Residuen helfen gut zu modellieren

MARKUS VOGEL, HEIDELBERG UND ANDREAS EICHLER, FREIBURG

Zusammenfassung: *Beim Modellieren von Daten gilt: Das Modell hat zu den Daten zu passen und nicht umgekehrt. Die Streuung realer Daten um ein Modell, mit dem ein erkannter Datentrend beschrieben wird, ist keine bloße lästige Abweichung von den vermeintlich „echten" Modellwerten. Die Differenzen zwischen den Daten und den Modellwerten, die Residuen, werden bei der Datenanpassung bewusst in Kauf genommen – nur der Trend der Daten, nicht die Daten selbst sollen erfasst werden. Die Residuen geben Aufschluss über die Modellgüte. Aber auch darüber hinaus enthalten Residuen wertvolle Informationen. In diesem Beitrag wird die Bedeutung der Residuen beim Modellieren von Daten aus verschiedenen Perspektiven beleuchtet.*

1 Residuen geben der Datenanpassung ein Maß

Der Schulunterricht bietet auf allen Alterstufen zahlreiche Möglichkeiten, um mit Daten Zusammenhänge aus Alltag, Natur und Technik zu untersuchen: Ergebnisse beim Sportunterricht, Wetter- und Temperaturanalysen, Zusammenhänge in Umfrageergebnissen, Gesetzmäßigkeiten beim Wachstum von Hefen oder Pilzen, Zusammenhang zwischen Geschwindigkeit und Bremsweg, Schallgeschwindigkeit, etc. Bei der Analyse solcher Daten geht es darum, dass die Schülerinnen und Schüler Gesetzmäßigkeiten oder Trends in Daten ausfindig machen und diese durch Funktionen modellieren.[1] Dabei können je nach Datenlage und Kontext einerseits und Vorwissen der modellierenden Person andererseits unterschiedliche Modellierungstechniken zum Einsatz kommen (vgl. Engel & Vogel, 2006). Für die Schule bietet die Arbeit mit elementaren Funktionen besondere Möglichkeiten, die Leitideen *Daten und Zufall*, *Funktionaler Zusammenhang* und die Kompetenz *Modellieren* miteinander zu vernetzen und so den Schülerinnen und Schülern den umwelterschließenden Aspekt (vgl. Vollrath, 2003) des Funktionsbegriffs erfahrbar werden zu lassen.

Für das Auffinden einer geeigneten Funktion ist das Streudiagramm von wesentlicher Bedeutung: Hier werden die Daten und eine mögliche Modellfunktion in einer graphischen Einheit abgebildet. Auf dieser Ebene betrachtet lautet die Aufgabe: Finde eine Funktion, deren Graph „möglichst gut" zu der Punktwolke der Daten „passt".

In der Formulierung „möglichst gut passen" wird eine inhaltliche Vorstellung über die Modellierungsgüte in der Frage ausgedrückt, wie und in welchem Maß Funktion und Daten zueinander in Beziehung stehen. Mit den Residuen, definiert als Differenz $r_i = y_i - f(x_i)$ zwischen den Datenpunkten $(x_i|y_i)$ und den jeweils entsprechenden Punkten $(x_i|f(x_i))$ der modellierenden Funktion $f(x)$, steht ein einfaches Abweichungsmaß zur Verfügung. Im Residuendiagramm werden die Differenzen als Lotabstände zwischen Datenwerten und Funktionswerten entlang einer Nulllinie abgebildet, welche die gleiche Skalierung wie das zugehörige Streudiagramm aufweist. Hier lässt sich weiter spezifizieren, was eine „möglichst gute" Datenanpassung meint: Nach der begründeten Entscheidung für eine Modellfunktion soll dieses so angepasst werden, dass die Residuen möglichst klein sowie zufällig im Sinne von trendfrei sein sollten und sich insgesamt nach oben und unten ausgleichen (vgl. Biehler & Schweynoch, 1999).

