Mustererkennung zur Klassifizierung verölter Wasser- und Vogelproben

Von John C. Ranger

1. Einleitung

Die in diesem Beitrag erläuterten Mineralöle werden in zwei Hauptgruppen zusammengefaßt: nämlich in die der Rohöle und in die der Produktöle. Rohöle wie Produktöle bestehen teilweise aus den gleichen chemischen Verbindungen bzw. Stoffklassen, wie z.B. Aromaten, Arkanen, Tripertenen, Phystan, Pristan u.a.m. Charakteristisch für jeden Öltyp sind die Verhältnisse der relativen Konzentrationen bzw. Stoffmengen gewisser oben erwähnter Verbindungen. Die in der jeweiligen Probe vorkommenden absoluten Konzentrationen der Verbindungen werden durch Fluoreszenz-, Gaschromatographie- und Massenspektrographie-Messungen, die als Muster (engl.: pattern) bezeichnet werden, ermittelt. Jeder in der Meßreihe vorkommende Meßwert entspricht einem im Muster enthaltenen Merkmal (engl.: feature). Im übrigen werden jene Muster, die irgendwie eine Zugehörigkeit zueinander aufweisen, in Gruppen bzw. Klassen (engl.: classes oder categories) zusammengefaßt. Die Muster lassen sich als Punkte in einem Koordinatensystem oder als Vektoren, d.h. n-Tupel der Form $x^* = (x_1, x_2, x_3, \ldots, x_n)$, veranschaulichen (siehe Abb. 1 und 2). Die Klassifizierung der als Feldproben zusammengefaßten unbekannten Wasser- und Vogelproben erfolgt dadurch, daß man diese Proben mit Vergleichs- bzw. Schiffsproben bekannten Ursprungs vergleicht. Je nach Anzahl der Vergleichsproben kann man ein nichtstatistisches Verfahren oder ein statistisches Verfahren zur Anwendung bringen. Bei einer relativ kleinen Anzahl von Vergleichsproben wende ich ein nichtstatistisches Verfahren, nämlich das im Hauptprogramm ARTHUR enthaltene Unterprogramm KNN (engl.: K-th Nearest Neighbor oder Potential Function Method) bzw. HIER (engl.: Hierarchial Clustering) an. Fallen große Mengen von Daten an, setze ich ein statistisches Verfahren ein, nämlich das Unterprogramm BAYES (engl.: Bayes Decision Theory). KNN klassifiziert eine gegebene Ölprobe mittels der am häufigsten in unmittelbarer Nähe vorkommenden Klasse. HIER bietet eine weitere Möglichkeit an, Mustergruppen (engl.: clusters) bzw. Klassen zu bestimmen. Es werden dort Mustermappen übereinander auf ein Protokoll gedruckt und miteinander auf verschiedenen Ähnlichkeits-Stufen über gedruckte Äste gebunden. Näheres über KNN und HIER findet man in ANDREWS (1972), DUEWER et al. (1975) und FREUND (1962), KILLEEN et al. (1976) und SACHS (1984). Ein Maß für die Ähnlichkeit zwei beliebiger Ölprobenmuster (x_j) und (x_l) läßt sich z.B. durch den Kehrwert des Euklidischen Abstandes $D_{j,l}$ im Merkmalraum mit n Dimensionen berechnen:

$$D_{j,l} = ((x_{1,l} - x_{1,j})^2 + (x_{2,l} - x_{2,j})^2 + \ldots + (x_{n,l} - x_{n,j})^2)^{1/2}.$$

