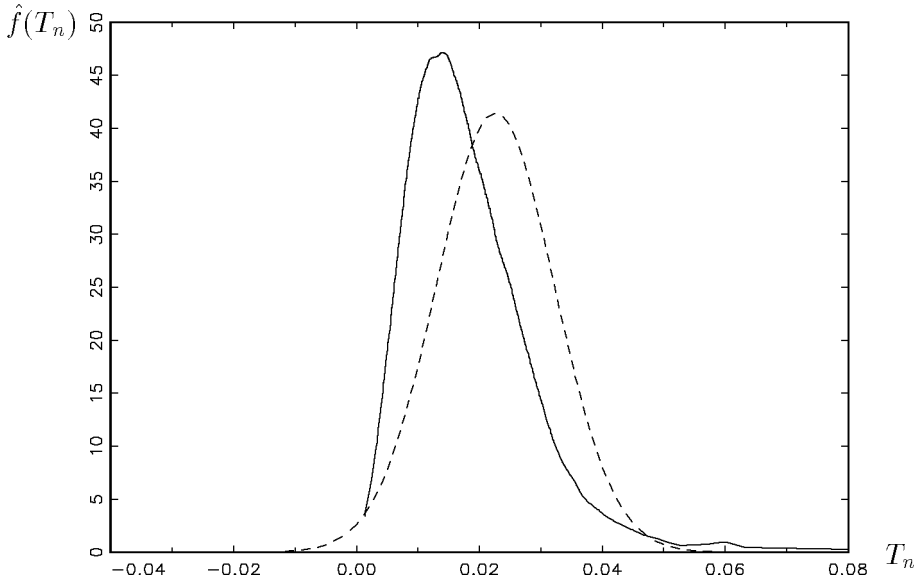
Der ausführliche Beweis ist bei Mammen (1992) nachzulesen.

Wie schnell die Konvergenz stattfindet, läßt sich durch Simulationen feststellen. Dazu sei folgendes Beispiel verwendet. Die Zufallsvariable X sei gleichverteilt auf dem Intervall $[0, 1]$. $m(x)$ sei gleich x und für die bedingte Varianz gelte $\sigma^2(x) = 0.01$. Y sei dementsprechend normalverteilt mit $\mu = x$ und $\sigma^2 = 0.01$. Gleichzeitig wird ebenfalls eine lineare Abhängigkeit des Y von X unterstellt, mithin trifft die Nullhypothese zu.
Für eine unterschiedliche Zahl von Beobachtungen werden dann sowohl über Monte–Carlo–Simulation als auch über die o. a. Normalapproximation die Verteilungen von T_n ermittelt und graphisch mit Hilfe von Kern–Dichteschätzungen dargestellt ($\hat{f}(x)$ ist dabei der übliche Kernschätzer (z. B. B. W. Silverman (1986)). Für die Wahl des Glättungsparameters h soll auf die Verwendung eines optimierenden Verfahrens verzichtet werden. Vielmehr soll die in vielen Arbeiten (z. B. M. Bonneu, M. Delecroix, E. Malin (1993)) anzutreffende Vereinfachung $h = \frac{\sigma}{n^{0.2}}$ auch hier zur Anwendung kommen. Insbesondere wird die in der Literatur zur Dichteschätzung und zur nichtparametrischen Regression enthaltene

Bedingung für optimale Konvergenz ($h \sim n^{-0.2}$) eingehalten. Als Kernfunktion soll der Quartic–Kern Verwendung finden.

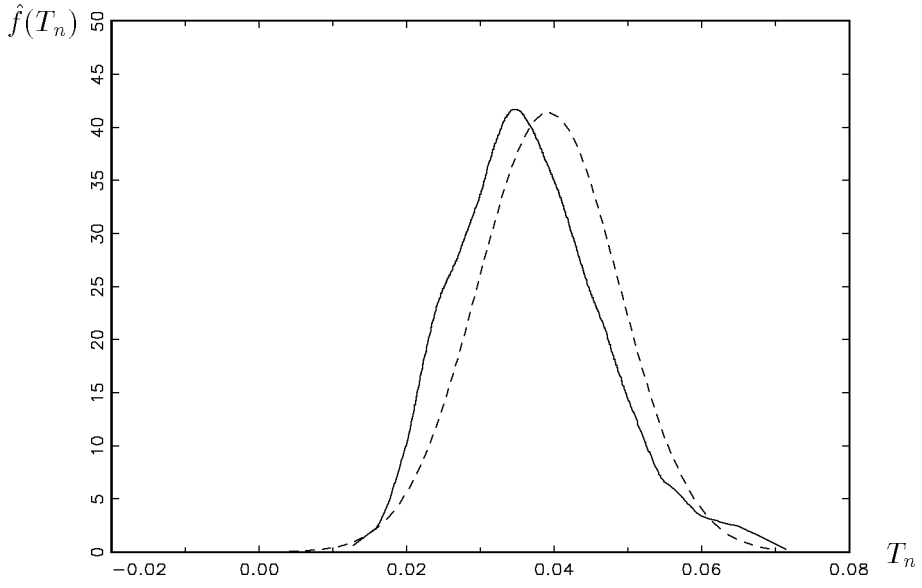
Leider läßt sich feststellen, daß die Konvergenz nur sehr langsam stattfindet. Dies gilt insbesondere für den Erwartungswert der beiden Verteilungen. Die Varianz hingegen ist bereits bei kleinen Stichprobengrößen etwa gleich. Für den Fall $n = 200$ ist die Normalapproximation zudem deshalb noch nicht zu gebrauchen, da nicht unerhebliche Masse auf dem negativen Teil der Achse liegt, dieses aber offensichtlich nicht zutreffen kann. Wie

