Beiträge zur Hydrogeologie / 48/II / Seiten 171-188 / Graz 1997

Multiphase Regression-Verfahren zur Abflußkomponententrennung

Multiphase regression method for hydrograph separation

C. Bäck, K. Fuchs[1]), K. D. Wernecke[2]) & J. Fank[3])

Inhalt

1. Aufgabenstellung

In den meisten alpinen Kleineinzugsgebieten sind die detaillierten Abflußverhältnisse aufgrund fehlender Meßstellen mit kontinuierlichen Aufzeichnungen kaum oder nur sehr ungenügend erfaßt. Im Zuge von vielen Projekten mit wasserwirtschaftlicher Fragestellung (z. B. Bau von Kleinkraftwerken, Ressourcenschätzung für die künftige Trink-

[1]) Dipl.-Ing. C. Bäck & Dipl.-Ing. Dr. K. Fuchs, Institut für Angewandte Statistik und Systemanalyse, JOANNEUM RESEARCH Forschungsgesellschaft mbH, Steyrergasse 25a, A-8010 Graz.
[2]) Univ.-Doz. Dr. K.-D. Wernecke, Medizinische Fakultät Charieté der Humboldt Universität, Schuhmannstraße 20, 1040 Berlin.
[3]) Dr. J. Fank, Institut für Hydrogeologie und Geothermie, JOANNEUM RESEARCH Forschungsgesellschaft mbH, Elisabethstraße 16/II, A-8010 Graz.

wassernutzung) ist aber eine genaue Kenntnis vor allem der abflußfähigen Wassermengen nach längeren Trockenperioden vonnöten, um eine sinnvolle hydrologische Beurteilung durchführen zu können.

1.1. Untersuchungsgebiet

Jeglicher in einem Einzugsgebiet fallender Niederschlag fließt bei fehlendem Grundwasserabstrom am orographisch tiefsten Punkt aus diesem ab. Bei Übereinstimmung des orographischen mit dem tatsächlichen läßt sich das Einzugsgebiet aus der topographischen Karte anhand der Isohypsen oder aus einem digitalen Geländemodell abgrenzen. Figur 1 zeigt eine Teileinzugsgebietsgliederung und Meßstellenverteilung der oberen Pöllauer Safen (Österreich, Steiermark). Dies ist ein orographisch sehr gut abgrenzbares, abgeschlossenes, nur nach SE offenes Einzugsgebiet mit einer zentralen Entwässerung. Das Gesamteinzugsgebiet hat eine Fläche von 58,77 km^2, die Flächen der Teileinzugsgebiete liegen zwischen 0,14 km^2 und 5,89 km^2.

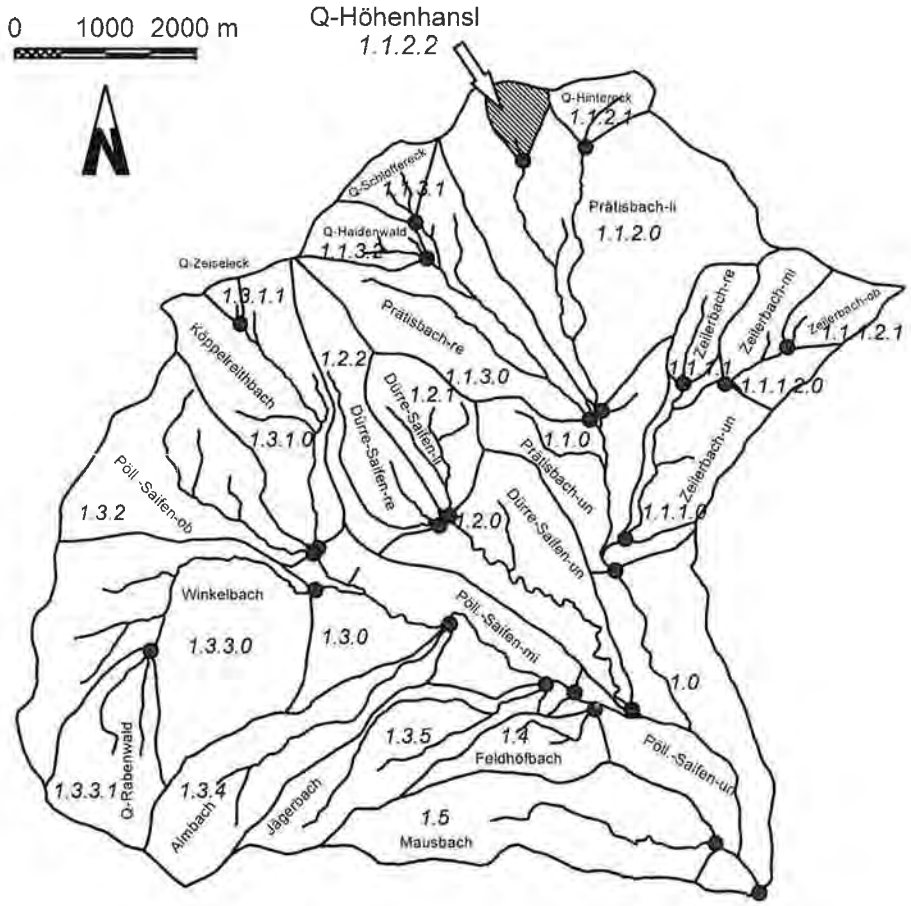

Fig. 1: Teileinzugsgebietsgliederung des Einzugsgebietes Pöllauer Safen oberhalb des Basispegels. Tributary catchments to the Pöllauer Safen watershed.

In der vorliegenden Arbeit ist das relativ kleine Teileinzugsgebiet Höhenhansl (vgl. Fig. 1, Einzugsgebiet 1.1.2.2) wegen der nahezu lückenlosen Erfassung von Abflußdaten der Jahre 1991–1996 Gegenstand der Untersuchungen. In diesem Einzugsgebiet sind die hydrologischen Randbedingungen aufgrund früherer Arbeiten (J. FANK et al., 1993) gut bekannt.

1.2. Vorherrschende Speichertypen

Die während eines Niederschlag-Abflußereignisses gesamt abfließende Wassermenge eines Einzugsgebietes besteht in der Regel aus mehreren Komponenten, welche unterschiedlichen Speichern entstammen. Die Analyse von Abfluß-Rückgängen (Rezessionen) erlaubt eine Auftrennung des Gesamtabflusses in Abflußkomponenten und zudem Rückschlüsse auf das Speichervermögen der einzelnen zugrundeliegenden Speicher. Bei Oberflächengerinnen lassen sich zumeist zwei oder drei unterschiedliche Abflußanteile voneinander trennen:

- **Basisabflußkomponente:** Diese Abflußkomponente entspricht dem Anteil des länger gespeicherten Wassers (Grundwasserkomponente). Es handelt sich dabei um jenen Abflußanteil, der im wesentlichen bei längeren Trockenperioden zum Abfluß gelangt.
- **Zwischenabflußkomponente:** Dieser Abflußanteil durchfließt nur kurzfristig die obersten Bodenschichten und erreicht nach kurzer Verweilzeit im Untergrund den Vorfluter.
- **Direktabflußkomponente:** Bei diesem Anteil handelt es sich um Wasser, das ohne längere Verzögerung nach einem Niederschlagsereignis (oder einsetzender Schneeschmelze) den Pegel erreicht.

1.3. Schätzung von Speicherkenngrößen

Der Verlauf der Abflußganglinie während einer Hochwasserwelle spiegelt neben dem auslösenden Niederschlags- oder Tauwetterereignis auch die geologischen, morphologischen, hydrologischen und klimatischen Eigenschaften des Einzugsgebietes wider. Der Verlauf des ansteigenden Teiles einer Hochwasserganglinie hängt vor allem von der Geometrie des Einzugsgebietes sowie der räumlichen und zeitlichen Verteilung des Niederschlages ab. Der fallende Rezessionsteil ist durch die Entleerung der zum Abfluß beisteuernden Speicher, das heißt v. a. durch deren hydrogeologische Eigenschaften bestimmt (Fig. 2).