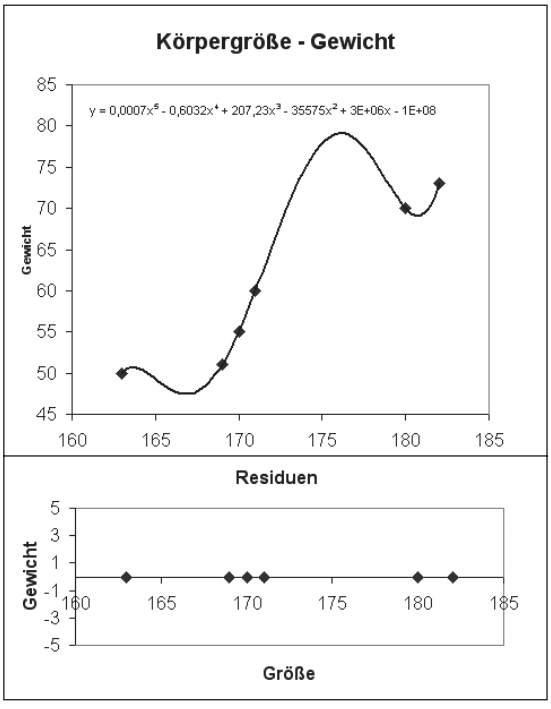

Abb. 1: Perfekte Datenanpassung?

Bei dem Versuch, die Residuen durch die Funktionseinpassung möglichst klein und zufällig hin zu bekommen, ist in modellierungstechnischer Hinsicht wesentlich: Die Minimierung darf nicht um den Preis einer sachgerechten funktionalen Modellierung geschehen. Im Extremfall hieße das, dass die bestmög-

liche funktionale Datenmodellierung immer darin besteht, dass alle Abweichungen zwischen Modell und Daten – die Residuen – verschwinden und der Funktionsgraph genau durch alle n Datenpunkte $(x_1, y_1), \ldots, (x_n, y_n)$ verläuft. Über die Interpolation durch ein Polynom der Form $p(x) = a_0 + a_1x + a_2x^2 + \ldots + a_kx^k$ wäre mit $k = n - 1$ eine solche Funktion $p(x)$ für n Datenpunkte prinzipiell immer zu finden. Diese Datenanpassung ist jedoch im Allgemeinen nicht sinnvoll zu interpretieren wie das Beispiel in Fig. 1 zeigt. Durch die vollständige Datenerfassung verliert das funktionale Modell die Möglichkeit, einen sachlich sinnvoll begründbaren Zusammenhang zwischen Körpergröße und Gewicht trendgemäß abzubilden. Dadurch büßt es ein zentrales Merkmal ein, welches ein Modell nach Stachowiak (1973) kennzeichnet: das Abbildungsmerkmal. Es ist ein wichtiges Ziel, dass die Schülerinnen und Schüler lernen, zwischen der kontextfreien Punkte-Erfassung und der funktionalen Anpassung zu unterscheiden, welche die kontextuelle Bedeutung der Datenwolke im Blick hat.

2 Residuen verweisen auf den Modellcharakter

Daten sind Zahlen mit Kontext. Sie quantifizieren Beobachtungen von Phänomenen der natürlichen, technischen oder sozialen Umwelt. In dem seit TIMSS und PISA vielfach rekurrierten mathematischen Modellierungskreislauf[2] sind die Daten als Realmodell eines zugrunde liegenden Phänomens einzuordnen (vgl. Vogel, 2006 und Eichler & Vogel, 2009). In den Daten manifestiert sich der Aspekt des Ausgangsphänomens, welcher für eine strukturelle Betrachtung wesentlich und zugänglich ist. Der Mathematisierungsvorgang besteht darin, im Nebel der Datenwolke Strukturen ausfindig zu machen und diese Strukturen auf der mathematischen Modellebene durch eine mathematische Funktion zu beschreiben. Aus dem, was Borovcnik (2005) als Strukturgleichung bezeichnet, lässt sich folgende Modellierungsgleichung ableiten (Vogel, 2006):

