GKSS
FORSCHUNGSZENTRUM
in der HELMHOLTZ-GEMEINSCHAFT

Ein Bewertungsmaß für die amplitudentreue regressive Abbildung von verrauschten Daten im Rahmen einer iterativen „Errors in Variables"-Modellierung (EVM)

Autoren:

B. Thees
L. Kutzbach
M. Wilmking
E. Zorita

ISSN 0344-9629

HELMHOLTZ
| GEMEINSCHAFT

wissen
schafft
nutzen

GKSS 2009/8

GKSS 2009/8

Ein Bewertungsmaß für die amplitudentreue regressive Abbildung von verrauschten Daten im Rahmen einer iterativen „Errors in Variables“-Modellierung (EVM)

Autoren:

B. Thees

(Ernst-Moritz-Arndt-Universität Greifswald)

L. Kutzbach

(Universität Hamburg)

M. Wilmking

(Ernst-Moritz-Arndt-Universität Greifswald)

E. Zorita

(GKSS, Institut für Küstenforschung)

GKSS-Forschungszentrum Geesthacht GmbH · Geesthacht · 2009

Die Berichte der GKSS werden kostenlos abgegeben.
The delivery of the GKSS reports is free of charge.

Anforderungen/Requests:

GKSS-Forschungszentrum Geesthacht GmbH
Bibliothek/Library
Postfach 11 60
21494 Geesthacht
Germany
Fax.: +49 4152 87-17 17

Als Manuskript vervielfältigt.

ISSN 0344-9629

GKSS-Forschungszentrum Geesthacht GmbH · Telefon (04152) 87-0
Max-Planck-Straße 1 · 21502 Geesthacht / Postfach 11 60 · 21494 Geesthacht

GKSS 2009/8

Ein Bewertungsmaß für die amplitudentreue regressive Abbildung von verrauschten Daten im Rahmen einer iterativen „Errors in Variables“- Modellierung (EVM)

Barnim Thees, Lars Kutzbach, Martin Wilmking, Eduardo Zorita

24 Seiten mit 8 Abbildungen und 6 Tabellen

Zusammenfassung

Es wird ein Bewertungsmaß für die Amplitudentreue einer regressiven Abbildung vorgestellt. Als Ergänzung zu den klassischen Gütemaßen einer Modellierung, wie z.B. dem „Root Mean Square Error“ (RMSE), könnte es benutzt werden, um ein „bestes“ regressives Modell im Rahmen einer „Errors in Variables“-Modellierung (EVM) zu schätzen: Ein Modellergebnis eines mit einem unbekannten Fehlerrauschen überlagerten Prozesses mit einem minimalen Fehler im Punkt (RMSE) und einer gleichzeitig maximalen Amplitudentreue der Periodizitäten. Anhand von Testrechnungen mit einem mittelwertfreien, unabhängigen und weißen Fehlerrauschen überlagerten sinusförmigen Prozesses und einer derart verrauschten Klimamodellrechnung (Pseudoproxy) wird gezeigt, dass dieses Maß im Rahmen einer iterativen EVM-Modellierung näherungsweise zum Auffinden eines unbekannten Rauschverhältnisses des Fehlerrauschens benutzt werden kann, da es bei den bisher durchgeführten Testrechnungen ein Maximum in dessen Nähe besitzt. Dies wäre von herausragender Bedeutung, da eine Anwendung der allgemeingültigeren EVM-Modellierung häufig an der Unkenntnis des Fehlerrauschens in den Daten scheitert.

A metric to assess variance conservation by 'Error in Variables' regression models

Abstract

A new metric to assess the skill of regression methods to conserve the true amplitude of the predictand variance in regression analysis is defined. Additionally to the classical measure of skill, e.g. the root-mean-square-error, this new metric can be used to identify a 'best' regression model within the family of error-in-variables (EIV) models. This optimal model represents a process with superimposed noise with unknown amplitude and a simultaneous conservation of the amplitude of the signal. Tests based on pseudo-proxies derived by adding to the predictor a disturbed sinus wave or with pseudo-proxies derived from a climate model simulation show that this metric can be used to estimate the unknown amplitude of the noise within the framework of an EIV model. This can be of relevance since, generally, the successful application of EIV models is hindered by the lack of knowledge of the true amplitude of the noise.

Manuscript received / Manuskripteingang in TFP: 1. Oktober 2009

Inhalt

1 Einleitung

Bei der Beobachtung eines (unbekannten) Prozesses (y_i,x_i) in der natürlichen Umwelt kann im Gegensatz zur „Klimakammerphysik“ nicht davon ausgegangen werden, dass die i=1,...,N Beobachtungsergebnisse $(y_i^{'},x_i^{'})$ dann „fehlerfrei in den Variablen“ y, x vorliegen. Neben den unvermeidbaren zufälligen Fehlern von etwaigen Messgeräten existieren häufig Fehler in der Beobachtung, die dadurch entstehen können, wenn z.B. nicht vorhandene Temperaturwerte (y) eines Baumstandortes durch die Temperaturmessungen einer Klimastation ersetzt werden. Daneben haben natürlich eine unbekannte Anzahl von weiteren Prozessen einen möglichen Einfluss auf die Beobachtungsergebnisse, so spiegelt sich beispielsweise neben dem Einfluss der Temperatur auch der Niederschlag, die Versorgung mit Nährstoffen usw. in den Baumringdicken (x) wieder. Die Beobachtungsergebnisse des Prozesses zeigen somit in der Darstellung: $y_i^{'}=f(x_i^{'})$ eine durch dieses Fehler- und Prozessrauschen verstärkt streuende Punktwolke.
Hierbei wird angenommen, dass das Fehler- und Prozessrauschen: (δ_i,ε_i) den unverrauschten Prozess additiv überlagert: $y_i^{'}=y_i+\delta_i$ und $x_i^{'}=x_i+\varepsilon_i$

Mit Hilfe der Ausgleichsrechnung (Regressionsrechnung) kann der unbekannte Prozess, im einfachsten Fall als linearer Prozess:

$$y=c_0+c\cdot x \tag{1}$$

oder auch invers : $x=a_0+a\cdot y$ (1b)

aus den Messergebnissen dann geschätzt werden (z.B. in [2], [4]):

1.1 „Ordinary Least Squares“ Modellierung (OLS)

Im Rahmen einer „Ordinary Least Squares“(OLS)-Schätzung wird eine bestmögliche Gerade:

$$\hat{y}_i^{'}=\hat{c}_0+\hat{c}\cdot x_i^{'} \tag{2}$$

oder invers : $\hat{x}_i^{'}=\hat{a}_0+\hat{a}\cdot y_i^{'}$ (2b)

derart in die Punktwolke $(y_i^{'},x_i^{'})$ gelegt, dass die mittlere quadratische Abweichung von dieser Geraden minimiert wird (bei Differenzbildung in Richtung der abhängigen Größe):

$$\sum_{i=1}^{N}(y_i^{'}-\hat{y}_i^{'})^2=\sum_{i=1}^{N}(y_i^{'}-\hat{c}_0-\hat{c}\cdot x_i^{'})^2=f(\hat{c}_0,\hat{c})=MIN \tag{3}$$

$$\sum_{i=1}^{N}(x_i^{'}-\hat{x}_i^{'})^2=\sum_{i=1}^{N}(x_i^{'}-\hat{a}_0-\hat{a}\cdot y_i^{'})^2=f(\hat{a}_0,\hat{a})=MIN \tag{3b}$$

Die Bedingungen (3) oder (3b) führen dann bekanntlich zu den Schätzgleichungen für die Modellparameter (c_o,c) oder (a_o,a) :

$$\hat{c}=\frac{S_{x',y'}}{S^2_{x'}} \qquad (4) \qquad \hat{c}_0=\bar{y}'-\hat{c}\cdot\bar{x}' \qquad (5)$$

$$\hat{a}=\frac{S_{x',y'}}{S^2_{y'}} \qquad (4b) \qquad \hat{a}_0=\bar{x}'-\hat{a}\cdot\bar{y}' \qquad 5b)$$

wobei die Ausgleichsgeraden (2) und (2b) sich in: $(\bar{y}', \bar{x}')$ schneiden und das Produkt aus beiden Anstiegen gleich dem Korrelationskoeffizienten ist:

$$r^2=\hat{a}\cdot\hat{c} \qquad (6)$$

Das bedeutet dann, nur wenn $r^2=1$, d.h. ein deterministischer Zusammenhang besteht, ist $\hat{c}=1/\hat{a}$.