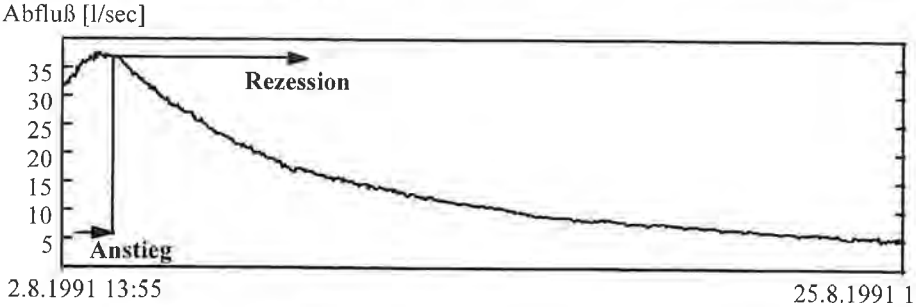

Fig. 2: Ganglinie des Abflusses an einem Pegel aus Datensammleraufzeichnungen.
Runoff hydrograph evaluated from a gauging station with automatic data recording.

Das erwartete Auslaufen eines einzelnen Speichers in Trockenwetterperioden kann durch die sogenannte Maillet Funktion (E. MAILLET, 1905) dargestellt werden:

$$E[\text{Abfluß}_t] = Q_t = q_o \exp(-\alpha t);\ q_o,\ \alpha > 0. \tag{1.1}$$

Der erwartete Abfluß Q_t zum Zeitpunkt t eines Speichers ergibt sich somit aus einem Abflußwert q_o zum Zeitpunkt t=0 und dem Auslaufkoeffizienten α welcher auch Leerlauf- oder Austrocknungskoeffizient genannt wird.

Während nach einer langen Trockenwetterperiode nur mehr Basisabfluß angenommen werden kann, setzt sich der Gesamtabfluß der Hochwasserwelle aus mehreren Abflußkomponenten zusammen. Interpretiert man den Rezessionsvorgang als Entleerung mehrerer Einzelspeicher, so kann man den Erwartungswert des Gesamtabflusses als Summe mehrerer Einzelspeicher modellieren (R. K. LINSLEY et al., 1975):

$$E[\text{Gesamtabfluß}_t] = \sum_i q_{oi} \exp(-\alpha_i t);\ q_{oi},\ \alpha_i > 0. \tag{1.2}$$

Grundlegend für die hydrogeologische Beurteilung eines Einzugsgebietes ist die Schätzung der Parameter q_{oi} und α_i des zusammengesetzten Speichermodells (1.2) aus dem zugehörigen nichtlinearen Regressionsansatz

$$\text{Gesamtabfluß}_t = \sum_i q_{oi} \exp(-\alpha_i t) + \varepsilon;\ q_{oi},\ \alpha_i > 0 \tag{1.3}$$

mit einem Fehlerterm ε, für den $E[\varepsilon]=0$ gilt. Zur Lösung dieses nichtlinearen Regressionsproblems steht eine Reihe iterativer Verfahren zur Verfügung, welche ausgehend von vorgegebenen Startwerten eine gegebene Zielfunktion minimieren. Durch die Nichtlinearität der Zielfunktion ist allerdings nicht gewährleistet, daß die berechnete Lösung einem globalen Optimum entspricht. Daher sollten die Startwerte möglichst nahe am globalen Optimum liegen.

In der Folge beschäftigt sich die vorliegende Arbeit mit Methoden zur Berechnung günstiger Startwerte und deren Einsatz anhand der aufgezeichneten Meßdatenreihen des Einzugsgebietes Höhenhansl. Besonderes Augenmerk wird dabei auf die Stabilität und Geschwindigkeit der Algorithmen gelegt, da die Meßreihen Umfänge bis zu 60 000 Meßpunkten erreichen und zudem systematische Modellverletzungen (z. B. Niederschlag, Autokorrelationen) zu erwarten sind. Da diese Modellverletzungen zumeist nicht die Schätzung aller Einzellinearspeicher beeinträchtigen, sollten die eingesetzten Verfahren zumindest insofern robust sein, daß die Parameter der ungestörten Speicherkomponenten des Modells (1.3) unverzerrt geschätzt werden.

2. Methodik zur Schätzung der Speicherkenngrößen

2.1. Nichtlineare Regression

Eine Vorlaufstudie zeigte, daß sich das Iterationsverfahren von Levenberg-Marquardt (N. R. DRAPER & H. SMITH, 1981) hinsichtlich der Stabilität und der Genauigkeit der errechneten Lösungen am besten zur Schätzung der Parameter q_{oi} und α_i des Modells (1.3) eignet.

Anhand einer Simulationsstudie wurden die Iterationsverfahren von Gauss, Gauss-Newton und Levenberg-Marquardt sowie das Steepest Descent Verfahren erprobt. Zudem wurden die Verfahren von R. G. CORNELL (1962) und von M. AGHA (1971) implementiert und untersucht. Die zuletzt genannten Algorithmen stellen eine Alter-

native zu den obigen Iterationsverfahren dar. Aufgrund der vorliegenden Datenumfänge und der häufig auftretenden Modellstörungen treten bei diesen nichtiterativen Verfahren allerdings sehr große numerische Ungenauigkeiten auf, wodurch die Gültigkeit der Lösungen nicht gewährleistet wird.

2.2. Startwertsuche – Startwertberechnung über Knickpunkte

Bisher wurden die für das Verfahren von Levenberg-Marquardt benötigten Startwerte visuell (manuell) aus den vorliegenden Meßdaten bestimmt. Die dazu erforderliche Vorgangsweise wird im folgenden anhand von Abflußdaten des Einzugsgebietes Pöllau, Pegel Höhenhansl erklärt. Figur 3 zeigt die Rohdaten, welche in Form von Abflußaufzeichnungen als geordnete Meßdatenreihe vorliegen.

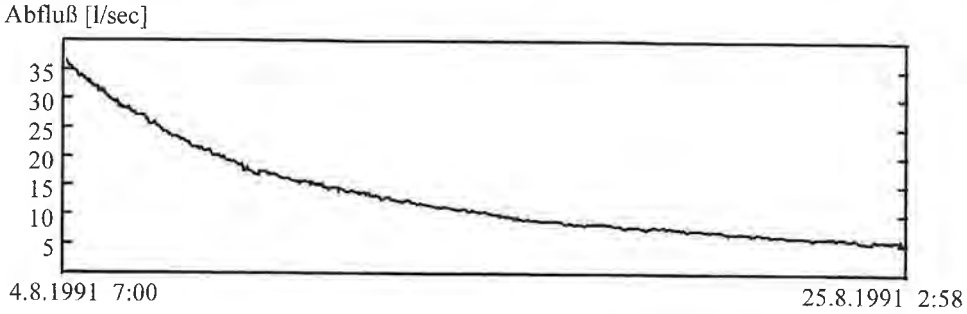

Fig. 3: Meßwertreihe der Rezession 910803.
Measured runoff data from the recession 910803.

Da die Verweilzeiten in den einzelnen Speichern Unterschiede aufweisen, ist der letzte Teil der Rezessionskurve nur mehr durch das Auslaufen eines einzigen Speichers charakterisiert. Daher kann dieser durch einen einzigen Exponentialterm $q_{o1} \exp(-\alpha_1 t)$ modelliert werden. Dies ist insbesondere aus Fig. 4, welche die logarithmierten Werte visualisiert, abzulesen. Der lineare Teil des zuletzt auslaufenden Speichers ist in Fig. 4

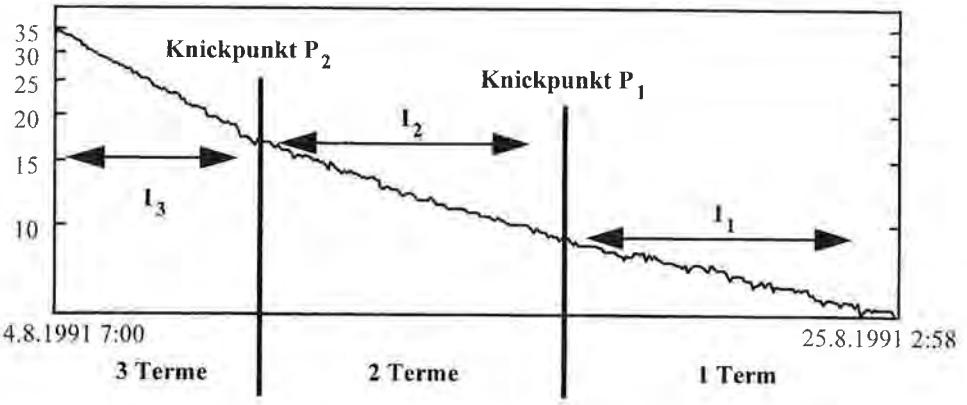

Fig. 4: Meßwertreihe der Rezession 910803, logarithmierte Darstellung.
Measured runoff data from the recession 910803, logarithmic.

durch das Intervall I_1, welches mit dem Knickpunkt P_1 beginnt und mit dem letzten Meßwert endet, gekennzeichnet. Aufgrund der unterschiedlichen Verweilzeiten in den einzelnen Speichern gibt es immer ein Intervall I_1, das nur durch das Auslaufen eines einzigen Speichers charakterisiert ist.