Daten = Funktion + Residuen

Die modellierende Funktion steht als „deterministisches Modellierungskonzentrat“ für die erklärte Variabilität der Daten. Demgegenüber zeugen die Residuen (= Daten – Funktion) von der nicht erklärten Variabilität der Daten. Diese Differenz kann mit dem Konstrukt Zufall beschrieben und als stochastische Komponente (z. B. Engel, 1998) mathematisch erfasst werden. Die reale und die mathematische Modellebene unterscheiden sich durch die Residuen. Gleichsam sind beide Modellebenen durch die Residuen verknüpft – sie vermitteln beim Modellierungsvorgang zwischen Daten und Funktion. Dies ist in Fig. 2 anhand des Beispiels der CO_2-Daten dargestellt, die auf Mauna Loa im Zeitraum von 1995 bis 2005 gemessen wurden. Der jahreszeitlich unabhängige Anstieg ist durch eine gleitende Mittelwertkurve modelliert. Übrig bleiben die Residuen, in denen saisonale und durch Zufall erklärte Schwankungen enthalten sind.

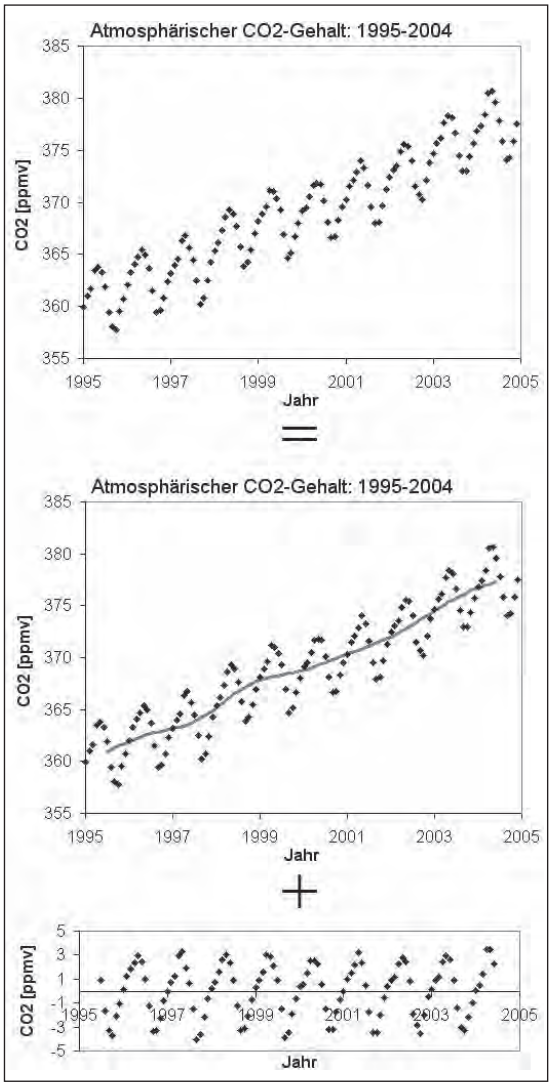

Fig. 2: Residuen vermitteln zwischen Daten und Funktionen

Der Unterschied, für den die Residuen stehen, ist in didaktischer Hinsicht dann bedeutsam, wenn auf diese Weise der Unterschied zwischen Modell und Realität betont werden kann. Häufig verwechseln Schülerinnen und Schüler Modell und Realität miteinander (z. B. Humenberger & Reichel, 1995). Im mathematisch-naturwissenschaftlichen Unterricht taucht dieses Problem oft beim Nach-Entdecken von Gesetzmäßigkeiten, wie z. B. $s = \frac{1}{2}g \cdot t^2$ oder $F = m \cdot a$ auf: Unterrichtliche Experimente dienen in der Regel dazu, diese Gesetzmäßigkeiten als gültig

dadurch nachzuweisen, dass sie sich wiederholt aufzeigen lassen. Ist aber vor der Analyse von Messdaten bereits bekannt, was an Trend zu finden ist, und tritt die modellierende Person nur mit dieser „Trendbrille“ an die Daten heran, dann liegt die Gefahr nahe, dass die Daten als etwas Störendes und vom vermeintlich „wahren“ Trend Ablenkendes wahrgenommen werden.