Da die Differenzbildung bei der Minimierung der Quadrate in (3) und (3b) in Richtung der jeweils abhängigen Größe erfolgt, kann die OLS- Ausgleichsrechnung nur unverzerrte Schätzungen der Modellparameter liefern, wenn die jeweils unabhängigen Variablen fehlerfrei vorliegen, d.h. für y = f(x) dann: $\varepsilon_i \equiv 0$ (!) , so dass dann gilt:

$$y_i'=\hat{y}_i'+res_i=\hat{c}_0+\hat{c}\cdot x_i+res_i \qquad (7)$$

Das Regressionsmodell (2) ist nur dann eine beste Schätzung für (1) solange die Residuen: res_i eine Gauß-Normalverteilung (GNV) befolgen.

Diese Residuen können dabei aufgefasst werden als:

Messfehler von y_i , wenn der lineare Prozess (1) wirklich (im deterministischen Sinne) existiert, z.B.: bei einer („Klimakammer“-)Kalibrierung, unter der Annahme von Fehlerfreiheit der x_i .

oder als:

Modellfehler von $\hat{y}_i$, wenn ein unbekannter Prozess in der natürlichen Umwelt abgebildet werden soll. In diesem Falle existieren aber immer auch Beobachtungsfehler (s.o.).

1.2 „Errors in Variables“-Modellierung (EVM)

Bei den Ausgleichsrechnungen im Rahmen einer „Errors in Variables“-Modellierung (EVM) werden nun, sozusagen verallgemeinernd, Fehler in beiden Variablen (y, x) zugelassen. Die Minimierung der mittleren quadratischen Abweichungen, jetzt auch als „Total Least Squares“ bezeichnet, führen dann zu unverzerrten Schätzungen $(\hat{c}_{k0}, \hat{c}_k)$, falls über das Rauschen $(\delta_i, \varepsilon_i)$ angenommen werden kann, dass es mittelwertfrei ist :

$$\bar{\delta}_i, \bar{\varepsilon}_i \equiv 0 \qquad (8a)$$

unabhängig untereinander und zu den Variablen, d.h. entsprechend kovarianzfrei ist:

$$S_{\delta\varepsilon}, S_{\delta y,}, S_{\delta x}, S_{\varepsilon y}, S_{\varepsilon x} \equiv 0 \qquad \text{(für alle i (hier weggelassen)} \qquad (8b)$$

und außerdem weiß, d.h. ohne Erhaltungsneigung ist:

$$S_{\delta i \delta j}, S_{\varepsilon i \varepsilon j} \equiv 0 \qquad \text{(für: } i \neq j \text{)} \tag{8c}$$

Mit diesen Bedingungen können dann z.B. die folgenden Schätzgleichungen für (c_o, c) (z.B.: in [2]) erhalten werden:

Für den Anstieg c die Gleichung

$$\hat{c}_k = \frac{S^2_{y'} - \lambda \cdot S^2_{x'}}{2 \cdot S_{x'y'}} + \frac{\sqrt{((N-1)\cdot S^2_{y'} - \lambda \cdot (N-1) \cdot S^2_{x'})^2 + 4 \cdot \lambda \cdot ((N-1) \cdot S_{x'y'})^2}}{2 \cdot (N-1) \cdot S_{x'y'}} \tag{9}$$

und für c0 dann entsprechend:

$$\hat{c}_{0k} = \bar{y}' - \hat{c}_k \cdot \bar{x}' \tag{10}$$

Bei Kenntnis des Rauschverhältnisses (RV) des Fehlerrauschens:

$$\lambda = S^2_\delta / S^2_\varepsilon \tag{11}$$

kann somit mit (9) und (10) eine unverzerrte Schätzung der Modellparameter erhalten werden.

Anhand eines vorgegebenen künstlichen und dann verrauschten Prozesses sollen nun einige Probleme bei dessen regressiver Modellierung näher beschrieben und mögliche Lösungsansätze aufgezeigt werden:

2 Der Effekt der Amplitudenreduktion

Die folgende lineare Abbildung einer Sinusschwingung :

$y_i = c_0 + c \cdot x_i$ mit: $y = 2.1 \cdot \sin$ und: $x = \sin$; i=1...N=120 für eine Sinusschwingung

bei der die Variablen y und x mit einem weißen, mittelwertfreien und unabhängigen Zufallsrauschen überlagert sind: $y'_i = y_i + \delta_i$; $x'_i = x_i + \varepsilon_i$
kann dann nur im Rahmen einer Errors-in-Variables-Modellierung (EVM)„amplitudentreu“ ($\hat{c}_k = 2.1$) abgebildet werden, sofern die Streuungen des Rauschens (bzw. das Rauschverhältnis RV= λ) bekannt sind und das Rauschen mittelwertfrei, weiß, unabhängig untereinander und zu den Variablen ist.

Dagegen erfolgt bei derart verrauschten x- und y-Daten im Rahmen einer Ordinary Least Squares (OLS) Modellierung grundsätzlich immer eine Unterschätzung der Amplituden (z.B. in [1],[3],[4]) :

$$abs(\hat{c}) < abs(\hat{c}_k) = abs(c) < abs(1/\hat{a}) \tag{12}$$

Bei einem Fehlerrauschen mit den obigen Eigenschaften ändert sich nämlich in (4) die Kovarianz nicht: $S_{x'y'}=S_{xy}$, wobei sich aber die Streuung dann natürlich erhöhen muss: $S^2_{x'}>S^2_x$, so dass das Verhältnis aus beiden kleiner wird und sich somit eine betragsmäßig unterschätzte Steigung der OLS-Ausgleichsgeraden ergibt.

Bei der korrekten EVM- Schätzung bewegt sich dann die geschätzte Steigung (9) zwischen dem Wert für die OLS-Schätzung (4) (für: $\lambda\rightarrow\infty$) als Untergrenze und dem Wert für den Kehrwert der inversen OLS-Schätzung (4b) (für $\lambda\rightarrow 0$) als Obergrenze.

In diesem Zusammenhang existiert aber nun das folgende Problem:

Die rauschkorrigierte regressive Abbildung (EVM) wird gegenüber der unkorrigierten (OLS) im Punkt „schlechter“, d.h. der Standardfehler wird größer, dafür werden aber die Amplituden möglicher (zeitlicher) Schwankungen in der Datenreihe besser abgebildet (siehe dazu auch der RMSE des Beispiels in Tabelle 2).

Die „klassischen“ Bewertungsmaße (Standardfehler, RMSE,...) für die Güte einer Modellierung sind somit nicht brauchbar, wenn man gerade an einer möglichst amplituden-treuen Abbildung der Daten interessiert ist.

Die Daten müssen dann quasi „unscharf im Detail“ aber mit Augenmerk auf ihr „typisches“ (periodisches) Verhalten hin analysiert, verglichen werden. Die Modelldaten sollen dann möglichst alle Periodizitäten mit den „wahren“ Amplituden der Ausgangs(Mess-)Daten wiederspiegeln.