Durch den linearen Verlauf im Intervall I_1 können zunächst die Parameter q_{o1} und α_1 dieses Exponentialterms geschätzt werden (lineare Regression). Alle weiteren, noch zu bestimmenden Intervalle I_i (i=2,...,c) sind durch das gleichzeitige Auslaufen mehrerer Speicher, nämlich gerade i Speicher, charakterisiert. Die Endpunkte dieser Intervalle sind in der logarithmierten Darstellung in Form eines Knickes zu erkennen und bedeuten das Einsetzen des Auslaufens eines zusätzlichen Speichers. Eine analytische Auftrennung der Einzelspeicher ist für diese Intervalle nicht möglich, da ein lineares Verhalten der logarithmierten Meßreihe in den Intervallen I_i (i=2,...,c) nicht vorausgesetzt werden kann.

Eine gute Schätzung für die Parameter der restlichen Exponentialterme erhält man dennoch durch sukzessives Extrapolieren der bereits geschätzten Parameter $q_{o\,i-1}$ und α_{i-1} auf das gesamt untersuchte Intervall I zur Schätzung eines weiteren Parameterpaares q_{oi} und α_i. Dabei wird in jedem Schritt i (i=2,...,c) der Beitrag des bereits bekannten Exponentialterms $q_{o\,i-1} \exp(-\alpha_{i-1} t)$ vom Gesamtabfluß abgezogen. Dadurch erhält man ein weiteres Intervall I_i mit linearem Verlauf, aus welchem die Parameter q_{oi} und α_i geschätzt werden können.

Dieser Vorgang wird solange wiederholt, bis die disjunkten Intervalle I_i (i=1,...,c) das Gesamtintervall I abdecken.

Figur 4 zeigt eine Partition der Datenreihe in drei Intervalle (I_1, I_2, I_3), getrennt durch die beiden Knickpunkte P_1 und P_2. Die Anzahl (c–1) der gewählten Knickpunkte bestimmt die Anzahl c der Intervalle I_i und der Speicher (Exponentialterme) im zusammengesetzten Modell. Für das gewählte Beispiel wurden somit c=3 Exponentialterme angesetzt:

$$E[\text{Gesamtabfluß}_t] = q_{o1} \exp(-\alpha_1 t) + q_{o2} \exp(-\alpha_2 t) + q_{o3} \exp(-\alpha_3 t);\ q_{oi},\ \alpha_i > 0. \qquad (2.1.)$$

Mit dieser Vorgangsweise können die Startwerte direkt aus den Daten berechnet werden, sobald die Knickpunkte in der logarithmierten Darstellung festgelegt sind. Die manuelle Bestimmung der Knickpunkte stellt allerdings ein sehr aufwendiges Verfahren dar, da die Meßreihen in der Regel Umfänge von 15 000 bis 60 000 Meßpunkten aufweisen. Zudem sind die Knickpunkte in der logarithmierten Darstellung visuell oft kaum wahrnehmbar. Folglich erfordert die Knickpunktsuche ein sehr hohes Fachwissen des bearbeitenden Hydrologen und führt daher in der Regel bei unterschiedlichen Bearbeitern auch zu unterschiedlichen Ergebnissen. Um die Festlegung der Knickpunkte zu objektivieren und die damit verbundenen zeitintensiven Tätigkeiten auf ein Minimum zu reduzieren, wurde ein Algorithmus zur automatischen Bestimmung der Knickpunkte entwickelt.

3. Knickpunkte – Modellierung mit Multiphase Regression-Modell

Die Schätzung der Parameter q_{oi} und α_i (i=1,...,c) des Modells 1.3 mit dem Verfahren von Levenberg-Marquardt setzt das Vorliegen günstiger Startwerte voraus. Solche Startwerte können über die Knickpunkte der logarithmierten Datenreihe mit Hilfe der in Abschnitt 2.2 vorgestellten Vorgangsweise berechnet werden. Obwohl in den

einzelnen Intervallen zwischen den Knickpunkten nicht unbedingt Linearität vorherrscht, ist es sinnvoll, ein stückweises lineares Modell anzusetzen und die Knickpunkte aus den Schnittpunkten der geschätzten Geraden zu berechnen. Der durch etwaige Nichtlinearität entstehende Fehler kann vernachlässigt werden, da die Zeitpunkte der Knicke nicht unbedingt mit jenen der vollständigen Entleerung der Speicher übereinstimmen. Zudem stellen die Knickpunkte keine endgültige Lösung dar, sondern dienen der Berechnung einer günstigen Ausgangsposition für die Optimierung nach Levenberg-Marquardt.

3.1. Multiphase Regression

Lineare Beziehungen zwischen einer Zielgröße y und Einflußgrößen $X=(x_1,...,x_p)$ lassen sich durch das allgemein bekannte lineare Modell $E[y] = X\beta$ darstellen, wobei β den gesuchten Parametervektor darstellt.

Liegen in unterschiedlichen Intervallen I_i (i=1,...,c) der Regressoren $x_1,...,x_p$ unterschiedliche lineare Beziehungen vor, so spricht man von einer mehrphasigen Regression (Multiphase Regression):

$$E[y] = \begin{cases} X\beta_1 & X \in I_1 \\ X\beta_2 & X \in I_2 \\ \vdots & \vdots \\ X\beta_c & X \in I_c \end{cases} \qquad (3.1)$$

Für jedes Intervall I_i wird in (3.1) ein eigenes lineares Modell angesetzt, wobei die einzelnen Modelle durch sogenannte Knickpunkte (Changepoints) voneinander getrennt sind und β_i den Parametervektor für das Modell i im Intervall I_i (i=1,...,c) darstellt. Für die logarithmierten Abflußdaten (vgl. Fig. 4) wird daher das Modell

$$y_k = x_k\beta_{i1} + \beta_{i2} + \varepsilon \quad \text{mit} \quad x_k \in I_i;\ k \in \{1,...,n\};\ i = 1,...,c \qquad (3.2)$$

angesetzt. Dabei stellt $x=(x_1,...,x_n)^T$ den Vektor der Zeitpunkte dar, zu welchen die Abflußwerte, die im Vektor $y=(y_1,...,y_n)^T$ enthalten sind, gemessen wurden.

Zur Schätzung der Parametervektoren β_i (i=1,...,c) schlagen G. A. F. SEBER & C. J. WILD (1989) vor, bei einer bekannten Anzahl von Changepoints die Modellparameter β_{i1}, β_{i2} (i=1,...,c) für alle möglichen Kombinationen von Changepoints zu schätzen und schließlich jenen Parametersatz als Lösung zu wählen, welcher den größten Likelihood (größte Wahrscheinlichkeit) hat. Dieser Ansatz ist vor allem bei kleinen Datenmengen vorteilhaft, im gegenständlichen Fall aufgrund der großen Datenumfänge jedoch nicht einsetzbar.

Auch R. E. QUAND (1972) geht von einer bekannten Anzahl von Changepoints aus. Die vorgeschlagene Methode setzt weiters voraus, daß die Fehlerterme ε unabhängig und normalverteilt sind. Durch das Einführen eines zusätzlichen Parametervektors $(\lambda_1,...,\lambda_c)$ mit $\Sigma\lambda_{i=1}$ wird die Wahrscheinlichkeit des Zutreffens des Modells i beschrieben. Die Likelihoodfunktion der Stichprobe enthält neben den Parametervektoren β_i zusätzlich die unbekannten Wahrscheinlichkeiten λ_i. Durch die Maximierung der logarithmierten Likelihoodfunktion erhält man die gesuchten Parametervektoren β_i, und somit die vorliegenden Changepoints. Das Maximieren der Likelihoodfunktion stellt allerdings wieder ein nichtlineares Problem dar, das lediglich iterativ gelöst werden kann und dessen Startwerte nicht leichter als im ursprünglichen Problem bestimmt werden können.

3.2. Clustering mit Fuzzy c-Means

Die Methoden der unscharfen Clusteranalyse basieren auf der Minimierung von Gütekriterien, wobei das zugrundeliegende Optimierungsproblem iterativ gelöst wird. Daher zählen diese Verfahren auch zu den iterativen Clustering-Methoden. Die Optimierung der Gütefunktionale wird vor allem dadurch erleichtert, daß diese für unscharfe Partitionen stetige, differenzierbare Funktionen darstellen.