In Schulbüchern werden leider häufig Abbildungen gezeigt, bei denen Messwerte genau auf entsprechenden Funktionsgraphen liegen. So wird diese falsche Sichtweise noch zusätzlich unterstützt. Auch der in diesem Zusammenhang übliche Terminus „Fehlerrechnung“ wirkt hier didaktisch sehr unglücklich: Er legt die Sichtweise nahe, dass Daten etwas Fehlerbehaftetes sind, wohingegen der identfizierte Trend als etwas Fehlerbereinigtes erscheint. Tritt man dagegen „vor-Urteils-frei“ an die Daten heran, dann stellen die Residuen eine wichtige Informationsquelle für den Modellierungsvorgang dar. Sie enthalten noch die Informationen von den Originaldaten, welche beim (ersten) Modellieren nicht berücksichtigt wurden.

3 Was Residuen zu sagen haben – Beispiele

Das Residuendiagramm lässt sich mit der didaktischen Metapher einer „Modellierungs-Lupe“ erklären: Hier wird das vergrößert anvisiert, was bei der funktionalen Modellierung – die mit dem Trend eher das „Ganze“ global in den Blick nimmt – übrig geblieben ist. Bei genauer Betrachtung lässt sich auch hieraus noch einiges herauslesen:

3.1 Unterschiedlich große Streuung in einer Datenmesssung (Heteroskedastizität)

Bei der Messung von gleichartigen Objekten oder mathematischen Größen (z. B. Länge) verschiedener Größenordnung lässt sich beobachten, dass bei einer funktionalen Anpassung mit größeren Messwerten größere Residuen einhergehen können.

In Fig. 3 ist die lineare Modellierung eines Datensatzes abgebildet, der die Messung von Längen und Breiten von Buttermuscheln dokumentiert. Die Gerade steht für ein durchschnittliches Verhältnis von Länge und Breite. In dem Residuendiagramm ist eine Zunahme der Streuungen zu sehen. Diese Beobachtung lässt sich bei der entsprechenden Vermessung anderer Objekte, wie z. B. Ahorn-Blätter, wiederholen. Auch bei der nichtlinearen Modellierung eines Datensatzes zum Phänomen des freien Falls (Biehler, 2006) zeigt sich diese Form der Heteroskedastizität in den Residuen.