Abbildung 1: *Vergleich der Verteilung der Teststatistik für $n = 200$. Die MC–Simulation ist durch die durchgezogene Linie und die Normalapproximation durch die gestrichelte Linie gekennzeichnet.*



oben gezeigt, wächst T_n mit steigendem n. Dies erkennt man ebenfalls an der asymptotischen Verteilung, denn b_n ist offensichtlich umgekehrt proportional zu dem verwendeten h, welches bei der oben erwähnten Regel mit wachsendem n sinken sollte. Daher sinkt der Anteil der Masse, der bei der Normalapproximation im negativen Bereich liegt mit wachsendem n. Bei $n = 50000$ ist die Approximation schon recht gut, auch wenn der Erwartungswert der beiden Verteilungen offensichtlich noch recht stark divergiert.

Abbildung 2: *Vergleich der Verteilung der Teststatistik für $n = 50000$. Die Kennzeichnung der Kurven erfolgt analog zu Abbildung 1.*



4 Approximation durch Bootstrap

Wie der vorige Abschnitt zeigt, ist insbesondere für kleine n die Normalverteilung nicht als Prüfverteilung heranzuziehen. Dieses gilt umso mehr, als bei der Normalapproximation im konkreten Anwendungsfall mit $\sigma(x)$ und $f(x)$ zwei unbekannte Terme auftreten, die gesondert geschätzt werden müssen und im Zweifel die Qualität der abgeleiteten kritischen Werte sinken läßt. Es zeigt sich im folgenden, daß über den Bootstrap verläßlichere Aussagen über kritische Werte zu treffen sind. Dabei stellt sich allerdings die Frage, welche Form des Resampling vorzuziehen ist. In der Literatur sind im wesentlichen zwei Formen bekannt (R. Cao-Abad und W. Gonzalez–Manteiga, 1993):

- Resampling aus den Wertepaaren (x, y)
- Resampling aus den Residuen.

Dabei wird in beiden Fällen das Resampling über alle Beobachtungen durchgeführt. Wie in den Simulationen gleich gezeigt wird, geht dadurch die bedingte Verteilung von Y für $X = x$ allerdings verloren. Formal ausgedrückt heißt das, daß die Regressionsfunktion

eben gerade nicht der bedingte Erwartungswert der Beobachtungen unter der resampleten Verteilung ist:

$$E^*\left(Y_i^* - m_{\hat{\theta}}(X_i^*)|X_i^*\right) = Y_i^* - m_{\hat{\theta}}(X_i^*) \neq 0$$

Vielmehr wird mit Hilfe dieses "naiven" Bootstraps die Verteilung (X_i, Y_i) reproduziert. Wu (1986) hat bereits gezeigt, daß im Falle des linearen Modells mit variabler bedingter Varianzfunktion der "naive" Bootstrap–Schätzer zu inkonsistenten Schätzungen führt. Er schlägt daher das "wild–bootstrap" vor (vgl. Härdle, 1990), welches eben folgendes Ziel verfolgt:

$$E^*(Y_i^* - m_{\hat{\theta}}(X_i^*)|X_i^*) = 0$$

Dazu wird für jedes X_i ein Residuum $\hat{\varepsilon}_i = Y_i - \hat{m}_h(X_i)$ gebildet und ausschließlich auf der Basis dieses einen Wertes die Berechnung des resampleten Y_i^* durchgeführt. Es wird nun jeweils eine Zwei–Punkt–Verteilung $\hat{F}_i$ gebildet, die folgenden Bedingungen genügen muß:

$$\begin{aligned} E(Z) &= 0 \\ E(Z^2) &= (\hat{\varepsilon_i})^2 \\ E(Z^3) &= (\hat{\varepsilon_i})^3 \end{aligned}$$

Die Parameter der sich aus diesen Bedingungen ergebenden Zwei–Punkt–Verteilung sind z. B. (wobei a und b die beiden Massepunkte sind und $P(X = a) = \gamma$ gelten soll)

$$\begin{aligned} a &= \hat{\varepsilon}_i \left(1 - \sqrt{5}\right)/2 \\ b &= \hat{\varepsilon}_i \left(1 + \sqrt{5}\right)/2 \\ \gamma &= \left(5 + \sqrt{5}\right)/2 \end{aligned}$$

Aus diesen Verteilungen werden dann die ε_i^* resamplt und die Paare $(X_i, Y_i^* = m_{\hat{\theta}}(X_i) + \varepsilon_i^*)$ berechnet.