Weiters ist durch die Kompaktheit des Raumes aller unscharfen Partitionsmatrizen die Existenz von Minima und Maxima gesichert. Dies ermöglicht die Entwicklung von effizienten Iterationsverfahren und zudem Aussagen über deren Konvergenzeigenschaften.

Ist $X=\{x_1,...,x_n\}$ eine Menge von n Objekten aus dem R^P, c eine ganze Zahl mit $1<c<n$ und $V_{(c\times n)}$ der übliche Vektorraum aller reellen (c×n) Matrizen, dann läßt sich die Menge aller scharfen Partitionen von X in c nichtleere Klassen (c-Partitionen) als die Menge P_{cn} aller (c×n) Matrizen $U=(u_{ik})\in V_{(c\times n)}$ darstellen (J. C. BEZDEK, 1981), für die gilt:

$$P_{cn} = \{U \in V_{(c\times n)} \mid u_{ik} \in \{0,1\} \ \forall i,k; \ \Sigma_{i=1}^{c} u_{ik} = 1 \ \forall k; \ 0 < \Sigma_{k=1}^{n} u_{ik} < n \ \forall i\}. \quad (3.3)$$

Scharfe Partitionen sind dadurch gekennzeichnet, daß jedes Objekt x_k eindeutig einer Klasse C_i zugeordnet wird. Im Gegensatz dazu beinhalten unscharfe Partitionen detaillierte Information über die Ähnlichkeit eines Objektes x_k zu allen Klassen C_i $(1\le i\le c)$. Dies geschieht in Form von Gewichten u_{ik}, die den Grad der Zugehörigkeit eines Objektes x_k zu jeder der Klassen C_i angeben. Für die Menge P_{fcn} aller (c×n) Matrizen $U=(u_{ik})$, welche unscharfe Partitionen darstellen, gilt dann:

$$P_{fcn} = \{U \in V_{(c\times n)} \mid u_{ik} \in [0,1] \ \forall i,k; \ \Sigma_{i=1}^{c} u_{i=ik} = 1 \ \forall k; \ 0 < \Sigma_{k=1}^{n} u_{ik} < n \ \forall i\}. \quad (3.4)$$

Die Erweiterung auf unscharfe Partitionen bedeutet also einen Informationsgewinn gegenüber scharfen Partitionen.

Um möglichst kompakte Klassen zu erreichen, setzt J. C. BEZDEK (1981) im Fuzzy c-Means Algorithmus folgendes Zielkriterium für unscharfe Partitionen $U \in P_{fcn}$ ein:

$$J_m(U,v) = \Sigma_{k=1}^{n}\Sigma_{i=1}^{c} (u_{ik})^m \, d(x_k, v_i), \ 1 < m < \infty. \quad (3.5)$$

Die Gewichte u_{ik} entsprechen den Einträgen in der unscharfen Partitionsmatrix U, v_i stellt das Klassenzentrum der Klasse C_i dar, $d(x_k,v_i)$ ist ein zu definierendes Distanzmaß. Die Minimierung des Funktionals $J_m(U,v)$ liefert gut getrennte Klasseneinteilungen und repräsentative Klassenzentren v_i (siehe R. J. HATHAWAY & J. C. BEZDEK, 1988). Der Wichtungsparameter m steuert dabei die Schärfe der gesuchten Partition. Mit m=1 und der Einschränkung von U auf scharfe Partitionen reduziert sich die Minimierung von J_m auf den ISODATA Algorithmus (R. DUDA & P. HART, 1973) für scharfe Partitionen. Mit wachsendem m und unscharfen Partitionen U wächst auch die Unschärfe der berechneten Partitionen.

Für ein lokales Optimum (U^*, v^*) des Optimierungsproblems $Min_{U,v}\{J_m(U,v)\}$ gilt (J. C. BEZDEK, 1981):
Für k mit: $d_{ik}^* > 0 \ \forall i,k$:

$$u_{ik}^* = \left(\Sigma_{i=1}^{c} \left(\frac{d_{ik}^*}{d_{jk}^*} \right)^{\frac{1}{m-1}} \right)^{-1} \quad \forall i, \ m \in (0,1), \quad (3.6a)$$

für k mit: $\exists i$ mit $d^*_{ik} = 0$:

$$u^*_{ik} \in [0,1] \text{ mit } \Sigma^c_{i=1} u^*_{ik} = 1 \text{ und } u_{ik} = 0 \text{ falls } d^*_{ik} > 0 \qquad (3.6b)$$

mit einem geeigneten Distanzmaß: $d^*_{ik} = d(x_k, v^*_i)$.

Der Fuzzy c-Means Algorithmus baut auf den Bedingungen von (3.6) auf. Ausgehend von einer Startpartition $U^{(0)}$ und dem daraus berechneten Vektor $v^{(0)}$ der Klassenzentren, werden mit Formel (3.6) sukzessive neue Partitionen $U^{(r)}$ und daraus neue Klassenzentren $v^{(r)}$ berechnet, bis J_m $(U^{(r)},v^{(r)})$ einen stationären Punkt erreicht. Die Iterationen bilden somit eine Folge $\{(U^{(0)},v^{(0)}), (U^{(1)},v^{(1)}), \ldots, (U^{(r)},v^{(r)})\}$, wobei das Kriterium $J_m(U,v)$ in jedem Schritt reduziert wird. Zur Minimierung des Kriteriums J_m (U, v) wird nun eine Fixpunktiteration, das unscharfe Analogon zum ISODATA Algorithmus, vorgestellt (J. C. BEZDEK, 1981).

Fuzzy c-Means Algorithmus (FCM)

(1) Im ersten Schritt müssen folgende Parameter festgelegt werden:
die Klassenanzahl c, $1<c<n$,
der Wichtungsparameter m, $1 \leq m < \infty$,
eine geeignetes Distanzmaß $d(x_k,v_i)$,
eine geeignete Matrixnorm $\|.\|$ und eine Schranke ε als Abbruchkriterium,
eine Startpartition $U^{(r)}$ $(r = 0)$.
(2) Berechnen der Klassenzentren $\{v_i^{(r)}\}$, $1 \leq i \leq c$ aus $U^{(r)}$.
(3) Aktualisieren von $U^{(r)}$ durch (3.6) und $v_i^{(r)}$ zu $U^{(r+1)}$.
(4) Vergleich von $U^{(r)}$ und $U^{(r+1)}$ mit einer geeigneten Matrixnorm:
Stop, wenn gilt: $\|U^{(r+1)} - U^{(r)}\| \leq \varepsilon$,
sonst: $r := r+1$, und weiter mit Schritt (2).

3.3. Berechnung der Knickpunkte mit Fuzzy Clustering

Ausgangspunkt für den folgenden Algorithmus (FCRM) stellt der bereits in Abschnitt 3.2. vorgestellte Fuzzy c-Means Algorithmus (FCM) dar. Das eingesetzte Verfahren berechnet eine vorgegebene Anzahl von Knickpunkten aus den Meßwertpaaren $\{(x_1,y_1),\ldots, (x_n,y_n)\}$, aus welchen die Startwerte für Levenberg-Marquardt direkt berechnet werden können (vgl. Abschnitt 2.2.). Dabei entsprechen die Skalare x_k (k=1,...,n) des Vektors x den Meßzeitpunkten, die Skalare y_k (k=1,...,n) des Vektors y den logarithmierten Meßwerten zu diskreten Zeitpunkten.

Die Klassenzentren v_i werden durch die einzelnen linearen Modelle des Multiphase Regression-Modells (3.1) repräsentiert, die Zuordnung der Meßwertpaare (x_k,y_k) zu den einzelnen Klassen erfolgt über eine unscharfe Partitionsmatrix $U=(u_{ik}) \in P_{fcn}$.

Figur 5 zeigt die Startpartition, wie sie für die logarithmierten Daten der Rezession 910803 (Fig. 4) berechnet wurde (Abschnitt 3.3.1.). Gesucht ist eine Auftrennung der Datenreihe in drei lineare Modelle, gleichbedeutend mit der Bildung von drei Klassen. Daher ist für jede Klasse eine Zugehörigkeitsfunktion in der Graphik dargestellt. In horizontaler Richtung sind die Meßzeitpunkte aufgetragen, in vertikaler Richtung jeweils die Zugehörigkeitswerte einer Klasse, welche bei unscharfen Partitionen Werte im Intervall [0,1] annehmen können. Dieser Wertevorrat für die einzelnen Zugehörigkeitsfunktionen wird jedoch durch die Berechnung von scharfen Startpartitionen nicht ausgeschöpft.