3 Direkter Vergleich der Ausgangs- und Modelldaten durch Vergleich der Amplituden „in Wellenzügen“ der Datenreihe

Hierzu wird eine Datenreihe mit N-Werten mit Hilfe der (zeitlichen) Abfolge ihrer lokalen Extrema in eine Abfolge von p-Wellenzügen zerlegt, wobei die Amplitude des jeweiligen Wellenzuges zwischen den jeweiligen Maxima und Minima direkt gebildet wird:

Definition lokales Extremum:

Lokales Maximum von y: $Max(y)\equiv y^o_j:\overrightarrow{Def}\ y_{i-1}<(y^o_j:=y_i)>y_{i+1}$
Lokales Minimum von y: $Min(y)\equiv y^u_j:\overrightarrow{Def}\ y_{i-1}>(y^u_j:=y_i)<y_{i+1}$

mit : i = 1 , ... , N - (Mess)Daten
und : j = 1 , ... , p - Wellenzügen

Die Schwingungsweite, d.h. die doppelte Amplitude des Wellenzuges, z.B. zwischen zwei Minima, ist dann:

$$y^A_j:=((y^o_j-y^u_j)+(y^o_j-y^u_{j+1}))/2$$

Dieser Wellenzug schwingt dann mit seiner Amplitude um den Wert:

$$y^m_j:=y^o_j-y^A_j/2$$

Alle (q) y_i-Werte *innerhalb* eines (jeden) der p-Wellenzüge werden nun den Parametern des jeweiligen Wellenzuges zugeordnet. Für das obige verrauschte y = 2.1* SIN-Beispiel ergibt sich dann z.B. das folgende Zuordnungsschema:

I	$y'(i)$	$y_j^o(i)$	$y_j^u(i)$	$y_j^A(i)$	$y_j^m(i)$	$f_j^o(i)$	$f_j^u(i)$
1	-0.31631	0.88067	-0.31631	0.71591	0.52272	0.88067	0.16476
2	0.28707	0.88067	-0.31631	0.71591	0.52272	0.88067	0.16476
3	0.31948	0.88067	-0.31631	0.71591	0.52272	0.88067	0.16476
4	0.88067	0.88067	-0.31631	0.71591	0.52272	0.88067	0.16476
5	0.64584	0.66142	0.64584	0.05490	0.63397	0.66142	0.60652
6	0.66142	0.66142	0.64584	0.05490	0.63397	0.66142	0.60652
7	0.56720	1.10188	0.56720	0.32355	0.94011	1.10188	0.77833
8	1.10188	1.10188	0.56720	0.32355	0.94011	1.10188	0.77833
9	1.07068	1.10188	0.56720	0.32355	0.94011	1.10188	0.77833
10	0.98946	1.51168	0.98946	0.48221	1.27058	1.51168	1.02947
11	1.51168	1.51168	0.98946	0.48221	1.27058	1.51168	1.02947
12	1.06948	1.83281	1.06948	0.56766	1.54898	1.83281	1.26515
13	1.18476	1.83281	1.06948	0.56766	1.54898	1.83281	1.26515
14	1.83281	1.83281	1.06948	0.56766	1.54898	1.83281	1.26515
15	1.46083	…	…	…	…	…	…
16	1.60430						
…							

Tabelle 1: Zuordnungsschema Bildung Amplitudenflächen

Mit: y' *- verrauschte Sinuswerte:* $y = 2.1 \cdot \sin$ *;* y_j^o *- Lokales Maximum (der* y' *- Daten)*

y_i^u *- Lokales Minimum ;* y_j^A *- Schwingungsweite Wellenzug*

y_j^m *- Schwingungsschwerpunkt Wellenzug ;* f_j^o *- Obergrenze Amplitudenfläche*

f_j^u *- Untergrenze Amplitudenfläche*

Durch diese Zuordnung entstehen somit Rechteckflächen in: y = f(i), wobei die f_j^o (i) die Obergrenze der Rechteckfläche und f_j^u (i) die Untergrenze der Rechteckfläche repräsentieren.
Jede dieser j-Rechteckflächen (Amplitudenflächen) repräsentiert somit je einen zugeordneten Wellenzug mit der Amplitude $y_j^A/2$, der um den Wert y_j^m schwingt und die Periode q besitzt.

Die Ausgangsdaten y(i = 1...N = 120) können nun durch diese Zuordnung durch j = 1,...,p Wellenzüge beschrieben werden, die dann jeweils die lokalen Extrema der Ausgangsdaten abbilden:

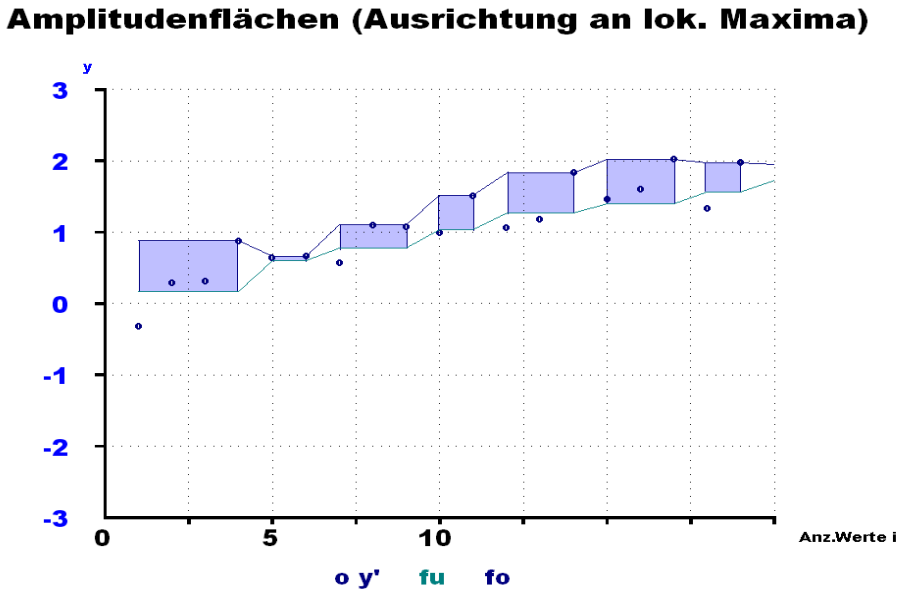


Abb. 1: Bildung Amplitudenflächen
mit: y' -verrauschte Sinuswerte ; f_o , f_u - Ober- Untergrenze Amplitudenfläche

Nach dem gleichen Verfahren können nun auch die Modelldaten : X(i) $:\equiv \hat{y}_i' = f(\hat{c}_k)$ behandelt werden. Hierdurch entsteht jetzt die Möglichkeit die Güte der Modelldaten X bzgl. ihrer Amplitudentreue zu den Ausgangsdaten y (neu) zu bewerten:

Postulat**:**
Die Modelldaten X(i) bilden die Ausgangsdaten y(i) korrekt ab, wenn die jeweiligen (j) Rechteckflächen (der Wellenzüge) übereinander liegen. Eine Teilabbildung ist vorhanden, wenn sie ein Schnittfläche besitzen, d.h. bei keiner Schnittfläche kann das Modell X die Ausgangsdaten y nicht abbilden.

4 Definition eines Bestimmtheitsmaßes für die (Flächen)Abbildung: BMF

Fy sei die nun durch Fx abzubildende (Rechteck)Fläche, dann gilt:

$$Fy = q \cdot (fy^o - fy^u) \qquad \text{und :} \qquad Fx = q \cdot (fx^o - fx^u)$$

Die Differenz zwischen den Flächen ist dann:

$$Fy - Fx = q \cdot ((fy^o - fy^u) - (fx^o - fx^u))$$

mit Definition folgender (Teil)Differenzen:

Differenz „oben“: $d_o = fy^o - fx^o$ Differenz „unten“: $d_u = fx^u - fy^u$
Differenz „Rand (1)“: $dr_1 = fx^o - fy^u$ Differenz „Rand (2)“: $dr_2 = fy^o - fx^u$

gelten dann die folgenden Aussagen:

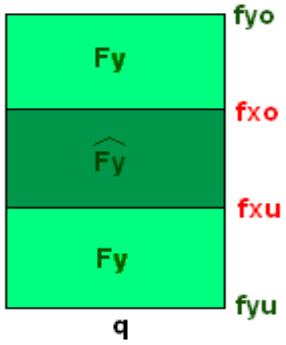
Die Abbildung von Fy durch Fx ist *korrekt*, wenn:

$ABS(d_o)+ABS(d_u)\equiv 0$ **und:** $\operatorname{sgn}(dr_1)>0$, $\operatorname{sgn}(dr_2)>0$

Es gibt *keine Abbildung* von Fy durch Fx , wenn:

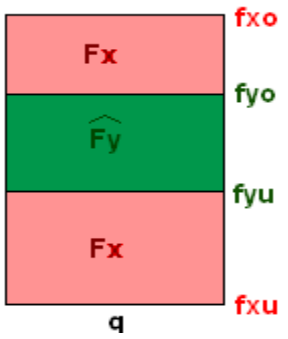
$\operatorname{sgn}(dr_1)\le 0$ **oder:** $\operatorname{sgn}(dr_2)\le 0$