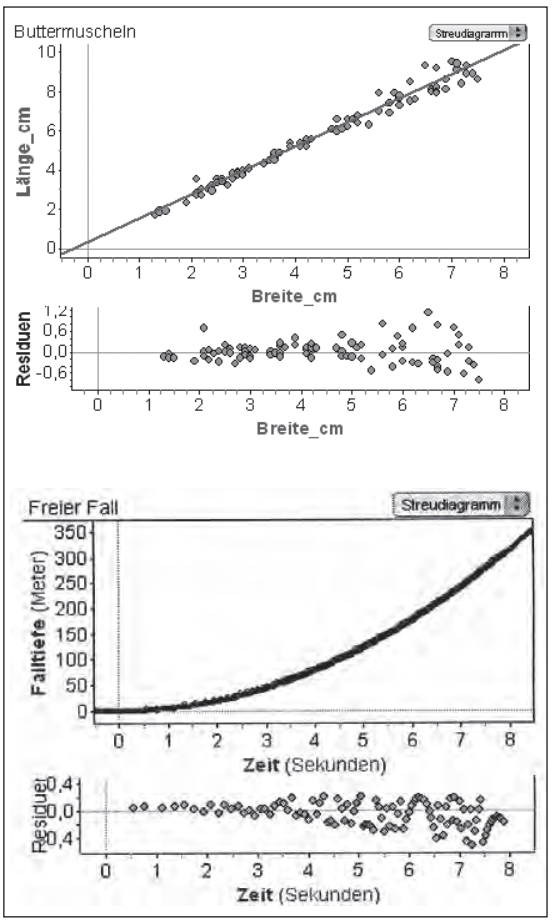

Fig. 3: Größere Datenstreuung bei größeren Messwerten

Während bei den Buttermuscheln die größere Formvariabilität größerer Muscheln als plausible Ursache erscheint, erklärt im zweiten Beispiel die zunehmende Messungenauigkeit bei größeren Falltiefen die Zunahme der Datenstreuung. Etwas schwieriger zu entdecken, aber in ähnlicher Weise zu erklären, ist die Beobachtung bei einer Zeitreihenanalyse zum atmosphärischen CO2-Gehalt (Eichler & Vogel, 2009) für die Frühjahrs- und Herbstmonate: in diesen Zeiträumen sind durch die Be- und Entlaubung die stärksten Veränderungen des Vegetationseinflusses auf den atmosphärischen CO2-Gehalt zu verzeichnen. Auch dies zeichnet sich in den Residuen ab.

3.2 Zunehmende „Verschlechterung“ eines Trendmodells

Aus der Modellierungsgleichung Daten = Funktion + Residuen folgt unmittelbar, dass eine zunehmend stärker werdende systematische Residuenabweichung (nach oben oder unten) darauf schließen lässt, dass die gewählte Modellfunktion zunehmend an Erklärkraft verliert. Diese Beobachtung ist beispielsweise dann zu machen, wenn der Zerfall eines radioaktiven Präparats ohne Berücksichtigung einer vorhandenen

Grundstreuung funktional modelliert wird. In Fig. 4 ist der radioaktive Zerfall einer Probe metastabilen Bariums 137 (Ba-137m) dokumentiert.

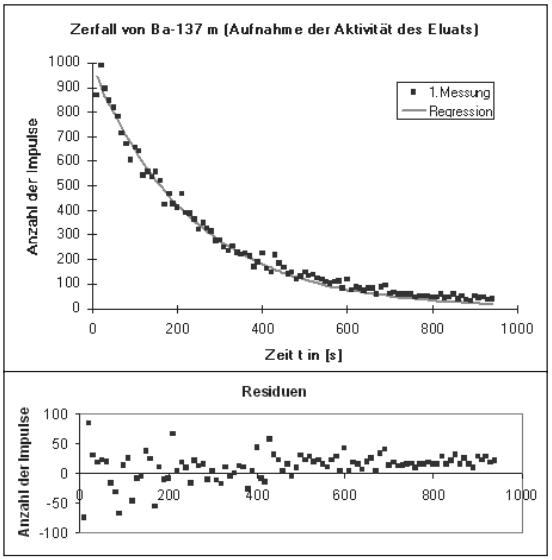

Fig. 4: Zunehmende „Verschlechterung“ des Trendmodells

Zunächst ist sehr deutlich die oben beschriebene Heteroskedastizität zu sehen. In diesem Beispiel wird allerdings die Streuung mit kleineren Messwerten zunehmend kleiner. Außerdem zeigt das Residuendiagramm deutlich, wie die gewählte funktionale Beschreibung einer exponentiellen Abnahme zunehmend weniger greift, obwohl sich dafür gute Gründe anführen lassen: Für die momentane zeitliche Änderung der radioaktiven Kerne gilt (*N(t)* ist die Anzahl der nach Ablauf der Zeit *t* noch nicht zerfallenen Kerne, λ ist die sog. Zerfallskonstante):

$$\frac{dN(t)}{dt} = -\lambda \cdot N(t)$$

Umsortieren nach Variablen und beidseitige Integration ergeben (N_0 ist die Anzahl der anfangs vorhandenen Kerne):

$$[\ln N(t)]_{N_0}^{N(t)} = -\lambda \cdot [t]_0^t$$

Woraus sich nach weiteren Schritten schließlich das bekannte Zerfallsgesetz ergibt:

$$N(t) = N_0 \cdot e^{-\lambda \cdot t}$$

Entscheidend für die zunehmende „Verschlechterung“ der Modell-Exponentialfunktion ist, dass in den Residuen die so genannte „Nullrate“ eingeht. Die Streuung der Nullrate wird als konstant großes und zufälliges Rauschen angenommen. In diesem Rauschen geht der schwächer werdende Trend zunehmend unter.