Ausgehend von der Startpartition werden in den einzelnen Iterationsschritten des

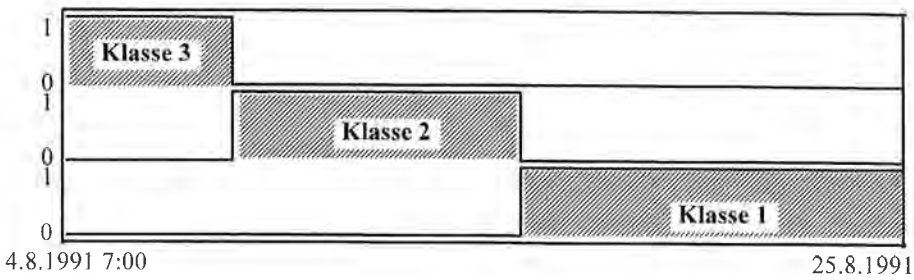

Fig. 5: Startpartition für die logarithmierte Meßwertreihe der Rezession 910803.
Starting partition for the logarithmic data of the recession 910803.

Algorithmus zunächst neue Klassenzentren berechnet. Als Klassenzentren v_i werden im vorliegenden Fall die Geraden des Multiphase Regression-Modells eingesetzt (vgl. Formel 3.2). Diese Geraden v_i lassen sich durch die Parametervektoren β_i (i=1,...,c) darstellen, wobei $\beta_i=\beta_i^{(r)}$ in jedem Iterationsschritt r mit einer durch die Zugehörigkeitswerte gewichteten linearen Regression geschätzt wird:

$$b_i^{(r)} = [x^T G_i^{(r)} x]^{-1} x^T G_i^{(r)} y. \qquad (3.7)$$

Dabei ist $G_i^{(r)}$ eine (n×n) Diagonalmatrix, deren Einträge der Zeile u_i der Partitionsmatrix $U^{(r)}$ entsprechen. Die Berechnung der Regressionsgeraden für drei Klassen aus der Startpartition für die Rezession 910803 (Fig. 5) ist in Fig. 6 visualisiert. Die horizontale Achse stellt wieder die Zeitachse dar, die vertikale Richtung entspricht der Skala der logarithmierten Daten. Da die Regressionsgeraden mit den Gewichten einer scharfen Partitionsmatrix ($u_{ik}^{(r)} \in \{0,1\}$) berechnet wurden, läßt sich deren Lage über die Daten gut nachvollziehen. Die drei Datenabschnitte, welche für die Berechnung der einzelnen Regressionsgeraden verantwortlich sind, sind in der Graphik durch vertikale Linien getrennt und zusätzlich mit ihrer Klassenzugehörigkeit gekennzeichnet. Wie aus der Graphik ersichtlich, passen sich diese ersten Regressionsgeraden aufgrund der Wahl der Startpartition schon sehr gut an die Daten an.

Aus den Residuen der einzelnen Regressionen läßt sich im nächsten Schritt mit Formel (3.6) eine neue Partitionsmatrix berechnen, welche das Zielkriterium verbessert.

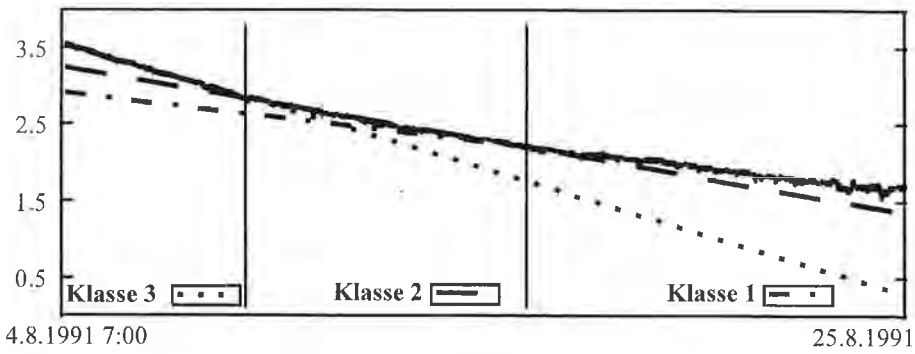

Fig. 6: Berechnung der Regressionsgeraden aus der Startpartition.
Calculation of the regression lines using the starting partition.

Dazu wird ein geeignetes Distanzmaß benötigt, welches durch die quadrierten Residuen definiert ist:

$$d_{ik} = (y_k - xb_i^{(r)})^2. \tag{3.8}$$

Dadurch ist gewährleistet, daß der Zugehörigkeitswert eines Datenpunktes für die Regressionsgerade (Klasse), welcher er am nächsten liegt, am größten ist. Durch den Einsatz von Formel (3.6) beschränkt sich die neu berechnete Partitionsmatrix nicht mehr auf scharfe Partitionen, wie auch in Fig. 7 ersichtlich.

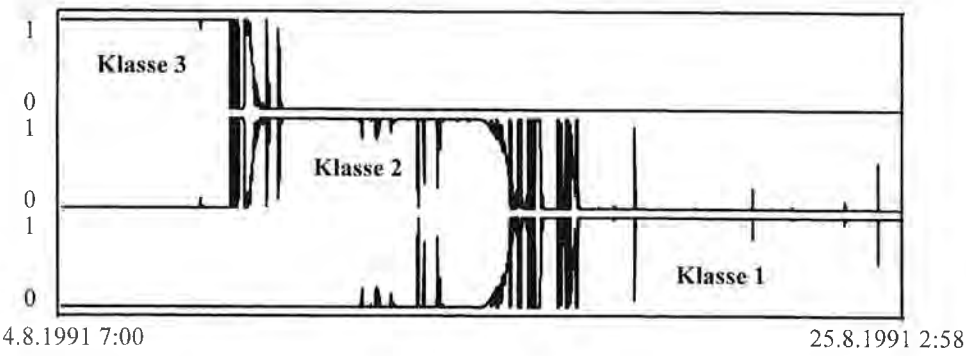

Fig. 7: Partition nach der ersten Iteration.
Partition after the first step of iteration.

Dieser Vorgang der Berechnung neuer Klassenzentren und Partitionen wird solange wiederholt, bis Stabilität der in aufeinanderfolgenden Iterationen berechneten Partition erreicht ist. Da infolge der großen Datenmengen durch das Abbruchkriterium über die Partitionsmatrizen $U^{(r)}$ und $U^{(r+1)}$ (vgl. FCM) die Konvergenz des Algorithmus nicht gesichert war, wurde das Abruchkriterium über die Regressionskoeffizienten $b^{(r)}$ und $b^{(r+1)}$ definiert:

$$\|b_i^{(r+1)} - b_i^{(r)}\|_{Max} < \varepsilon. \tag{3.9}$$

Die Ergebnisse des FCRM können über mehrere Wahlparameter gesteuert werden. Dazu zählen die Wahl des Distanzmaßes, Berechnung der Startpartition sowie der Wichtungsparameter m. Im vorliegenden Anwendungsfall konnten durch den Einsatz verschiedenster Distanzmaße keine wesentlichen Verbesserungen erzielt werden. Ausschlaggebend für zufriedenstellende Lösungen waren vielmehr die Berechnung einer günstigen Startpartition und die Wahl des Wichtungsparameters m.

3.3.1. Bestimmung der Startpartition

Die Auswahl einer geeigneten Startpartition ist ausschlaggebend für Konvergenzverhalten und -geschwindigkeit des eingesetzten Algorithmus. Gemäß des Multiphase Regression-Modells sollen Lösungen berechnet werden, welche die Meßdatenreihe in zusammenhängende Zeiträume teilt, getrennt durch die gesuchten Knickpunkte. Durch den FCRM selbst ist dieser Zusammenhang der einzelnen Zeiträume nicht unbedingt gewährleistet. Eine günstige Wahl der Startpartition kann einerseits die Auftrennung der Daten in nichtzusammenhängende Datenstücke verhindern, andererseits die Konvergenzgeschwindigkeit erheblich verbessern.

Wird die Startpartition über die Zeitachse berechnet, indem die Meßdatenreihe in gleich große Abschnitte geteilt wird, so kann ein kurzer hoher Peak der Rezession damit nicht entsprechend berücksichtigt werden.