Je größer also die „oberen“ und „unteren“ Differenzflächen werden, desto schlechter kann dann Fy durch Fx „erklärt“ werden.
Wird nun mit: $\hat{F}\,y$ die durch Fx erklärte (Teil/Schnitt)Fläche von Fy bezeichnet, so lassen sich (prinzipiell) die folgenden 6 Fälle unterscheiden:



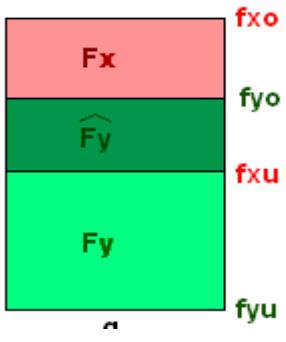
$Fy=Fx+q\cdot(d_o+d_u)$ $\operatorname{sgn}(dr_1)>0$; $\operatorname{sgn}(dr_2)>0$

$\hat{F}\,y=q\cdot(fx^o-fx^u)=Fx$



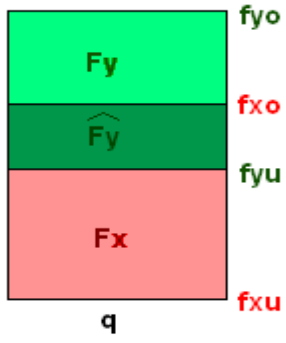
$Fy=Fx-q\cdot(d_o+d_u)$ $\operatorname{sgn}(dr_1)>0$; $\operatorname{sgn}(dr_2)>0$

$\hat{F}\,y=q\cdot(fy^o-fy^u)=Fy$



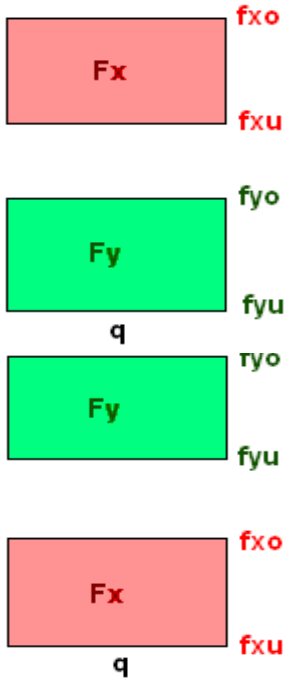
$Fy=Fx+q\cdot(d_u-d_o)$ $\operatorname{sgn}(dr_1)>0$; $\operatorname{sgn}(dr_2)>0$

$\hat{F}\,y=q\cdot(fy^o-fx^u)=q\cdot dr_2$



$Fy=Fx+q\cdot(d_o-d_u)$ $\operatorname{sgn}(dr_1)>0$; $\operatorname{sgn}(dr_2)>0$

$\hat{F}\,y=q\cdot(fx^o-fy^u)=q\cdot dr_1$



$\mathrm{sgn}(dr_1)>0 \;;\; \mathrm{sgn}(dr_2)<0(!)$

$\hat{F}\,y=0$

$Fy \neq f(Fx)$ $\quad \mathrm{sgn}(dr_1)<0(!)\;;\mathrm{sgn}(dr_2)>0$

$\hat{F}\,y=0$

Abb. 2: Definition Bestimmtheitsmaß für die Flächenabbildung(BMF)

Bestimmtheitsmaß für die Flächenabbildung:

$$BMF := \frac{2\cdot \hat{F}\,y}{(Fx+Fy)} \qquad \text{mit:} \quad Fx,Fy \neq 0 \tag{13}$$

→ $BMF \leq 1$

$BMF \equiv 0$ für : $\hat{F}\,y=0$

$BMF \equiv 1$ für : $\hat{F}\,y=Fx=Fy$

Der BMF-Wert wird für *jede Rechteck- Teilüberlappung*,_d.h. mit jeweils *gleichem q* von Fy (i) und Fx (i) gebildet.

Für den gesamten Abbildungsbereich (i=1,...,N) kann dann auch ein mittlerer BMF-Wert berechnet werden:

$$\overline{BMF} = 1/p\cdot\sum_{j=1}^{p} BMF(j) \tag{14}$$

5 Aussagen, Interpretation des BMF-Maßes

Eine andere (umgestellte) Form des BMF-Maßes:

$$\frac{2\cdot\hat{F}_y}{F_x+F_y} = c_{yx}\cdot\frac{\hat{F}_y}{F_x} + c_{xy}\cdot\frac{\hat{F}_y}{F_y}$$

mit: $c_{yx} = \frac{1}{(1+\frac{F_y}{F_x})}$; $c_{xy} = \frac{1}{(1+\frac{F_x}{F_y})}$

Hierbei ist dann : $c_{yx} + c_{xy} \equiv 1$ und: $0 \le c_{yx} \le 1$; $0 \le c_{xy} \le 1$

Wegen: $c_{yx} + c_{xy} = \frac{F_x \cdot 1}{F_x \cdot (1+\frac{F_y}{F_x})} + \frac{F_y \cdot 1}{F_y \cdot (1+\frac{F_x}{F_y})} = \frac{F_x}{F_x + F_y} + \frac{F_y}{F_y + F_x} = 1$

Somit könnte dann auch geschrieben werden:

$$\frac{2 \cdot \hat{F}_y}{F_x + F_y} = (1-c) \cdot \frac{\hat{F}_y}{F_x} + c \cdot \frac{\hat{F}_y}{F_y} \qquad (15)$$

mit : $$0 \le c = \frac{1}{(1+\frac{F_x}{F_y})} \le 1 \qquad (15a)$$

Für: $F_y > F_x \rightarrow \frac{F_x}{F_y} < 1 \rightarrow c > 0.5$

$F_y < F_x \rightarrow \frac{F_x}{F_y} > 1 \rightarrow c < 0.5$

Hiermit kann nun die Aussage des BMF-Maßes wie folgt beschrieben werden:

- Der BMF-Wert ist die gewichtete Summe aus der Schnittfläche $\hat{F}_y$ relativ zur Modellfläche Fx und aus der Schnittfläche $\hat{F}_y$ relativ zur Fläche der abzubildenden Daten Fy.

- Der Gewichtsfaktor c vor der Relation der abzubildenden Daten ist im Falle einer kleineren Modellfläche Fx als die Fläche der abzubildenden Daten Fy : 0.5 < c < 1. Somit ist dann der Faktor (1-c) vor der Relation zur Fläche der Modelldaten: 0 < (1-c) < 0.5

- Im Falle einer größeren Modellfläche Fx als die Fläche der abzubildenden Daten Fy gilt die entsprechende umgekehrte Aussage.

- Ist das Verhältnis: Fx / Fy sehr klein, d.h. hat das Modell im Vergleich zu den abzubildenden Daten nur eine geringe Strukturabbildung (kleine Flächen zwischen den lokalen Extrema), so wird der BMF-Wert dominiert durch das Verhältnis der Schnittfläche $\hat{F}_y$ zu den abzubildenden Daten Fy ($c \rightarrow 1$, d.h.: $(1-c) \rightarrow 0$)

Im umgekehrten Fall: $F_x / F_y \rightarrow \infty$ dominiert dann das Verhältnis der Schnittfläche $\hat{F}_y$ zu den Modelldaten Fx ($c \rightarrow 0$, d.h.: $(1-c) \rightarrow 1$)

- Nähern sich dagegen die Schnittfläche $\hat{F}_y$, die Modellfläche Fx und die der abzubildenden Daten Fy einander an, d.h. geht $BMF \to 1$, so geht $c \to 0.5$ und auch $(1-c) \to 0.5$, d.h. der BMF-Wert wird dann etwa zur Hälfte aus der Relation zu den abzubildenden Daten und etwa zur Hälfte aus der Relation zu den Modelldaten bestimmt.