Als Unterstützung für diese eher heuristische Erklärung sollen folgende Beispiele dienen. Analog zu dem vorigen Beispiel werden nun für $n = 200$ jeweils Polynome ersten bis vierten Grades unterstellt, in allen Fällen trifft somit die Nullhypothese zu. Für die Wahl des Glättungsparameters h und der Kernfunktion gelten die gleichen Angaben wie bzgl. des ersten Simulationsbeispiels. Die Anzahl der Bootstrap–Wiederholungen betrage $B = 200$. Die MC–Verteilung ist durchgezogen, die Verteilung über wild–bootstrap gestrichelt und die Verteilung über naiven Bootstrap gepunktet dargestellt.

Abbildung 3: *Vergleich der Bootstrap–Performance für $n = 200$, wobei ein Polynom ersten Grades unterstellt wird.*

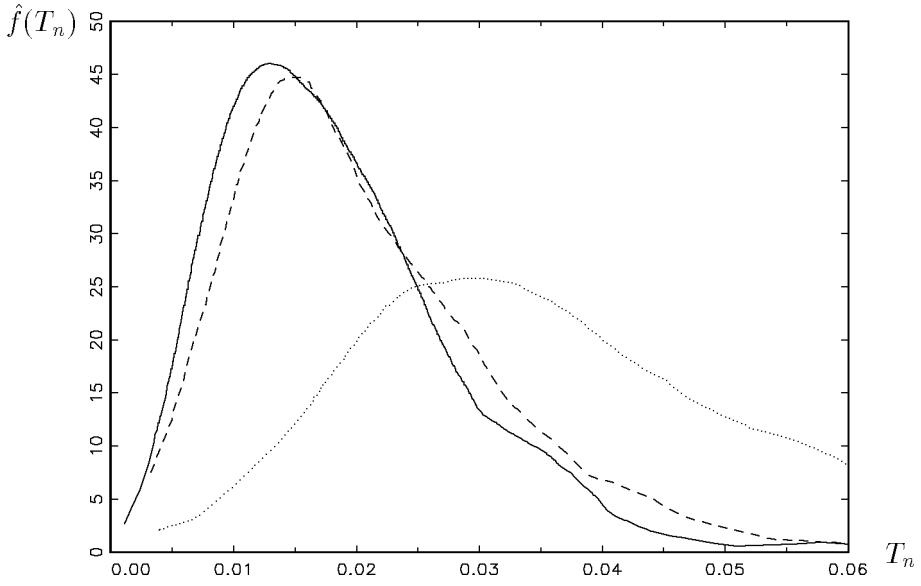


Abbildung 4: *Vergleich der Bootstrap–Performance für $n = 200$, wobei ein Polynom zweiten Grades unterstellt wird.*

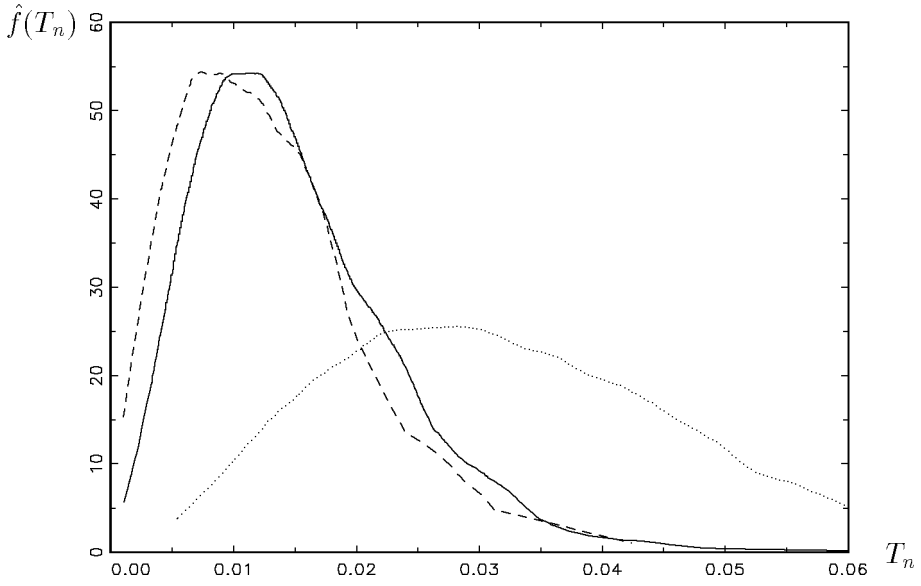


Abbildung 5: *Vergleich der Bootstrap–Performance für $n = 200$, wobei ein Polynom dritten Grades unterstellt wird.*