3.3 Residuen als Modellierungshilfe

Überlagern sich verschiedene Trends additiv, kann das Residuendiagramm helfen, diese schrittweise zu modellieren. In Fig. 5 ist ein Datensatz zur gleichmäßig beschleunigten Bewegung an einer schiefen Ebene abgebildet (vgl. Girwidz & Vogel, 2007). Die Daten wurden mit einer Luftkissenfahrbahn erhoben, wie sie sich in den meisten schulischen Physiksammlungen befindet. Bei dem Vorgang beschleunigte der Luftkissenschlitten gleichmäßig aus einer konstanten Anfangsgeschwindigkeit heraus. Es ist nicht schwer, sich mit einem Tabellenkalkulationsprogramm oder einem grafikfähigen Taschenrechner ein quadratisches Funktionenmodell für die Daten berechnen zu lassen. In der Schule hat dies allerdings den Nachteil, dass die Schüler den mathematischen Hintergrund des entsprechenden Tastendrucks auf die Rechnertaste oftmals nicht nachvollziehen können. Mit Hilfe des Residuendiagramms lassen sich alternativ die Daten per Schieberegler in einen reinquadratischen Teil (Beschleunigung) und einen (nahezu) linearen Teil (Anfangsgeschwindigkeit) für die Schüler anschaulich zerlegen.

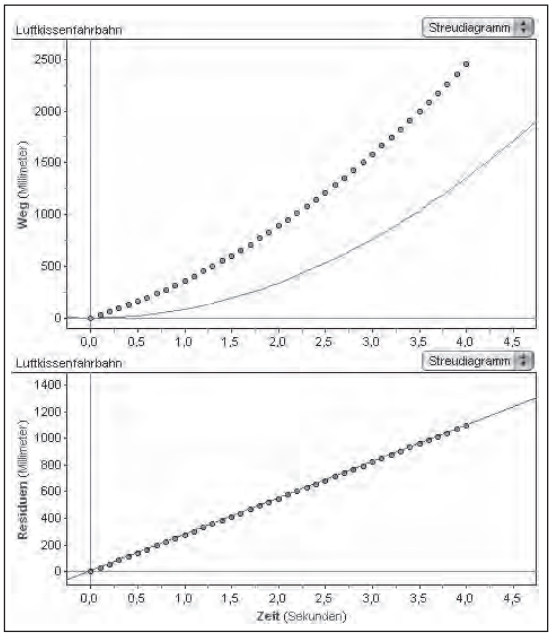

Fig. 5: Modellierung zu Daten einer gleichförmig beschleunigten Bewegung

Wenn in einem zweiten Modellierungsschritt der lineare Residuentrend funktional erfasst wird, dann lässt sich über die additive Verknüpfung beider Funktionen der gesamte Weg modellieren. Auch das, was dann noch übrig bleibt, wird wiederum im Residuendiagramm eingetragen. Erwartungsgemäß wird dann nur noch ein „zufälliges“ Muster zu sehen sein, das entsprechend des oben genannten Phänomens der Heteroskedastizität mit größeren Messwerten größer wird.

Das „Papierfaltproblem“ nach Biehler et al. (2007) zeigt sehr eindrücklich, wie der Blick auf die Residuen Ideen zu einer alternativen funktionalen Anpassung liefern kann. Die Aufgabenstellung besteht darin, von einem querliegenden DIN-A4-Blatt die linke obere Ecke auf eine beliebige Stelle der unteren Kante des Blattes zu falten. In dem variablen rechtwinkligen Dreieck, welches dadurch links unten entsteht, soll der Zusammenhang zwischen der variablen Grundseite und dem Flächeninhalt durch eine Funktion beschrieben werden. In den Streudiagrammen der Fig. 5 sind ein entsprechender Datensatz und eine funktionale Modellierung abgebildet.