6 Das Verhalten des BMF-Maßes bei einer iterativen EVM-Modellierung

Im Rahmen einer schrittweisen EVM-Modellierung zwischen RV → 0 (INV-Modell) und RV $\to \infty$ (OLS-Modell) bestehen nun die folgenden Eigenschaften:

Bei Variation des Rauschverhältnisses RV zwischen ∞ und 0 erfolgt ein „stetes Aufblähen" der modellierten Punktwolke, quasi aus der „Mitte der abzubildenden Punktwolke" heraus (der mittlere Fehler der Regressionsmodells ist Null !). Damit werden dann auch der RMSE und die Schnittflächen $\hat{F}_y$ zwischen den Flächen der lokalen Extrema des Modells (Fx) und den lokalen Extrema der abzubildenden Daten (Fy) im Mittel stetig größer.

Aufgrund der obigen Definition des BMF-Maßes besteht aber rechnerisch die Möglichkeit eines Maximums für den *mittleren* BMF-Wert. Numerische Experimente ergeben dann ein BMF(Max) in der Nähe des (unbekannten) RV bei „künstlich verrauschten" deterministischen Daten.

Für das obige weiß verrauschte Sinus-Beispiel ergibt sich dann:

RV(I)	RMSE	MWBMF	MW($\hat{F}_{yi}$)	
1E-6	0.69953	0.47008	0.43676	
0.1	0.69642	0.47172	0.43671	
0.2	0.69350	0.47324	0.43660	
0.3	0.69076	0.47476	0.43651	
...				
0.9	0.67748	0.48209	0.43526	
1.0	0.67569	0.48319	0.43507	
1.1	0.67401	0.48415	0.43483	
...				
3.6	0.65130	0.49245	0.42390	
3.7	**0.65084**	**0.49250**	**0.42349**	→ BMF(Max)
3.8	0.65039	0.49248	0.42305	
...				
9.8	0.63946	0.48988	0.40744	
9.9	0.63939	0.48983	0.40727	
10.0	0.63932	0.48979	0.40710	
...				
1E6	0.63513	0.47355	0.37971	

Tabelle 2: Iterationsergebnisse in RV= 0.1- Schritten

Mit: *RV -* *Rauschverhältnis*
MWBMF - *Mittelwert BMF nach Gleichung (14)*
MW($\hat{F}_{yi}$) - *Mittelwert der Schnittfläche von Fx und Fy*

→ BMF(Max)=0.4925 für Iterationsschritt RV=3.7

Aber:

→ MW($\hat{F}_{yi}$) besitzt kein Maximum, d.h. stetiger Anstieg für RV → 0

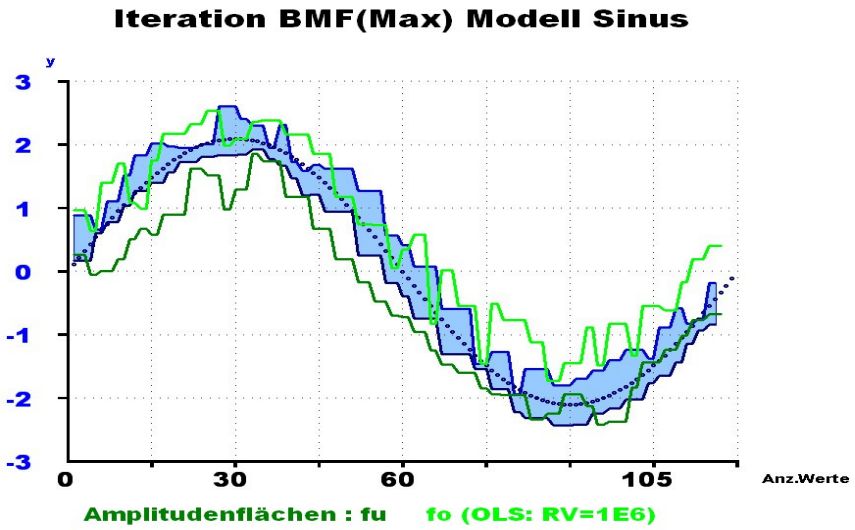


Abb.3: Iterationsschritt RV = 1E6 (OLS) ; BMF=0.473 ; C=1.776

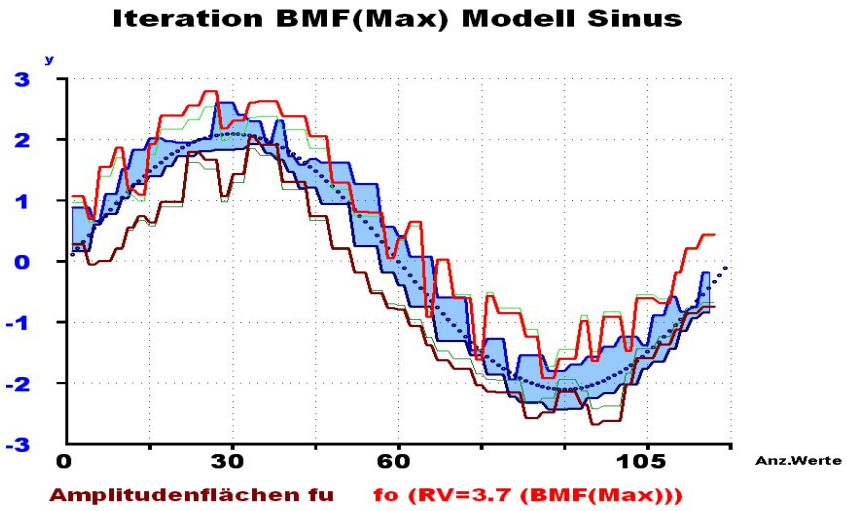


Abb.4: Iterationsschritt RV=3.7 ; BMF(Max)=0.493 → Ck=1.960

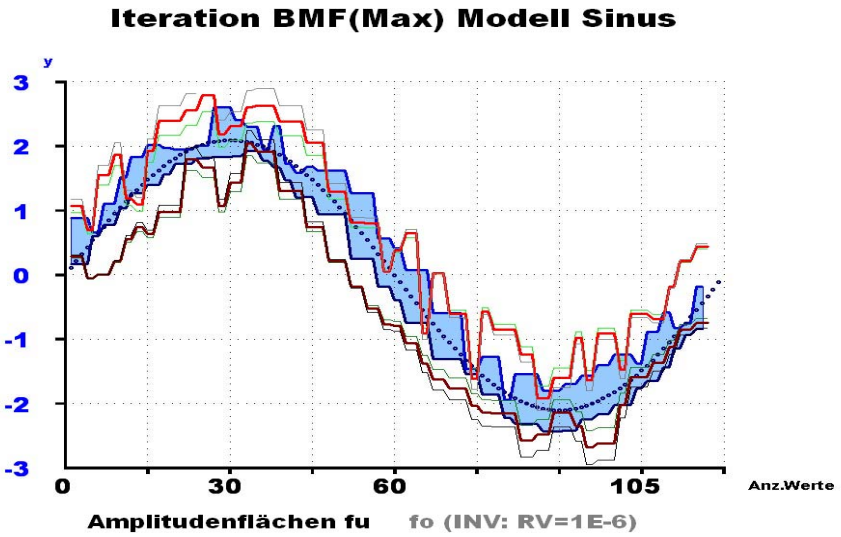


Abb.5: Iterationsschritt RV = 1E-6 (INV) ; BMF=0.470 ; C=2.16

Weitere Testberechnungen mit dann unterschiedlich stark verrauschten Daten sind in den folgenden Tabellen zusammengestellt:

	Eigenschaften ε_i					Eigenschaften δ_i				
Test	MW	S	SuAK	Cov1	Cov2	MW	S	SuAK	Cov1	Cov2
1a	8.5E-3	0.2921	1.333	7.7E-3	1.6E-2	7E-4	0.2964	1.457	1.1E-3	2.3E-3
1b	1E-17	0.2919	1.334	-2E-17	-1.4E-18	8E-10	0.2964	1.4457	-4E-17	-2.2E-17
2a	8.5E-3	0.2921	1.333	7.7E-3	1.6E-2	8E-17	1.4292	1.489	-0.109	-0.228
2b	1E-17	0.2919	1.334	-2E-17	-1.4E-18	6E-18	1.42096	1.469	-3E-16	-5E-16
3a	-4.1E-16	1.4413	1.346	0.081	0.168	8E-17	1.4292	1.489	-0.109	-0.228
3b	6E-17	1.4369	1.347	1.7E-16	-2E-16	6E-18	1.42096	1.469	-3E-16	-5E-16
4a	4.6E-18	0.3940	1.334	8.2E-3	1.7E-2	7E-4	0.2964	1.457	1.1E-3	2.3E-3
4b	-3.9E-18	0.3939	1.336	-9.4E-17	1.5E-17	8E-10	0.2964	1.4457	-4E-17	-2.2E-17

Tabelle 3: Eigenschaften des verwendeten Zufallsrauschens.