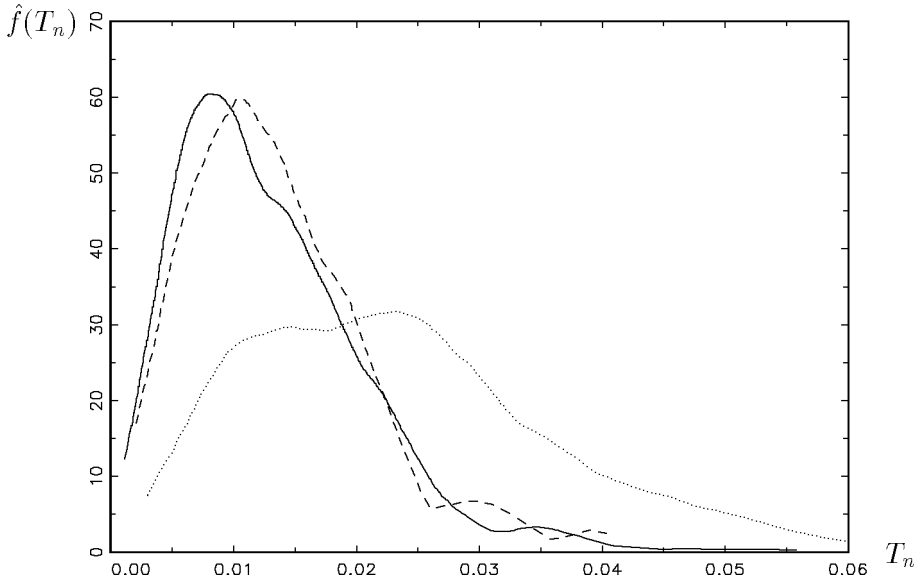
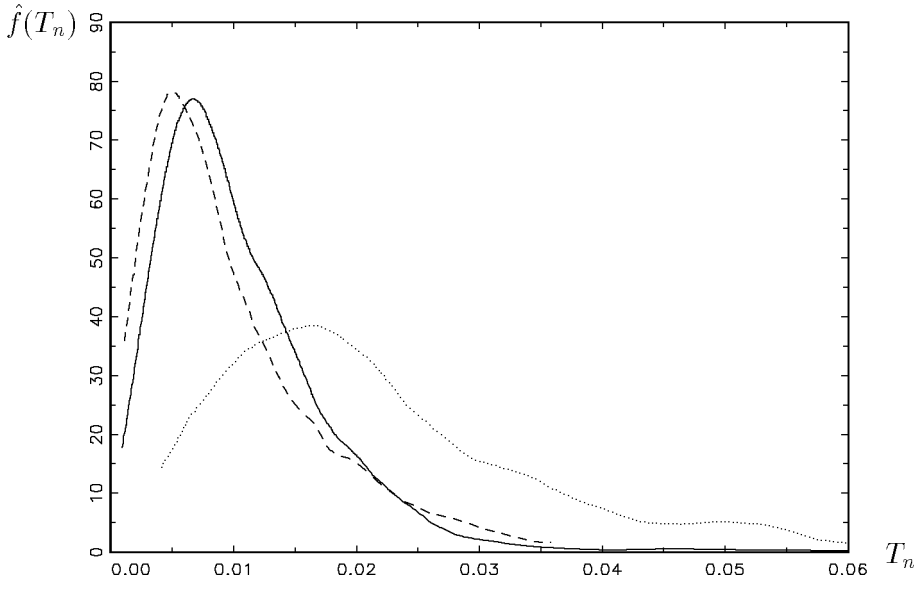


Abbildung 6: *Vergleich der Bootstrap–Performance für $n = 200$, wobei ein Polynom vierten Grades unterstellt wird.*



5 Monte–Carlo–Simulation der Gütefunktion

Um einen Eindruck von der Gütefunktion des vorgestellten Testfunktion zu erhalten, werden im folgenden für einzelne Punkte die Werte der Gütefunktion durch Monte–Carlo–Simulation ermittelt. Dazu soll folgendes Beispiel herangezogen werden:

$$m(x) = 2x - x^2 + c\left(x - \frac{1}{4}\right)\left(x - \frac{1}{2}\right)\left(x - \frac{3}{4}\right)$$

Das unterstellte parametrische Modell ist dabei ein Polynom zweiter Ordnung, so daß für $c = 0$ die Nullhypothese zutrifft. Mit steigendem c sollte daher die Ablehnungswahrscheinlichkeit zunehmen. Für die übrigen Komponenten gelten die gleichen Angaben wie in den bisherigen Simulationen. Bei $N = 1000$ Wiederholungen ergaben sich für verschiedene c, h–Kombinationen die in Tabelle 1 aufgelisteten Ablehnungswahrscheinlichkeiten für $\alpha = 0,05$. Offensichtlich hat die Wahl des jeweiligen Glättungsparameters einen er-

Tabelle 1: Ablehnungswahrscheinlichkeiten bei $N = 1000$ Simulationen

		c			
		0,0	0,5	1,0	2,0
h	0,1	0,088	0,104	0,278	0,876
	0,2	0,052	0,121	0,287	0,872
	0,25	0,038	0,128	0,230	0,795
	0,3	0,017	0,091	0,298	0,687

heblichen Einfluß auf die Ablehnungswahrscheinlichkeit. Ein mit Hilfe der Kreuzvalidierungsfunktion ermitteltes optimales h ergab sich mit $\hat{h} = 0,21$. Dementsprechend scheint die zweite Zeile die relevanten Ergebnisse zu beinhalten. Dies gilt umso mehr, als die dort ermittelte Ablehnungswahrscheinlichkeit für $c = 0$ nahe bei 0,05 liegt, was ja bei der Simulation vorausgesetzt wurde. Je größer h ist, desto geringer wird dabei wohl die Ablehnungswahrscheinlichkeit, was bei der allgemeinen Interpretation dieses Glättungsparameters nicht verwundert.