In dem 2. bis zum 4. Abschnitt kommen mehrere mathematische Gebilde vor, die

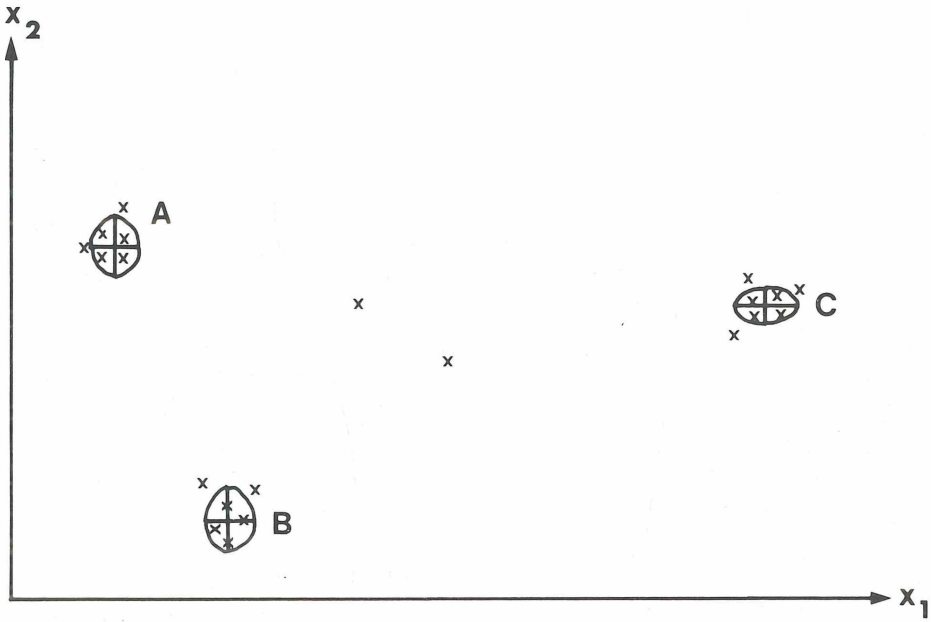

Abb. 1: Wasserproben-Situation
Die Ellipsen (bzw. Kreise) A, B und C stellen Schiffsproben, die Sterne Wasserproben dar (normiert).

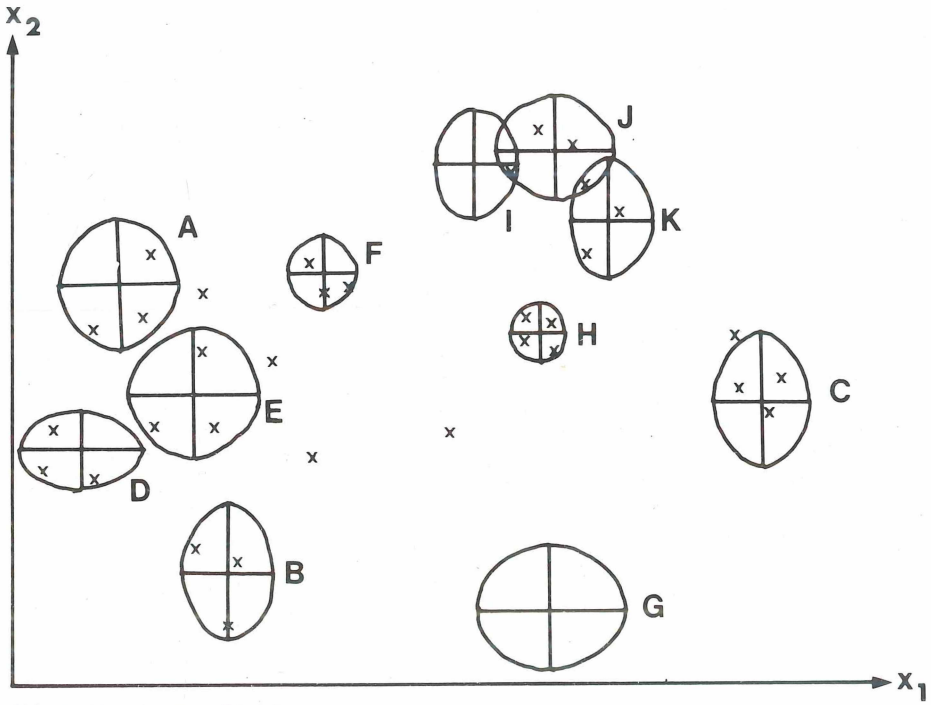

Abb. 2: Vogelproben-Situation
Die Ellipsen (bzw. Kreise) A bis K symbolisieren Vergleichsproben, die Sterne Vogelproben (normiert).

zunächst einmal einleitend erläutert werden müßten: Lateinische und griechische Kleinbuchstaben bezeichnen einzelne physikalische Meßgrößen – auch Skalare genannt – oder Funktionen, wie z. B. $x_{i,j,k}$ bzw. x_i, $g_{i,k}$, $w_{i,k}$, y_i und β_j. Rechts unten an den Kleinbuchstaben werden weitere Kleinbuchstaben, meistens die im lateinischen Alphabet auftretenden Buchstaben »i« bis »n«, angehängt. Diese Indizes dienen einer näheren Beschreibung der entsprechenden Meßgröße oder Funktion, wie z. B. das i.-Merkmal, das j.-Muster und die k.-Klasse, und nehmen als solche nur positive ganze Zahlen an. Es ist üblich, daß man sich auf die indizierten Meßgrößen auch als einzelne Meßwerte bzw. Merkmalwerte bezieht, obwohl sie, genaugenommen, als einzelne Variablen zu betrachten sind. Bei fehlendem Index symbolisiert ein Kleinbuchstabe eine vertikale – z. B. x – oder eine horizontale – z. B. $x^* = (x_1, x_2, x_3, \ldots, x_n)$ – Anordnung mehrerer solcher Größen. Eine derartige Anordnung von Meßgrößen wird Vektor genannt. Demzufolge kann man den Begriff Meßgröße als gleichbedeutend mit Mustermerkmal und den Begriff Vektor als gleichbedeutend mit Muster auffassen. Großbuchstaben – z. B. X und E_x – bezeichnen rechteckige, mit Spalten und Zeilen gekennzeichnete Anordnungen von Meßgrößen oder Funktionen und sind als Matrizen bekannt. Eine Matrix wird manchmal in der Kurzform »$(x_{i,j})$« geschrieben. In diesem Beitrag werden Vektoren bzw. Muster auch in der Form »(x_j)« angegeben. Anhand der Matrix-Addition und -Multiplikation lassen sich lineare Gleichungssysteme in der Form $y = X \cdot \beta + e$ darstellen. Dabei ist »X« eine Matrix, »y«, »β« und »e« sind Vektoren.