Während mit dem alleinigen Blick auf das Streudiagramm eine quadratische Datenanpassung durchaus zunächst plausibel erscheint (Fig. 6 oben), wirft der s-förmige Verlauf im zugehörigen Residuendiagramm Fragen auf. Die Eigenschaften, welches ein funktionales Modell für die Residuen hätte (drei Nullstellen, zwei Extrema), legen die Idee nahe, die quadratische Anpassung durch eine kubische Funktion additiv zu ergänzen. Ob dies im Unterricht mit Computerunterstützung eher systematisch probierend oder theoriegeleitet geschieht (vgl. Biehler et al., 2007) bleibt in der Entscheidung der Lehrkraft. Das Ergebnis ist eine funktionalen Anpassung, in deren Residuendiagramm keine vergleichbaren Auffälligkeiten zu entdecken sind (Fig. 6 unten). In diesem Beispiel kommt der erwähnte „Modellierungs-Lupen“-Charakter besonders anschaulich zum Ausdruck.

4 Zusammenfassung

Die vorausgehenden Überlegungen sollen veranschaulichen, dass die Residuen nicht als „modellierungstechnischer Abfall“ gering zu schätzen sind. Residuen enthalten wertvolle Informationen zu dem Phänomen, welches den Daten zugrunde liegt, und können im Modellierungsprozess wichtige Hinweise für die sukzessive Datenanpassung geben. Sie stehen für das Spannungsverhältnis zwischen Daten und dem, was an Trend in den Daten zu entdecken ist. So kennzeichnen die Residuen die funktionale Datenanpassung als das, was sie ist: eine Modellierung, die auf vereinfachenden Annahmen und Entscheidungen beruht, um im Überangebot der Dateninformation das wesentlich Erscheinende besser zu erkennen.

Seine Wertschätzung für Residuen formuliert Tim Erickson (2005, S. 85) folgendermaßen: „How close is close enough? A real function never goes through all the points. It only comes close. There is no firm rule, but we will learn about a tool here – the residual plot – that may be the most important piece of data analysis machinery since the slide rule.“