Trotz dieser allgemeinen Aussagen wäre im weiteren zu untersuchen, ob es im Rahmen dieses Tests speziellere Auswahlkriterien für den Glättungsparameter gibt. Nach Kenntnis des Autors liegen dort noch keine Ergebnisse vor.

6 Das Wachstum des Internet

Das Internet ist ein weltweites Computernetz, das ursprünglich für militärische Zwecke installiert wurde, sich aber, nachdem es 1972 der Öffentlichkeit zugänglich gemacht wurde, schnell zu einem Computerverbund amerikanischer Universitäten entwickelte und inzwischen Hochschulen und andere Institutionen in der ganzen Welt verbindet. Dabei nimmt der Anteil der akademischen Einrichtungen an diesem Kommunikationsnetz ständig ab. Im Januar 1995 lag der Anteil erstmals unterhalb dem Anteil kommerzieller Institutionen. Das Internet ist damit zu dem Kernstück der in naher Zukunft zu erwartenden weltumspannenden Datenautobahn geworden, die dann allerdings nicht nur electronic mails oder andere akademische Inhalte, sondern vielmehr kommerzielle Produkte wie Fernsehen etc. transportieren wird.

Damit ist die Verbreitung dieses Netzes ein geradezu ideales Anwendungsfeld für die klassischen endogenen Wachstumsmodelle. Einen Überblick über diese Modelle, die in der betriebswirtschaftlichen Marktforschung auch weiterhin eine große Rolle spielen, findet sich bei Mertens und Falk (1994).

Als konkretes Merkmal für die Verbreitung des Internet sind vier Merkmale herangezogen worden:

1) Anzahl der mit dem Netz verbundenen Rechner (Großrechner, Workstations, PC's u.a.). Diese erhalten im Netz eine eigene Identifikationsnummer, auch IP–Nummer genannt.

2) Anzahl der mit dem Netz verbundenen Sub-Netze. Das sind Netze, deren Rechner in der Systematik der IP–Nummern zu sogenannten Domains zusammengefaßt werden. Dabei handelt es sich z.B. um lokale Netze der einzelnen Universitäten.

3) Anzahl der auf dem Netz verschickten Datenpakete. Die auf dem Netz verschickten Informationen werden in der Regel in einzelne Pakete zerlegt und dann auf verschiedenen Wegen zum Adressaten geleitet und dort wieder zusammengefügt.

4) Anzahl der über das Internet verschickten Bytes. Dies ist die niedrigste Aggregationsstufe der versendeten Informationen (dabei wird die Anzahl der Bytes meist – wie auch hier – direkt in der Einheit Millionen gemessen).

Tabelle 2: Ausbreitung des Internet

Merkmal	Zeitraum	Anzahl Daten
Rechner	08/81–10/94	28
Sub–Netze	07/88–01/95	79
Pakete	01/88–10/94	82
Bytes	03/91–11/94	45

Tabelle 3: Endogene Wachstumsmodelle

Linear	$Y(t) = a + bt$
Exponentiell	$Y(t) = \exp\{a + bt\}$
Mod. exponentiell	$Y(t) = Y^* - \exp\{-a - bt\}$
Logistisch	$Y(t) = \frac{Y^*}{1+\exp\{a-bt\}}$
Gompertz	$Y(t) = Y^* \exp\{-bc^t\}$
Weblus	$Y(t) = \frac{Y^*}{1+\left(\frac{b}{t}\right)^c}$

Grundsätzlich ist es bei der Komplexität dieses Systems schwierig, an genaue Werte für diese Merkmale zu gelangen. Das Internet verfügt jedoch über einen Hauptleitungsstrang, das "NSF–Backbone". Der Betreiber dieses Backbones veröffentlicht Zahlen über die bei ihm angemeldeten Rechner und Subnetze sowie über die Nutzung seines Backbones und der Auslastung der Subnetze. Diese Zahlen sind jederzeit über die verschiedenen Dienste des Internet abrufbar. Die Verfügbarkeit der Daten ist in Tabelle 2 aufgeführt.

Mithilfe der KQ–Methode werden nun die in Tabelle 3 aufgeführten Wachstumsmodelle angepaßt. In der Folge wird der vorgestellte Test durchgeführt. Dazu wird zum einen auf der Basis der resampleten Werte (wobei 200 Bootstrap–Wiederholungen erfolgt sind) die Testgröße T_n ermittelt. Der kritische Wert ergibt sich dann als das 95%–Quantil der Verteilung der T_n, mithin wurde also in der Reihe der aufsteigend sortierten Werte der T_n der 190. Wert gewählt. Dann erfolgt auf der Basis der Ursprungswerte die Berechnung der Testgröße, welche dann als Prüfgröße mit dem kritischen Wert verglichen wird. Der vorgestellte Test führt dann zu den in Tabelle 4 bis Tabelle 7 aufgeführten Resultaten, wobei bei der Größenordnung der Werte zu beachten ist, daß das Niveau der Merkmalswerte offenbar großen Einfluß auf die Testgröße hat.