Figur 8 zeigt das Ergebnis einer Klassifikation der Rezession 920914 bei Einsatz einer solchen Startpartition $S=(s_{ik})$ mit:

$$s_{ik} = \begin{cases} 1 & k \in [(i-1)\frac{n}{3}; i\frac{n}{3}] \; i = 1, \ldots; 3. \\ 0 & \text{sonst} \end{cases} \tag{3.10}$$

Eine Auftrennung in Einzellinearspeicher ist mit dieser Berechnung der Startpartition nicht möglich. Die Ursachen dafür liegen in der großen Varianz der Meßdatenreihe 920914 und in der Modellstörung durch Niederschläge im ersten Abschnitt der Rezession. Um dennoch brauchbare Ergebnisse zu erzielen, werden die Startpartitionen nicht über die Zeitachse, sondern über die Wertachse berechnet. Ein erster Ansatz dazu ist mit $S=(s_{ik}) \in P_{fcn}$

$$s_{ik} = \begin{cases} 1 & y(x_k) \in [(i-1)\frac{y_{max} - y_{min}}{3} + y_{min}; i\frac{y_{max} - y_{min}}{3} + y_{min}] \; i = 1, \ldots, 3 \\ 0 & \text{sonst} \end{cases} \tag{3.11}$$

in Fig. 9 dargestellt, welche einen Ausschnitt der über die Meßwerte berechneten Startpartition für die Rezession 920914 darstellt. Deutlich erkennbar bilden die Klassen 2 und 3 in diesem Fall keine disjunkten Zeitabschnitte. Der Algorithmus kann die Ein-

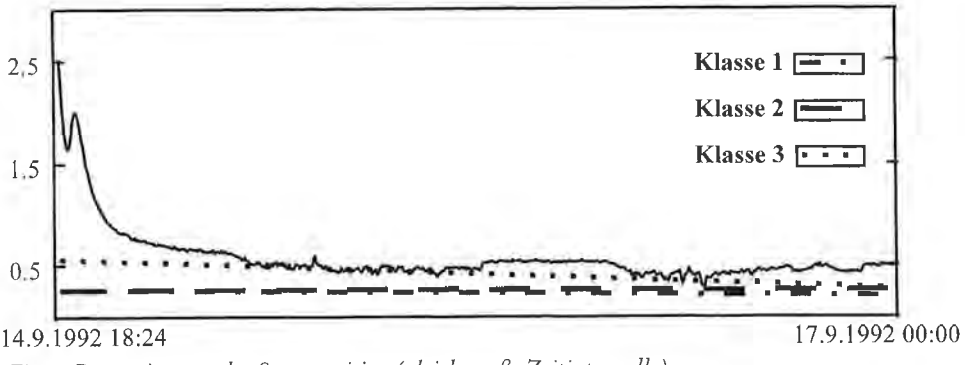

Fig. 8: Regression aus der Startpartition (gleich große Zeitintervalle).
Regression lines using a starting partition based on equally spaced time lags.

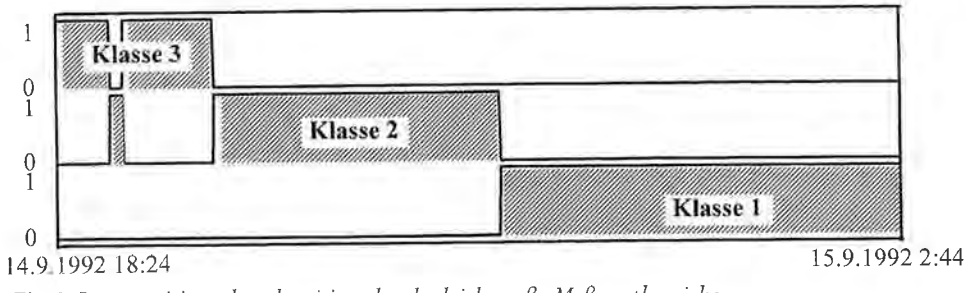

Fig. 9: Startpartition, charakterisiert durch gleich große Meßwertbereiche.
Starting partition based on the measured data.

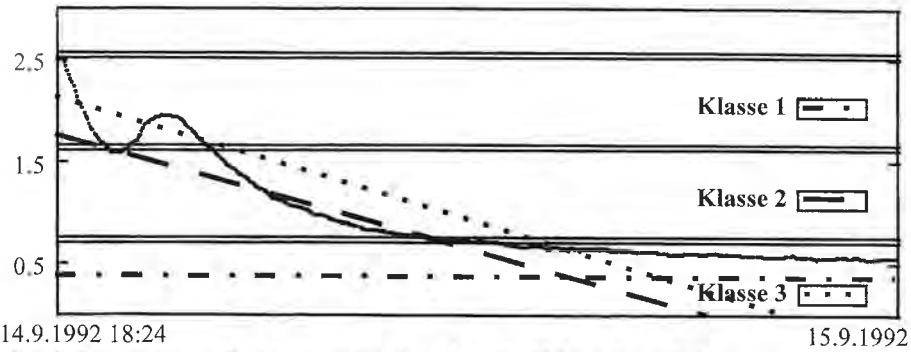

Fig.10: Regression aus der Startpartition (gleich große Meßwertbereiche).
Regression lines using a starting partition based on the measured data.

zelbereiche der Klassen aufgrund der Residuen nicht mehr zusammenführen und liefert somit keine verwertbaren Knickpunkte (Fig. 10).

Diese Schwierigkeiten können jedoch umgangen werden, wenn man unzusammenhängende Klassen schon in der Startpartition zusammenführt, wie in Fig. 11 dargestellt. Dabei wird ausgehend von der Startpartition in Fig. 9 der Klasse 3 nur noch das steilste Datenstück zugeordnet, um diesen Teil der Rezession bestmöglich zu modellieren. Das Ergebnis in Fig. 12 zeigt, daß mit Hilfe dieser Strategie auch bei durch kleine Niederschlagsereignisse gestörten Datenreihen und starker Variation Knickpunkte geschätzt werden können.

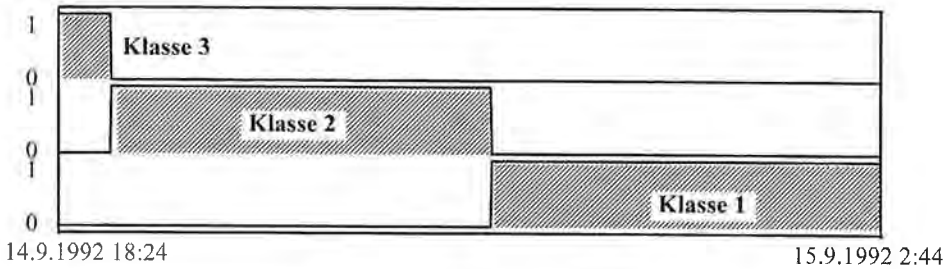

Fig.11: Startpartition, charakterisiert durch ungleich große Meßwertbereiche.
Starting partition based on the measured data, modified.

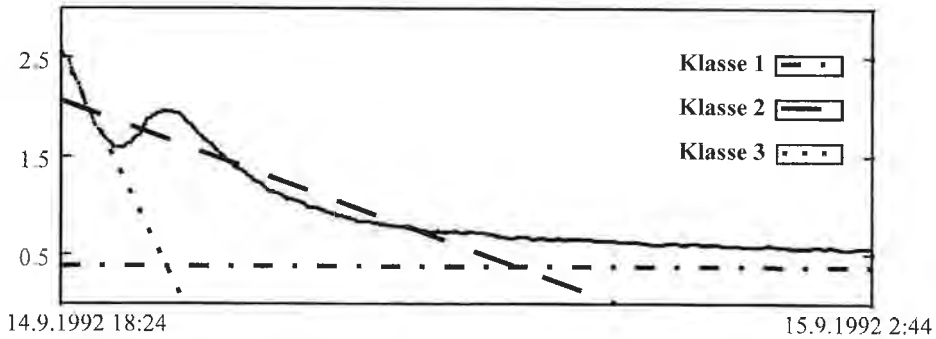

Fig.12: Regression aus der Startpartition (ungleich große Meßwertbereiche).
Regression lines using the modified starting partition based on the measured data.

3.3.2. Auswirkungen des Wahlparameters m

Neben der Voraussetzung einer günstigen Ausgangsposition für den FCRM-Algorithmus in Form einer geschickt gewählten Startpartition hat auch der in Formel (3.6a) eingesetzte Wichtungsparameter m eine wesentliche Wirkung auf die Ergebnisse des FCRM. Dieser Parameter steuert, in welcher Form die Residuen zu den einzelnen Regressionsgeraden in Zugehörigkeitswerte umgesetzt werden. Ein Wichtungsparameter von m=1,1 bis m=1,8 bewirkt, daß die in den Iterationsschritten berechneten Partitionen scharfen Partitionen sehr nahekommen, das heißt, jeder Punkt wird bevorzugt einer Klasse zugeordnet. Mit wachsendem m (m>1,8) sinkt die Schärfe der berechneten Partitionen und somit die bevorzugte Zuordnung zu einer einzigen Klasse. Eine solche unscharfe Partition zeigt Fig. 13. Eine klare Zuordnung der einzelnen Meßwerte zu zwei Klassen ist über weite Abschnitte nicht möglich.