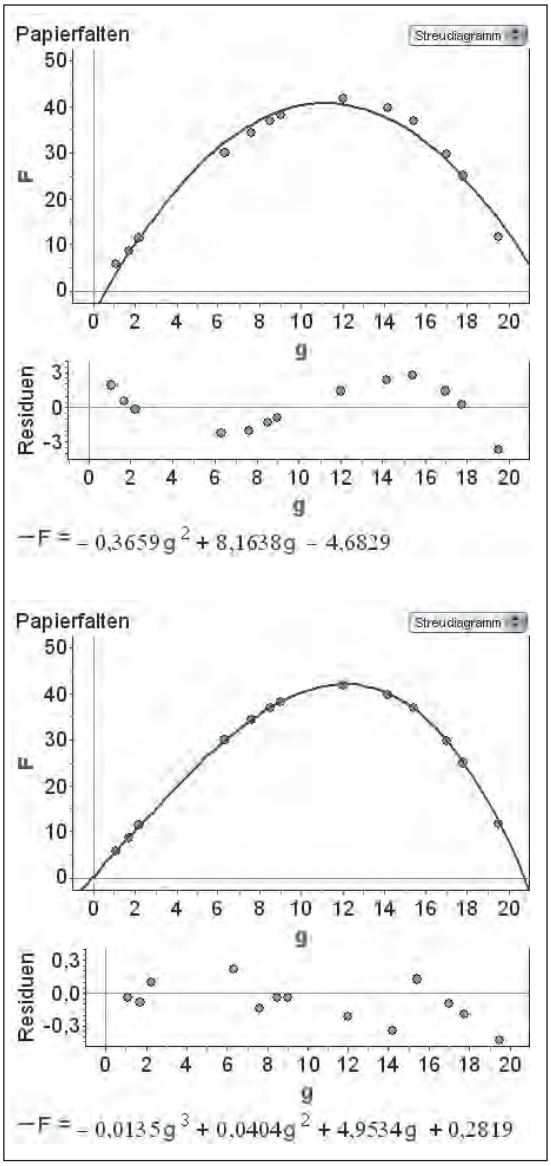

Fig. 6: Residuen bei funktionalen Modellierungen zum Papierfaltproblem

Anmerkungen

1 Bei univariaten Daten lässt sich ein Lagewert wie z. B. das arithmetische Mittel als konstante Funktion deuten.

2 In der didaktischen Literatur finden sich mittlerweile verschiedene Modellierungskreisläufe, die sich zwar hinsichtlich der verwendeten Begriffe und ihrer schematischen Anordnung z. T. unterscheiden, jedoch nicht in der Grundstruktur.

Literatur

Biehler, R. (2006): Leitidee „Daten und Zufall“ in der didaktischen Konzeption und im Unterrichtsexperiment. In: J. Meyer (Hrsg.), *Anregungen zum Stochastikunterricht* Band 3, Tagungsband 2004/2005 des Arbeitskreises „Stochastik in der Schule“ (S. 111–142). Hildesheim: Franzbecker.

Biehler, R. & Schweynoch, S. (1999): Trends und Abweichungen von Trends. *mathematik lehren*, 97, 17–22.

Biehler, R., Prömmel, A. & Hofmann, T. (2007): Optimales Papierfalten – Ein Beispiel zum Thema „Funktionen und Daten“. *Der Mathematikunterricht*, 53(3), 23–32.

Borovcnik, M. (2005*)*. Probabilistic and statistical thinking. URL: http://www.ethikkommission-kaernten.at/lesenswertes/Upload/CERME_Borovcnik_Thinking.pdf (Stand: 01.10.2009).

Eichler, A. & Vogel, M. (2009). Die Leitidee Daten und Zufall. Wiesbaden: Vieweg+Teubner

Engel, J. (1998). Zur stochastischen Modellierung funktionaler Abhängigkeiten: Konzepte, Postulate, Fundamentale Ideen. *Mathematische Semesterberichte*, 45, 95–112.

Engel, J. & Vogel, M. (2006). Funktionen in einer Welt voller Daten: Vernetzungen zwischen Stochastik, Algebra und Analysis. In: J. Meyer (Hrsg.), *Anregungen zum Stochastikunterricht* Band 3, Tagungsband 2004/2005 des Arbeitskreises „Stochastik in der Schule“ (S. 159–171). Hildesheim: Franzbecker.

Erickson, T. (2005). The model shop. Using data to learn about elementary functions. Oakland: eeps media. (Third Field-test Draft)

Girwidz, R. & Vogel, M. (2007), Modellieren mit interaktiven Arbeitsblättern in Excel, zur Veröffentlichung bei: *Beiträge zur Frühjahrstagung des Fachverbandes Didaktik der Physik der Deutschen Physikalischen Gesellschaft* 2007 (in press)

Humenberger, J. & Reichel, H.-C. (1995). Fundamentale Ideen der Angewandten Mathematik und ihre Umsetzung im Unterricht. Mannheim: BI-Wissenschaftsverlag.

Stachowiak, H. (1973). Allgemeine Modelltheorie. Berlin: Springer.

Vogel, M. (2006). Mathematisieren funktionaler Zusammenhänge mit multimediabasierter Supplantation. Hildesheim: Franzbecker.

Vollrath, H.-J. (2003). Algebra in der Sekundarstufe. Heidelberg: Spektrum

Anschrift der Verfasser

Markus Vogel
Fakultät III – Mathematik,
Institut für Datenverarbeitung/Informatik
Pädagogische Hochschule Heidelberg
Im Neuenheimer Feld 561
69120 Heidelberg
vogel@ph-heidelberg.de

Andreas Eichler
Institut für Mathematik und Informatik
und ihre Didaktiken
Pädagogische Hochschule Freiburg
Kunzenweg 21
79119 Freiburg
andreas.eichler@ph-freiburg.de