Tabelle 4: Anzahl der Pakete

Modell	krit. Wert	Prüfwert
Linear	1,0073E06	7,8087E08
Exponentiell	1,2011E06	8,5197E06
Mod. exponentiell	2,5496E06	8,1578E08
Logistisch	1,0434E06	4,2406E06
Gompertz	1,5452E06	3,6909E06
Weblus	1,2515E06	1,1657E07

Tabelle 5: Anzahl der Sub–Netze

Modell	krit. Wert	Prüfwert
Linear	57306	4,2763E08
Exponentiell	64839	7,9601E06
Mod. exponentiell	5,4935E05	4,4389E08
Logistisch	60059	7,5711E06
Gompertz	2,2229E05	1,1185E07
Weblus	1,5072E05	1,3246E07

Die Ergebnisse zeigen, daß der Test bei den beiden Merkmalen Sub–Netze und Daten–Pakete für alle sechs untersuchte Wachstumsmodelle zu einer Ablehnung der Nullhypothese führt. Die Ablehnung ist allerdings bei verschiedenen Modellen (insbesondere einfaches Exponential– und logistisches Modell) durchaus knapp. Bei einem Konfidenzniveau von 0.99 hätte sich in beiden Fällen die Nullhypothese nicht ablehnen lassen.

Bei den beiden Merkmalen Rechner und Bytes werden die einfache Exponentialfunktion sowie die logistische Funktion nicht abgelehnt (bei $1 - \alpha = 0.95$). Für die übrigen Modelle kommt es allerdings wieder zu einer Ablehnung der Nullhypothese.

In den beiden Abbildungen 7 und 8 wird die Entwicklung der Merkmale Daten–Pakete und Sub–Netze zusammen mit der nichtparametrischen (gestrichelt) und der exponentiellen (durchgezogen) Anpassung dargestellt. Besonders im Falle der Sub–Netze erkennt man, wie sehr der tatsächliche Verlauf von dem unterstellten exponentiellen Verlauf abweicht.