Mit: MW ; S - Mittelwert, Streuung; *Cov1 =* $S_{\varepsilon(\delta)x}$ *; Cov2 =* $S_{\varepsilon(\delta)y}$ *- Kovarianzen*

SuAK - Summe quadrierte Autokorrelationsfunktion; *RV - Rauschverhältnis* $\lambda = S_\delta^2 / S_\varepsilon^2$

Test	R^2	C(OLS)	C(INV)	RV	Ck(RV)	RV(I1)	Ck(I1)	BMF	RV(I2)	Ck(I2)	BMF
1a	0.824	1.776	2.155	1.03	2.072	3.7	1.960	0.493	1.9	2.024	0.493
1b	0.821	1.794	2.186	1.03	2.102	3.9	1.981	0.490	2.0	2.048	0.491
2a	0.401	1.589	3.961	23.94	1.849	7.5	2.386	0.486	17.8	1.940	0.507
2b	0.446	1.791	4.019	23.69	2.096	12.2	2.365	0.504	21.0	2.134	0.523
3a	0.130	0.425	3.269	0.98	1.583	1.4	1.171	0.490	1.0	1.558	0.490
3b	0.116	0.437	3.777	0.98	2.024	1.6	1.325	0.498	1.2	1.727	0.513
4a	0.722	1.572	2.179	0.566	2.090	5.0	1.791	0.450	1.4	1.992	0.455
4b	0.716	1.584	2.213	0.566	2.122	4.9	1.817	0.446	1.5	2.012	0.453

Tabelle 4: Testergebnisse der Regressionsrechnungen

Mit: R^2 - Korrelationskoeffizient

C(OLS) - Anstieg OLS- Ausgleichsgerade in Gleichung (2)

C(INV) -	*Kehrwert Anstieg der OLS- Ausgleichsgeraden in Gleichung (2b)*
Ck(RV) -	*Anstieg der EVM- Ausgleichsgeraden bei bekanntem RV (Glg. (9))*
Ck(I1) -	*Anstieg iterative EVM- Ausgleichsgerade (Glg.(9) mit(14),(15), (15a))*
Ck(I2) -	*Anstieg iterative EVM- Ausgleichsgerade (Glg.(9) mit(14),(16), (16a))*
BMF -	*BMF (Max) (Glg.(14)) der entsprechenden EVM- Iteration*

In der Tabelle 3 sind Kenngrößen für das verwendete Zufallsrauschen aufgeführt. Als Maß für die Bedingung (8c) (weißes Rauschen) wurde hier die Summe der quadrierten Autokorrelationsfunktion gewählt, die im Grenzfall einer verschwindenden Erhaltungsneigung gegen 1 gehen muss (siehe dazu auch im Anhang). Aus der Tabelle 3 folgt nun, dass die Bedingungen: (8a)-(8c) einer EVM-Modellierung in erster Näherung erfüllt werden, wobei in den Testreihen: _a noch „Restkovarianzen“ (8b) vorhanden sind, die dann in den Testreihen: _b beseitigt wurden. Die Stärke des Rauschens wurde so gewählt, dass sich für die verrauschten Abbildungen der Sinusschwingung Korrelationskoeffizienten zwischen etwa 0.9 und 0.1 ergeben, wobei hierzu die Variablen x als auch y mal mehr und mal weniger „stark verrauscht“ wurden. In der Tabelle 4 sind neben den Ergebnissen für die Anstiege: C nach der OLS- und der INV-Modellierung, als Unter- und Obergrenze des möglichen Anstieges der EVM-Modellierung auch die Ergebnisse der Anstiege mit dem jeweils gültigen Rauschverhältnis: RV aufgeführt: Ck(RV). Gleichzeitig werden die Ergebnisse der BMF(Max)-Iteration: Iterationsschritt RV(I) mit maximalem BMF-Wert und dem dazugehörenden Anstieg Ck (I) angegeben.

6.1 Auswertung der Testergebnisse:

Je höher das Fehlerrauschen der Daten absolut ($S_\delta^2, S_\varepsilon^2$) ist, um so stärker ist dann die Unterschätzung der OLS- Steigung und um so weiter liegen die Unter- und Obergrenze des EVM- Bereiches auseinander. Dies wird auch durch einen kleiner werdenden Korrelationskoeffizienten R^2 repräsentiert (siehe dazu auch Gleichung (6)).

Die EVM-Ergebnisse bei bekanntem Rauschverhältnis nach den Gleichungen (9) und (10) ergeben für die Testreihe _a mit Restkovarianzen (8b) noch leicht unterschätzte Steigungen. In der Testreihe _b ohne diese Restkovarianzen wird der unverrauschte Anstieg (2.1) dann deutlich besser „getroffen“.

Für die iterative EVM-Modellierung (I1) bei Maximierung des BMF-Maßes (14) existiert in allen Testreihen ein Maximum des BMF-Maßes und die hiermit erhaltenen Steigungen Ck(I1) liegen auch mehr oder weniger in der Nähe des unverrauschten Anstieges, wobei sie keinen so deutlichen Einfluss durch die Restkovarianzen (8b) zeigen. Für große RV (Test 2) liegt der iterative Anstieg Ck(I1) mit ~ 2.37 aber etwas oberhalb und für kleine RV und bei niedrigen R^2 (Test3 und 4) aber doch relativ weit unterhalb des unverrauschten Anstieges.

Optimierung des BMF-Maßes:

$$\frac{2\cdot\hat{F}_y}{F_x+F_y}=(1-c)\cdot\frac{\hat{F}_y}{F_x}+c\cdot\frac{\hat{F}_y}{F_y} \tag{16}$$

mit (nun) :
$$0\le c=\frac{1}{(1+\frac{S_y^2}{S_x^2}\cdot\frac{F_x}{F_y})}\le 1 \tag{16a}$$

und hierin mit: S_x^2 - Streuung der Modelldaten (X)

S_y^2 - Streuung der zu modellierenden Daten (y)

Durch Einführung dieses Optimierungsfaktors: S_y^2/S_x^2 in den Gewichtsfaktor in (16a) wird jetzt berücksichtigt, dass innerhalb des Iterationsbereiches von RV= $\lambda \to \infty$ (OLS) bis RV = $\lambda \to 0$ (INV) durch das „schrittweise Aufblähen" der modellierten Daten die Streuung der Modelldaten S_x^2 häufig über die der abzubildenden Daten S_y^2 hinausgehen muss, damit sich die Ausgleichssteigung entsprechend (weiter) anhebt und sich dadurch dann der unverrauschten annähern kann.

Für die iterative EVM-Modellierung (I2) bei Maximierung des BMF-Maßes (14) mit nun (16) und (16a) wird eine deutlich bessere Übereinstimmung mit dem unverrauschten Anstieg erhalten!

7 Eine erste Anwendung der iterativen EVM-Modellierung auf einen Pseudoproxy-Datensatz

Im obigen Beispiel wurde eine markante und regelmäßige Periodizität in Form einer Sinusschwingung verrauscht und mit Hilfe der iterativen EVM-Modellierung dann auch amplitudentreu abgebildet. Mit Hilfe von verrauschten Pseudoproxy-Daten soll dieses Iterationsverfahren nun in einem ersten Testbeispiel auch auf Daten mit realen, unregelmäßigen Periodizitäten angewendet werden. Bei diesen Daten handelt es sich um Klimamodellrechnungen (Modell „ERIK") der Jahresmitteltemperaturen (°K) einer Gitterzelle (Grid 1) vom Jahr 1000 bis 1990 [5].