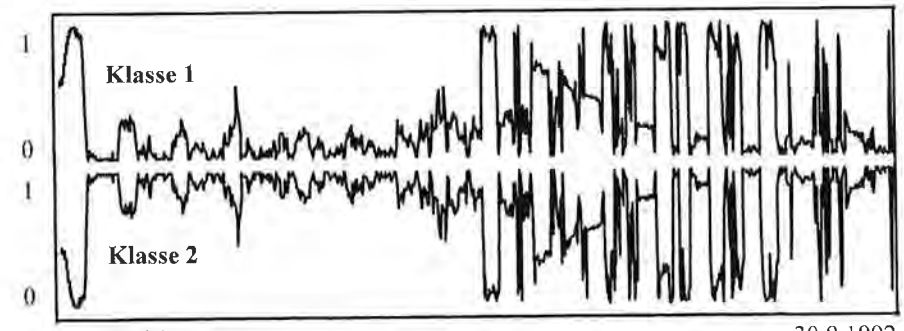

Fig.13: Ergebnispartition für die Rezession 920914 (m=2).
Resulting partition for the recession 920914, m=2.

Infolge der großen Variationen der Meßdatenreihe 920914 und der zu großen Wahl von m spalten sich die in der Startpartition noch zusammenhängenden Klassen sukzessive in nicht zusammenhängende Teilintervalle auf, bis schließlich beide Regressionsgeraden lediglich die Varianz in den Daten erklären und die Knickpunkte nicht mehr wie erwünscht darstellen (vgl. Fig. 14).

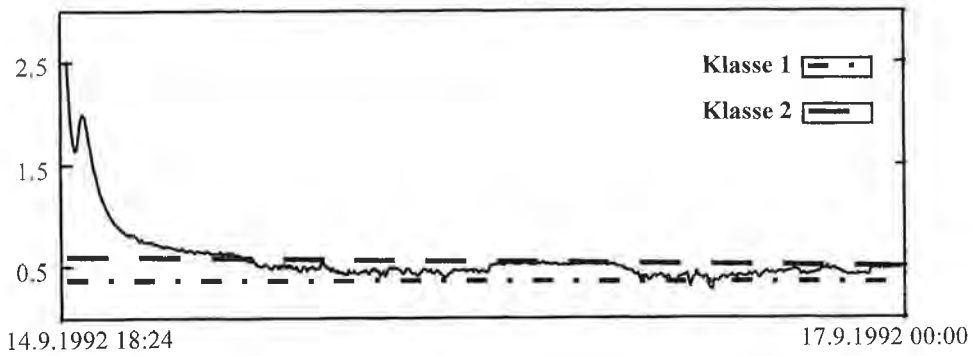

Fig.14: Ergebnis für die Rezession 920914 (m=2).
Resulting regression lines for the recession 920914, m=2.

4. Ergebnisse und Diskussion

Die Anwendung des FCRM auf Abflußdaten liefert im allgemeinen sehr gute Schätzungen für die gesuchten Knickpunkte, selbst dann, wenn die Modellvoraussetzungen durch Einsetzen von Niederschlag oder Wasserentnahmen in der Rezessionsphase verletzt sind.

Die Lösungen für die Rezessionen 910803 und 920914 sind in Fig. 15 bis Fig. 18 zusammengefaßt, Eckdaten zur Berechnung sind Tab. 1 zu entnehmen.

Die Ergebnisse zeigen deutlich die Robustheit der Methode gegenüber systematischen periodischen Einflüssen im Baseflow Bereich. Zudem ist die klare Auftrennung der Rezessionen in einzelne Abschnitte, getrennt durch Knickpunkte deutlich erkennbar.

Über die Knickpunkte gelangt man zur Ausgangslösung für das Verfahren nach Levenberg-Marquardt. Die Lösung dieses Verfahrens für die Rezession 920914 ist in Tab. 2 tabellarisch und in Fig. 19 graphisch visualisiert.

Aus hydrologischer Sicht steht allerdings nicht nur eine möglichst gute Anpassung des Modells 1.3 an die Daten im Vordergrund, sondern auch eine Reihe von Speicherkenngrößen, welche sich mit den geschätzten Parametern der Einzellinearspeicher berechnen lassen. Tabelle 3 zeigt die wichtigsten Kenngrößen, welche für die Rezession

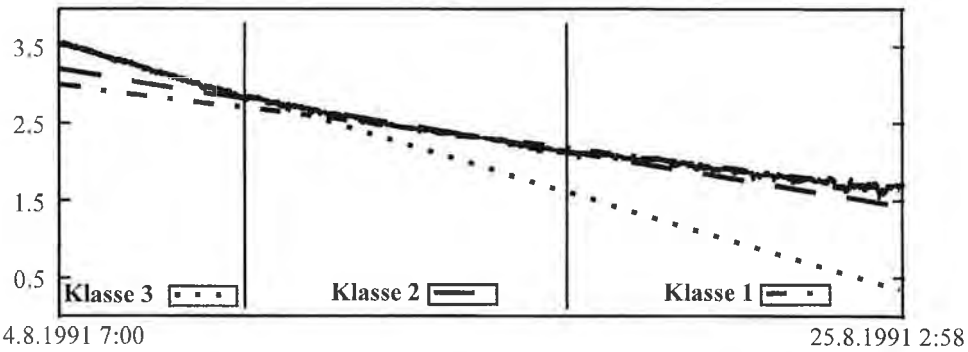

Fig.15: Ergebnisse des FCRM (Rezession 910803).
Resulting regression lines for the recession 910803.

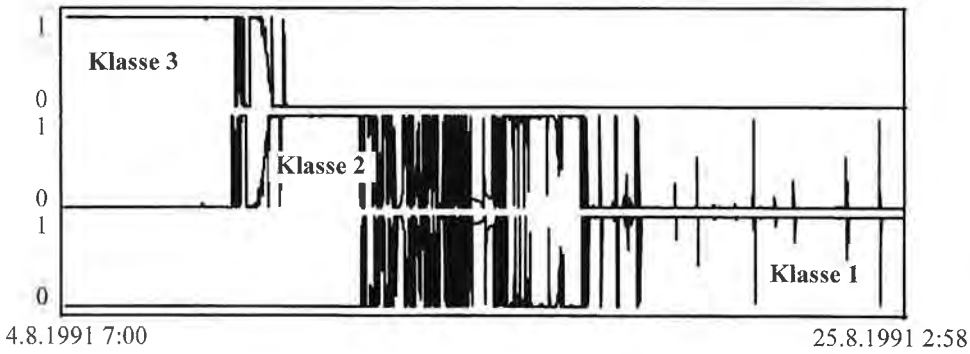

Fig.16: Ergebnispartition des FCRM (Rezession 910803).
Resulting partition for the recession 910803.

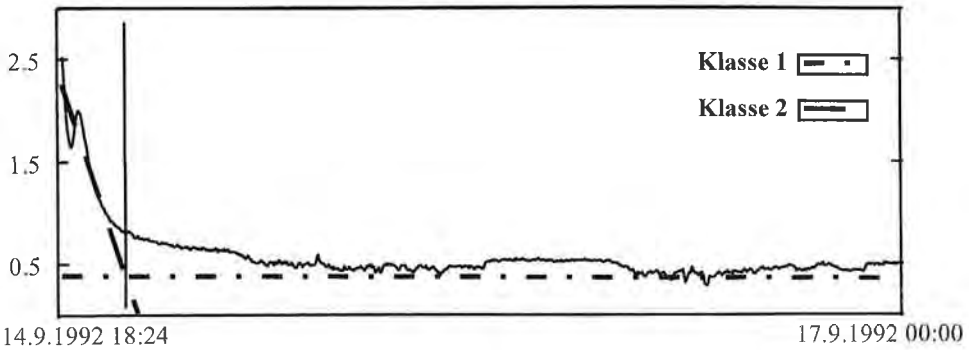

Fig.17: Ergebnisse des FCRM (Rezession 920914).
Resulting regression lines for the recession 920914.