Abbildung 7: *Entwicklung der Anzahl der Daten–Pakete (× 10.000)*

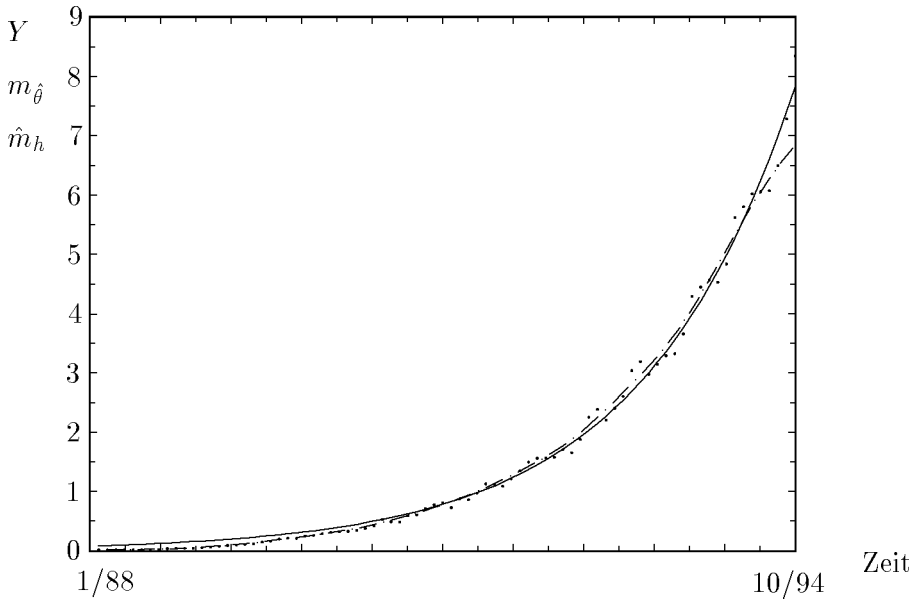


Abbildung 8: *Entwicklung der Anzahl der Sub–Netze (× 10.000)*

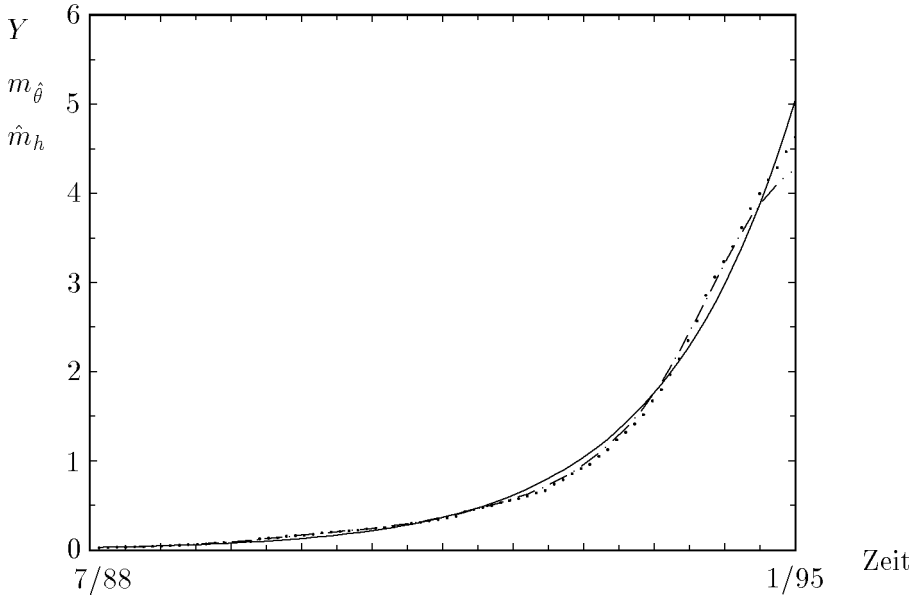


Tabelle 6: Anzahl der Rechner

Modell	krit. Wert	Prüfwert
Linear	3,7855E09	9,4858E11
Exponentiell	3,1417E9	2,3409E09
Mod. exponentiell	4,0984E09	9,5383E11
Logistisch	2,5781E09	2,3818E09
Gompertz	5,3831E09	8,3095E09
Weblus	3,1961E09	4,6321E09

Tabelle 7: Anzahl der Bytes

Modell	krit. Wert	Prüfwert
Linear	3,3238E05	1,2435E07
Exponentiell	3,1693E05	3,0339E05
Mod. exponentiell	3,5918E05	1,3529E07
Logistisch	3,5214E05	3,3280E05
Gompertz	3,0194E05	9,1963E05
Weblus	2,7997E05	5,0994E06

In den beiden Abbildungen 9 und 10 sind die logistischen Anpassungen (durchgezogen) und die nichtparametrischen Schätzungen (gestrichelt) für die Anzahl der Rechner und für die Anzahl der Bytes dargestellt. Die sich bei der logistischen Anpassung ergebenden Sättigungsgrenzen ergeben für die Anzahl der Rechner einen Wert von 3,4426E07 und für die Anzahl der Bytes einen Wert von 2,1925E05.

Die unterschiedlichen Ergebnisse haben im wesentlichen zwei Ursachen:

1) Bei den beiden Merkmalen Anzahl der Pakete und Anzahl der Sub–Netze handelt es sich um aggregierte Größen, deren Zusammensetzungen nicht eindeutig definiert sind. Weder die Größe der Daten–Pakete, die alleine von der Belastung des Netzes und dem Umfang der versendeten Informationen abhängt, noch die der Sub–Netze, die von den lokalen System–Administratoren bestimmt wird, sind jeweils gleich. Insofern kann es sehr leicht zu Sprüngen in der Entwicklung kommen, die die einfachen

Abbildung 9: *Entwicklung der Anzahl der Rechner (× 1.000.000)*

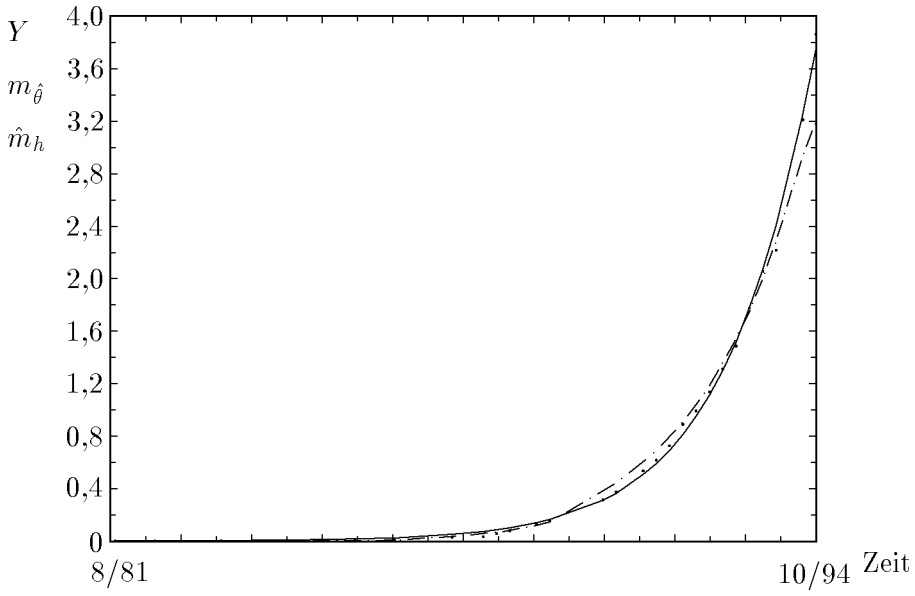
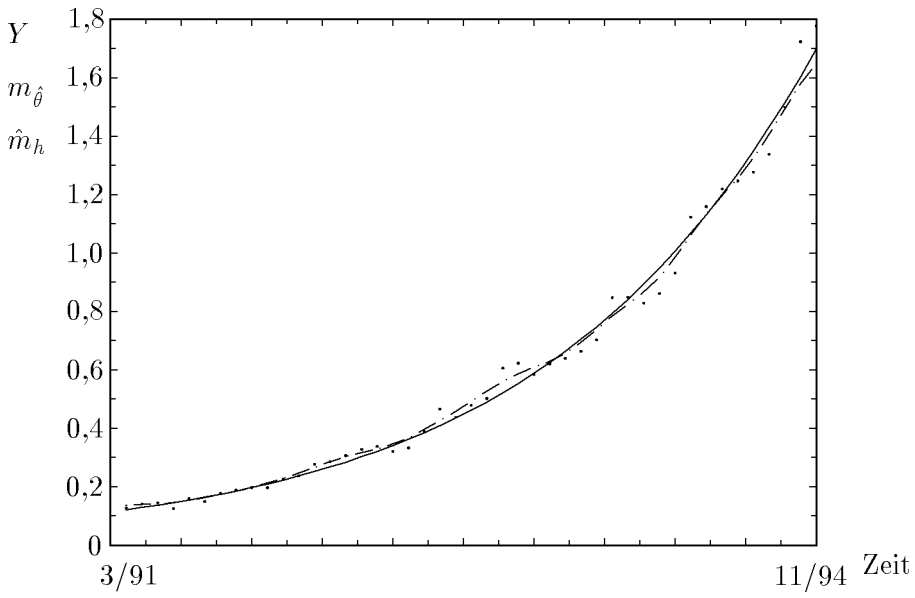