7.1 Simulation Pseudoproxy „Erik" Grid 1:

$y = c_0 + c \cdot x$ mit : $c_0 = 0$, $c = 1$

$x_i' = erik_{grid}1_i + \varepsilon_i$ $\qquad$ $y_i' = erik_{grid}1_i + \delta_i$

	Eigenschaften ε_i				Eigenschaften δ_i			
Test	MW	S	SuAK	$S_{\varepsilon x}$	MW	S	SuAK	$S_{\delta y}$
Grid1	0.1157	1.2894	1.420	-1.346E-2	-0.0285	0.55070	1.355	-3.155E-3

Tabelle 5: Eigenschaften des Rauschens $(\delta_i, \varepsilon_i)$ *im Kalibrierbereich (N=101, 1889-1990).* (*Die Bezeichnungen sind analog zu Tabelle 3)*

Die Bedingungen: (8a) - (8c) für ein mittelwertfreies, kovarianzfreies weißes Rauschen sind in erster Näherung erfüllt.

Bestimmung der Modellparameter $(\hat{c}_{k0}, \hat{c}_k)$ für die Daten im Anpassungszeitraum (Kalibrierbereich) 1889 - 1990

Test	R^2	C(OLS)	C0	C(INV)	C0	RV	Ck(RV)	C0
Grid1	0.295	0.394	166.1	1.335	-92.2	0.182	1.048	-13.2

Test	R^2	RV(I2)	C_k(I2)	C0	BMF(Max)
Grid1	0.295	0.3	0.906	25.7	0.605

Tabelle 6: Testergebnisse der Regressionsrechnungen (Bezeichnungen analog zu Tabelle 4)

Die Modellierung der verrauschten Temperaturdaten ergibt mit R^2 = 0.295 (~ 0.3) einen von der Größenordnung her typischen Korrelationskoeffizienten für entsprechende Abbildungen mit realen Proxy (z.B. mit Baumringdaten).

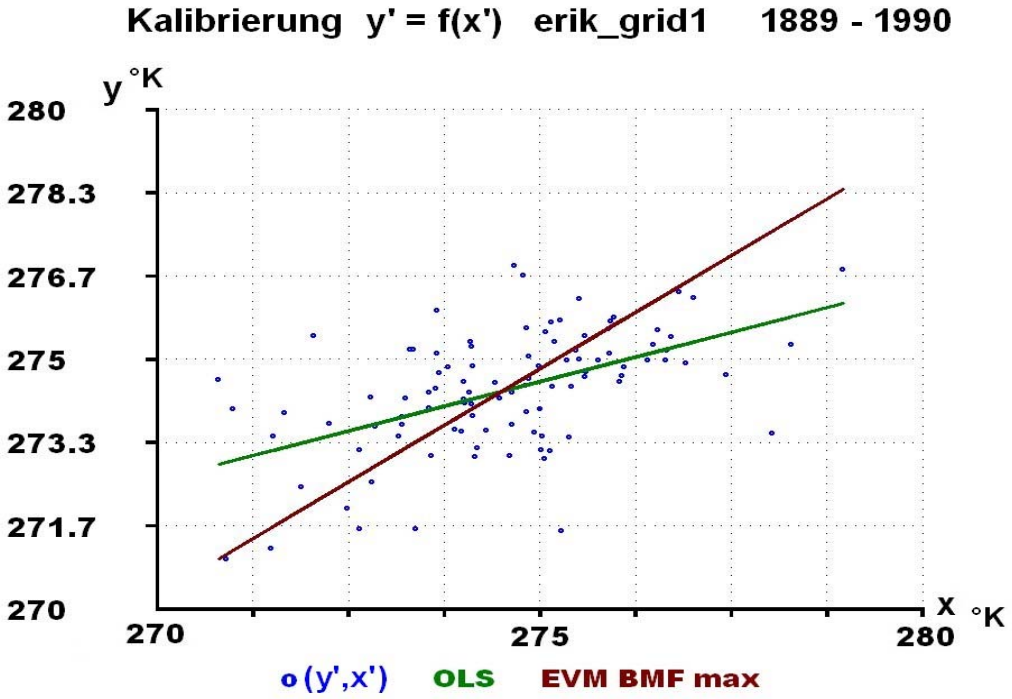


Abb. 6: Kalibrierung im Anpassungszeitraum 1889- 1990

mit: $(y'$ - *verrauschte Grid1- Modellwerte (s.o.)*

OLS - *Ausgleichsgerade (nach Glg.(4),(5))*
EVM BMFmax- *iterative EVM- Ausgleichsgerade (nach Glg.(14), (16a))*

Mit den im Anpassungszeitraum 1889 bis 1990 bestimmten Modellparametern (Tabelle 6) wird dann der gesamte Zeitraum 1000 bis 1990 entsprechend „rekonstruiert“.

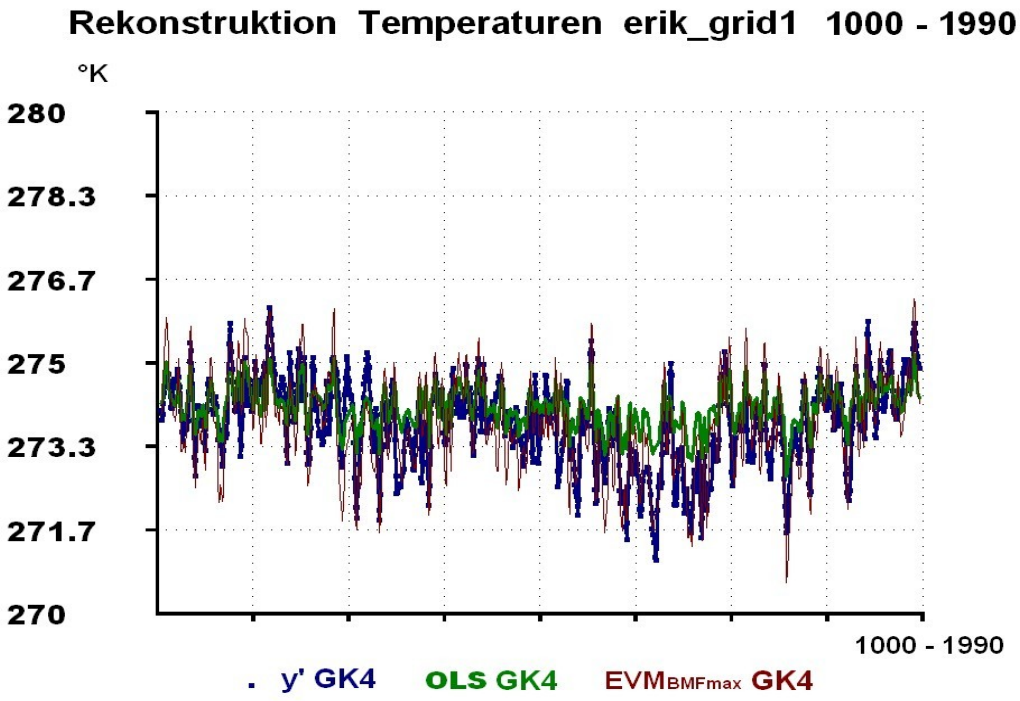


Abb. 7: Rekonstruktion 1000- 1990 nach „Gaussglättung“ über 4 Jahre (GK4)

In der Abbildung 7 wurden bereits alle Daten mit einem Gaussfilter über 4 Jahre (GK=4) geglättet. Solch eine Glättung von verrauschten Daten ist grundsätzlich immer notwendig, da durch ein weißes und unabhängiges Zufallsrauschen auch unvermeidbar zufällige Strukturen entstehen. Nach der binomialen Wahrscheinlichkeitsverteilung („Urnenmodell“) sind dann Strukturen mit Perioden im vierfachen Messwertabstand noch mit einer Wahrscheinlichkeit von ca. 10 % rein zufällig. Hierbei muss aber der Grundsatz gelten, dass so viel wie nötig, aber so wenig wie möglich geglättet wird, da jedes Tiefpassfilter entsprechend seiner Durchlasscharakteristik das gesamte Spektrum beeinflusst.
Auch in diesem Zusammenhang kann wieder das BMF-Maß benutzt werden, um durch eine schrittweise Gaussfilterung, beginnend mit GK=2 (Mindestglättung), den Punkt zu finden, bei dem das BMF-Maß wieder ein Maximum und (gleichzeitig) der RMSE ein Minimum besitzt :

7.2 Endergebnis der Rekonstruktion :

Die Suche nach der „optimalen“ Gaussfilterung unter den Bedingungen:

- möglichst geringe Glättung (Mindestglättung aber Gk=2 !)
- Maximum BMF und
- Minimum RMSE

ergibt dann eine (notwendige) Filterung für die abzubildenden Daten über Gk=13 und für das Rekonstruktionsmodell über Gk=16 (Jahre)

Bei dieser Glättung besitzt dann das Modell einen (minimalen) Fehler von: RMSE = 0.35 °K
bei einer gleichzeitigen (maximalen) Amplitudenabbildung von: BMF=0.566

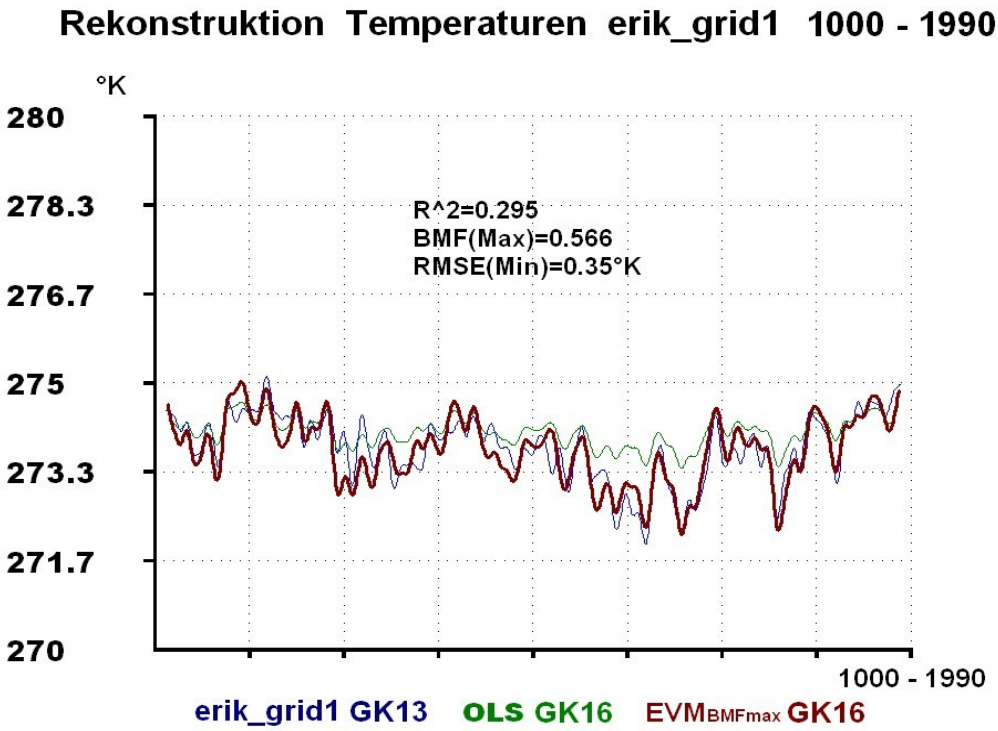


Abb.:8 Endergebnis der Rekonstruktion 1000- 1990

Mit: erik_grid1 GK13 - Grid1- Modellwerte (s.o.) mit Gaussglättung (13 Jahre)
OLS GK16 - OLS- Ergebnisse (Glg.(4),(5)) mit Gaussglättung (16 Jahre)
EVMBMFmaxGK16- iterative EVM- Ergebnisse (Glg.(14), (16a)) mit Gaussglättung (16 Jahre)

Auch für dieses Testbeispiel mit einem (realen) Spektrum an Schwankungen wurden mit Hilfe der iterativen EVM-Modellierung amplitudentreue Abbildungen erhalten.
Hierbei wurde das BMF- Maß zusätzlich benutzt, um im Rahmen einer schrittweisen Tiefpassfilterung (Gaussfilter) eine Glättung der Daten zu erhalten, die dann auf einen minimalen Modellfehler (RMSE) und gleichzeitig einer maximaler Amplitudenabbildung (BMF) hin optimiert wurde.

8 Zusammenfassung, Ausblick

Durch die Einführung eines Bestimmtheitsmaßes für die (Amplituden) Flächenabbildung (BMF): (13),(14) mit Hilfe von Wellenzügen, die aus den lokalen Extrema einer Datenreihe definierbar sind, kann dann die Amplitudentreue zwischen einer abzubildenden y(i) und einer Modelldatenreihe x(i) bewertet werden (Punkte 3,4,5). Durch die Zuordnung aller Daten innerhalb eines Wellenzuges zu einer sich aus der doppelten Amplitude und der „Länge“(Periode) des Wellenzuges in: y=f(i) definierbaren Fläche erfolgt eine Daten-reduktion. Über die Differenzbildung zwischen den Flächen zweier Datenreihen (y,x,(i)) werden quasi „unscharfe Differenzen“ realisiert, im Gegensatz zu den „exakten“ Differenzen im Punkt (i), wie sie üblicherweise, z.B. beim RMSE verwendet werden. Die mathematischen Eigenschaften solcher Differenzen, speziell zur Gültigkeit im „mathematischen Konsens“, sind jedoch unklar und müssten untersucht werden.

Das BMF-Maß kann im Rahmen einer iterativen EVM-Modellierung: (9,10) näherungsweise zum Auffinden eines unbekannten Rauschverhältnisses des Fehlerrauschens benutzt werden, da es bei den bisher durchgeführten Testrechnungen ein Maximum in dessen Nähe besitzt. Dies wäre von herausragender Bedeutung, da eine Anwendung der allgemeingültigeren EVM-Modellierung häufig an der Unkenntnis des Fehlerrauschens in den Daten scheitert. Eine Benutzung von Methoden, die nicht von Fehlern in (beiden) Variablen ausgehen führen dann zum erwähnten Amplitudenproblem (Punkt 2), z.B. in Rekonstruktionen klimatischer Bedingungen mit Hilfe von Proxy in der Paläöklimatologie (z.B.: [5]). Weitere Testrechnungen zur Gültigkeit und Güte dieser Näherungsmethode, vor allem in der Anwendung auf Daten im Rahmen von Klimarekonstruktionen mit Proxy , sind jedoch notwendig. Eine erste Testrechnung für Pseudoproxy im Punkt 7 zeigt zunächst gute Ergebnisse.

Bei der notwendigen Tiefpassfilterung von Modellierungen mit (stark) verrauschten Daten könnte das BMF-Maß als Ergänzung zum RMSE ebenfalls hilfreich sein, um neben einem kleinen Standardfehler auch einen möglichst geringen Verlust an Information in den Amplituden des Modellierungsergebnisses durch die Tiefpassfilterung zu erreichen. Auch hierzu sind weitere Tests notwendig.

Anhang:

RMSE (Root Mean Square Error) : RMSE = $\sqrt{1/N \cdot \sum_{i=1}^{N} (\hat{y}_i - y_i)^2}$

SuAK (Summe quadrierte Autokorrelationsfunktion) : $\sum_{k=1}^{N} R_y^2(k)$

mit:

Autokorrelationsfunktion von y zur Verschiebung (Lag) k : $R_y(k) = \frac{S_{yi,yi+k}}{S_{yi,yi+0}}$

und:

Autokovarianz (funktion) zum Lag k: $S_{yi,yi+k} = \frac{1}{(N-1)} \cdot \sum_{i=1}^{N} (y_i - \bar{y}) \cdot (y_{i+k} - \bar{y})$

Literatur

[1] Fuller, W, 1987: Measurement error models. Wiley; New York, 440 p.

[2] Hartung, J., 1999: Statistik, Lehr- und Handbuch der angewandten Statistik. R. Oldenbourg Verlag München Wien.

[3] Kutzbach, L., and Thees, B., 2009: Identification of linear relationships from noisy data using errors-in-variables models- relevance for reconstruction of past climate from tree-ring (and other) proxy information. Climatic Change, eingereicht.

[4] Stahel, W., 2006: Lineare Regression Seminar für Statistik, ETH Zürich, 126 S Vorlesungsscript: http://stat.ethz.ch/~stahel/courses/regression/reg1-script.pdf.

[5] von Storch, H., Zorita, E., Jones, J., Dimitriev, Y., Tett, S., González-Rouco, F. 2004: Reconstructing past climate from noisy data. Science 306: 679-682.