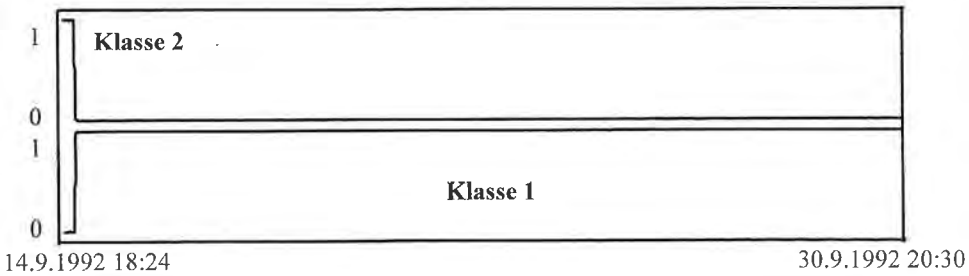

Fig.18: Ergebnispartition des FCRM (Rezession 920914).
Resulting partition for the recession 920914.

Tab. 1: Eckdaten für FCRM.
Parameters to be used by FCRM.

	Rezession 910803	Rezession 920914
Anzahl der Datenpunkte	32972	23166
Wichtungsparameter m	1,3	1,3
Abbruchkriterium ε	0,001	0,001
Anzahl benötigter Iterationen	13	13
Benötigte Zeit	~ 4 min	~ 2 min

Tab. 2: Eckdaten für Levenberg-Marquardt für die Rezession 920914.
Characteristic data for Levenberg-Marquardt for the recession 920914.

Ereignis	Datum, Zeit	Q [l/s]
Beginn	14.09.1992 17:40	1,69
Peak	14.09.1992 18:24	12,30
Ende	30.09.1992 20:30	1,17

Anzahl der Messungen	23166
Rezessionsdauer	16,1
Anzahl der Speicher	2
Residuenquadratsumme	365,07

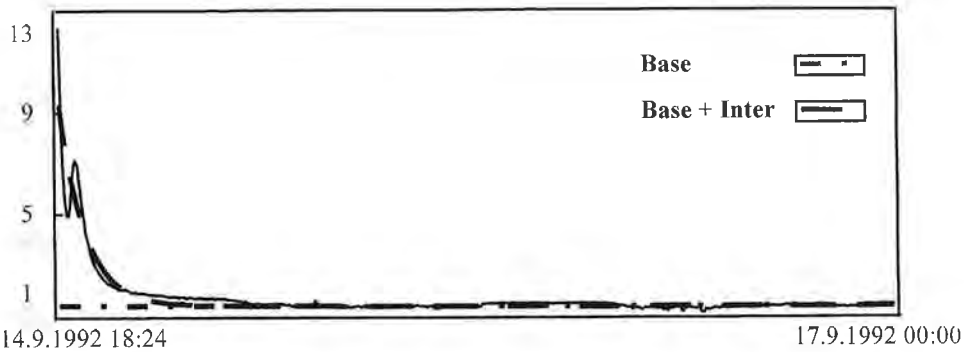

Fig.19: Ergebnisse von Levenberg-Marquardt für die Rezession 920914.
Results of Levenberg-Marquardt for the recession 920914.

920914 berechnet wurden. Die Kenngrößen sind darin nach Grundwasserabfluß (Baseflow), Zwischenabfluß (Interflow) und Gesamtabfluß getrennt. Neben den Schätzungen für q_{oi} (Q_{Peak}) und α_i (Auslaufkoeffizient) sind der Tabelle Angaben über das gesamt auslauffähige Volumen und die Auslaufzeit zu entnehmen.

Tab. 3: Speicherkenngrößen für die Rezession 920914.
Characteristic data of the recession 920914.

Ereignis 920914	Baseflow	Interflow	Baseflow + Interflow
Q_{Peak} [l/s]	1,46	8,10	9,56
Auslaufkoeff. α [d^{-1}]	–0,015	–15,165	
Volumen [m^3]	1800,76	45,65	1846,41
Volumen [%]	97,53	2,47	100,00
Zeit t mit $Q_t < 0,1$ l/s]	10. 03.1993 15:31	15. 09.1992 01:22	
Auslaufzeit [d]	176,8	0,3	

Zusammenfassung

Der Einsatz von automatischen Aufzeichnungssystemen mit kurzem Intervall für die Registrierung des Wasserstandes in Oberflächengewässern (im vorliegenden Fall Drucksonde und Datensammler; Intervall fünf Minuten) führt einerseits zu hervorragenden Datenbasen für die hydrologische Auswertung der Abflußrezessionen, andererseits zu extrem großen Datenmengen. Diese Datenmengen und der Wunsch nach objektiven Auswerteergebnissen erfordern die Entwicklung von mathematisch-statistischen Verfahren, welche die Eingangsdaten hydrologischer Modelle effizient und nachvollziehbar aus den Meßdaten ermitteln. Die vorliegende Arbeit zeigt dazu einen Ansatz, welcher die numerischen Anforderungen mit robusten Verfahren bewältigt und zudem eine Datenverarbeitung in einer für den praktischen Einsatz geeigneten Zeit ermöglicht.

Literatur

AGHA, M. (1971): A Direct Method for Fitting Linear Combinations of Exponentials.– Biometrics, Vol. **27**, 399–413, Washington (D.C.).

BÄCK, C., J. FANK, T. HARUM & M. HUSSAIN (1995): Erarbeitung von hydrogeologischen Speicherkenngrößen durch statistische Modellierung von Abflußrezessionskurven. Zwischenbericht 1. Projektjahr.– Unveröff. Bericht d. Inst. f. Angewandte Statistik und Systemanalyse und d. Inst. f. Geothermie und Hydrogeologie, JOANNEUM RESEARCH, Graz.

BEZDEK, J. C. (1981): Pattern Recognition with Fuzzy Objective Function Algorithms.– Plenum Press, New York.

CORNELL, R. G. (1962): A Method for Fitting Linear Combinations of Exponentials.– Biometrics, Vol. **18**, 104–113, Washington (D.C.).

DRAPER, N. R. & H. SMITH (1981): Applied Regression Analysis.– 709 S., New York (John Wiley & Sons).

DUDA, R. & P. HART (1973): Pattern Classification and Scene Analysis.– 390 S., New York (John Wiley & Sons).

FANK J., T. HARUM & H. STADLER (1993): Erfassung von Abflußvorgängen in kleinen Einzugsgebieten; Speicherverhalten kleiner Einzugsgebiete.– Unveröff. Bericht d. Inst. f. Geothermie und Hydrogeologie, JOANNEUM RESEARCH, Graz.

HATHAWAY, R. J. & J. C. BEZDEK (1988): Recent Convergence Results for the Fuzzy c-Means Clustering Algorithms.– Journal of Classification, Vol. **5**, 237–247, New York (Springer).

HATHAWAY, R. J. & J. C. BEZDEK (1993): Switching Regression Models and Fuzzy Clustering.– IEEE Transactions on Fuzzy Systems, Vol. **1**, No. **3**, 195–204, New York.

LINSLEY, R. K., M. A. KOHLER & J. L. H. PAULHUS (1975): Hydrology for Engineers.– New York (McGrawhill).

MAILLET, E. (1905): Mécanique et physique du globe. Essays d'hydraulique souterraine et fluviale.– Paris (Hermann).

QUAND, R. E. (1972): A New Approach to Estimating Switching Regressions.– Journal of the American Statistical Association, Vol. **67**, No. **338**, 306–310, Alexandria.

SEBER, G. A. F. & C. J. WILD (1981): Nonlinear Regression.– 768 S., New York (John Wiley & Sons).

Summary

To estimate the water ressources of a test area to guarantee the water supply it is necessary to know the discharge from the catchment area after long dry periods. The run off of an area normally consists of several components, which originate from different storage systems. So it is possible to interprete the run off as the discharge of several single storage systems. Its mean value E then mathematically can be expressed as a compounded model

$$E[\text{run off}]\beta_t = \sum_i q_{oi} \exp(-\alpha_i t);\ q_{oi}, \alpha_V > 0$$

depending on time t.

The parameters q_{oi} und α_i represent important hydrological parameters of the test area.

To estimate the seeked parameters q_{oi} and α_i of this nonlinear regression problem there exist several iterative procedures, f.e. the Levenberg-Marquart algorithm (N. R. DRAPER & H. SMITH, 1981). To get optimal estimators and quick convergence these iterative nonlinear regression procedures need realistic initial values. In this paper "good" initial values were calculated automatically with the help of multiphase regression models combined with fuzzy clustering algorithms (R. J. HATHAWAY & J. C. BEZDEK, 1993). The proposed procedure allows to control the calculations on the basis of the residuals.