Abbildung 10: *Entwicklung der Anzahl der Bytes (× 10.000)*



Wachstumsfunktionen nicht zu beschreiben vermögen. Die beiden anderen Merkmale hingegen weisen das jeweils niedrigste Aggregationsniveau auf. Ihre Entwicklung ist damit viel gleichmäßiger und kann daher eher von solch einfachen Modellen beschrieben werden.

2) Mit dem jeweiligen Aggregationsniveau verbunden ist allerdings auch das Problem der Datenerfassung. Es ist einfacher, Daten auf einem hohen als auf einem niedrigen Aggregationsniveau zu erfassen. Dies führt zu den unterschiedlichen Zeitreihenlängen. Dabei erkennt man, daß gerade die Entwicklung der Merkmale mit den längeren Zeitreihen nicht von den Modellen beschrieben werden kann, daß also der vorgestellte Test zu einer Ablehnung der Nullhypothese führt. Umgekehrt läßt sich damit argumentieren, daß in den beiden Fällen, in denen die Nullhypothese nicht abgelehnt werden konnte, das Datenmaterial möglicherweise für eine Ablehnung nicht ausreichend gewesen ist. Dem steht allerdings entgegen, daß der Test durchaus zu einer Differenzierung zwischen den Modellen fähig war.

7 Schlußbemerkung

Die hier verwendeten Modelle führten in den beschriebenen Fällen zu einer Ablehnung der Nullhypothese. Dies sollte dazu Anlaß geben, komplexere parametrische Wachstumsmodelle zu verwenden. Diese Komplexität kann entweder durch die Verwendung zusätzlicher Parameter oder aber durch die Berücksichtigung auch anderer exogener Variablen erfolgen. Der hier verwendete Test sollte auch für diesen Fall anwendbar sein, wobei dann auf der nichtparametrischen Seite allerdings additive Modelle zur Anwendung kämen, da die nichtparametrische Schätzung in höheren Dimensionen unter der "curse of dimensionality" leidet (Härdle, 1990). Dies und das Problem der Wahl des Glättungsparameters h speziell für diesen Fall sollten Gegenstand zukünftiger Forschung sein.

Literaturverzeichnis

Bonneau, M., M. Delecroix und E. Malin (1993): Semiparametric versus Nonparametric Estimation in Single Index Regression Model: A Computational Approach. Computational Statistics 8, S. 207–222

Cao-Abad, R. und W. Gonzalez–Manteiga (1993): Bootstrap Methods in Regression Smoothing. Nonparametric Statistics 2, S. 379–388

Härdle, W. (1990): Applied Nonparametric Regression.

Härdle, W., J. S. Marron (1990): Semiparametric Comparison of Regression Curves. Annals of Statistics 18, S. 63–89

Mammen, E. (1992): When does Bootstrap work? Asymptotic Results and Simulations. Lecture Notes in Statistics 77.

Mertens, P., J. Falk (1994): Mittel– und langfristige Absatzprognose auf der Basis von Sättigungsmodellen, in: P. Mertens (Hrsg.): Prognoserechnung.

Nadaraya, E. A. (1964): On Estimating Regression. Theory of Probability and its Applications 10, S. 186-190

Silverman, B. W. (1986): Density Estimation for Statistics and Data Analysis.