Meinen herzlichen Dank möchte ich an die folgenden Mitarbeiter im Deutschen Hydrographischen Institut und in der Fachgruppe »Seevögel und Robben«, Vogelwarte Helgoland, ausrichten: G. Dahlmann, P. Papisch, K. Jerzyki, E. Grün, W. Lange, G. Vauk, E. Vauk-Hentzelt, B. Reineking, B. Hansen, E. Hartwig und E. Schrey. Wegen des intensiven Gedankenaustausches danke ich P. Freimann und H. Theobald. Die Auswertung der gesammelten Daten erfolgte im Rahmen des UBA-Forschungsvorhabens Wasser 10204327 (Projektleiter: Dr. G. Vauk, Vogelwarte Helgoland).

2. Mathematisches Modell für Alterungsprozesse

Man stellt sich ein größeres Volumen einer Ölquelle, wie z. B. einen Bohrturm oder eine Raffinerie-Pipeline, einen Schiffstank usw., vor. Zweitens stellt man Überlegungen darüber an, wie wohl in einem relativ kurzen Zeitraum das Öl aus der Quelle ins freie Gewässer gelangen könnte. Drittens nimmt man an, daß der entstehende Ölfleck und eine oder mehrere verdächtige Ölquellen sofort aus der Luft von einem Beobachter im Hubschrauber gesichtet werden und daß, in wenigen Minuten, ein schnelles Boot an Ort und Stelle gelangt, damit man mehrere Quellen- und verölte Wasserproben nehmen kann. Ein solches Szenario stellt geradezu optimale Bedingungen für die Mustererkennung dar. Die zugehörigen Vergleichsproben- und Wasserprobenmuster lassen sich so, im wesentlichen, als feste Punkte im normierten Merkmalsraum – soweit keine spürbare Kontaminierung durch gealterte anthropogene oder biogene Kohlenwasserstoffe vorliegt – bestimmen. Solche Bedingungen, die ich hiernach als Wasserproben-Situation bezeichne, gewährleisten eine mit (sehr) hoher Wahrscheinlichkeit richtige und damit zuverlässige Klassifizierung der (unbekannten) Wasserproben. Auf der anderen Seite kommt es vor, daß eine (erheblich) längere Zeit (oder Entfernung) zwischen der Entstehung (oder Ort) des Ölflecks und der Probenahme bestehen kann. Die letztere nenne ich hiernach Vogelproben-Situation. Für solche Fälle betrachtet man die Vergleichsprobenmuster nicht mehr als feste Punkte, sondern als Punktewolken, deren Mittelpunkte ungefähr den bekannten Vergleichsproben entsprechen; denn über relativ längere Zeiten führen physikalisch-chemisch-biologische Einwirkungen dazu, daß die im Ölfleck enthaltenen Kohlenwasserstoff-Verbindungen abgebaut werden. Da diese Einwirkungen durch viele kleine, zeitlich und örtlich schwankende Einzelprozesse zu beschreiben sind, liegt es nahe, daß einerseits die gealterten Vergleichsmuster statistisch im Merkmalsraum eng um einen Mittelpunkt verteilt sind. Derartige kleine räumliche Abweichungen einer physikalischen Größe von ihrem Mittelpunkt sind in der statistischen Physik als Fluktuationen bekannt. In erster Näherung für die Häufigkeitsverteilung der Fluktuationen wird die Gauss-Verteilung verwendet (vgl. Landau und Lifschitz, 1979). Der oben erwähnte Mittelpunkt wird hiernach als $\bar{x}_{i,k}$ oder $x_{i,2,k}$ bezeichnet. Obwohl die Einzelprozesse deterministisch zu betrachten sind, läßt sich deren kollektives Verhalten unter quasi-gleichbleibenden physikalisch-chemisch-biologischen Zuständen andererseits auch statistisch-zeitlich erklären. Das wohl für die zeitliche Zersetzung sehr vieler Kohlenwasserstoff-Moleküle zulässige Analogon ist das in der Physik geltende radioaktive Zerfallsgesetz. Das Modell für die Alterungsprozesse sieht dann in etwa folgendermaßen aus: Man beginnt mit einem beliebigen Vergleichsprobenmuster $x^o_{i,1,k}$, d.i. das 1. Ausgangsmuster mit dem i.-Merkmal in der k.-Klasse. Nun erleidet die relative Konzentration $x^o_{i,1,k}$ eine kleine in der Zeit $t = t_1$ vollzogene negative Änderung, die zu dem Abbauprodukt mit der Konzentration $\bar{x}_{i,k} = x_{i,2,k}$ führt. Das auf diese Art und Weise entstandene Muster dient infolgedessen als Mittelpunkt $\bar{x}_{i,k}$ mehr oder minder stark gestreuter stochastisch-verteilter Alterungsmuster:

$$x_{i,j,k} = \bar{x}_{i,k} + \tilde{\delta} x_{i,j,k} \qquad (2.1)$$

mit $i = 1,2,3, \ldots, n_{Merkmal}$ und $j = 2,3,4, \ldots, n_{Muster} \cdot n_{Merkmal}$ bzw. n_{Muster} ist die Gesamtzahl der Merkmale bzw. Muster. Ähnliches gilt für k, wobei n_{Klasse} die im Klassifizierungsmodell vorkommende Gesamtzahl der bekannten Öltypen bezeichnet. Für das aus dem Ausgangsmuster entstandene Alterungsmuster gilt:

$$x_{i,2,k} = \bar{x}_{i,k} = x^o_{i,1,k} \cdot e^{-f_{i,k}(t)}. \qquad (2.2)$$

Hierzu nehme ich an, daß die Funktion $f_{i,k}(t)$ die folgende Form (vgl. Weidner et al., 1960) hat:

$$f_{i,k} = \alpha_{i,k} \cdot t, \text{ wobei} \qquad (2.3)$$

$\alpha_{i,k}$ der Abbaukoeffizient des i.-Merkmals in der k.-Klasse heißt. Die Zeit t hat z. B. die Einheit Stunde und ist die effektive Alterungsdauer. Gleichfalls ist $t_{1/2}$ die bekannte Halbwertzeit und definiert sich als die dafür notwendige effektive Alterungsdauer, in der die Konzentration $x^o_{i,1,k}$ um den Faktor 1/2 verringert wird. Der als physikalisch-chemisch-biologische Größe geltende Abbaukoeffizient $\alpha_{i,k}$ läßt sich provisorisch als Verhältnis zweier Funktionen, nämlich:

$$\alpha_{i,k} = \frac{g_i(p_{1k}, p_{2k}, p_{3k}, \ldots)}{c_i(q_{1k}, q_{2k}, q_{3k}, \ldots)} \qquad (2.4)$$

mit $g_i \geqq 0$ und $c_i > 0$, schreiben. Hier wird g_i als das theoretische Alterungsgewicht und c_i als der Stabilitätsmodul der i.-Kohlenwasserstoff-Verbindung bezeichnet. Demzufolge stellen die Parameter $p_{l,k}$ umweltabhängige – wie z. B. statischer Druck, Temperatur, von der Wellenlänge abhängige Lichtintensitäten, Konzentrationen der Reagenz-Verbindungen und Populationsdichten der Flora und Fauna – und die Parameter $q_{l,k}$ umweltunabhängige bzw. latente Größen dar, wie z. B. Molekularstrukturen, Löslichkeiten, Flüchtigkeiten, Affinitäten zur chemischen Transformation und Immunitäten bzw. Widerstandsfähigkeiten gegen biologische Aktivitäten.

Zum Zweck der Mustererkennung ist es von bedeutendem Interesse, die endlichen Wertebereiche der Koeffizienten $\alpha_{i,k}$ abzugrenzen. Hierfür ist es nützlich, eine um die oben erwähnte Punktewolke abgeschlossene Sicherheitskurve bzw. -fläche einzuführen. Auch wenn empirisch schwer durchführbar, ist es denkbar, daß man für einen Öltyp eine monoton steigende Zahlenfolge bestimmen kann, wofür gilt:

$$0 \leqq a < \alpha_{i,k} < b \qquad (2.5)$$

für alle möglichen oder relevanten theoretischen Alterungsgewichte $g_i = g_i(p_{1k}, p_{2k}, p_{3k}, \ldots)$. Bei dem Intervall (2.5) geht es darum, eine obere Grenze $\alpha^{max}_{i,k}$ derart zu bestimmen oder zu schätzen, daß man, bei denjenigen Kohlenwasserstoff-Verbindungen mit den größten Stabilitäten c_i, eine möglichst große Ausdehnung der obengenannten Punktewolke erzielt. Bei $n_{Merkmal} = 2$ im normierten Merkmalraum nimmt die Sicherheitskurve die Gestalt einer Ellipse mit den Halbachsen $(r_s)_{i,k}$ an. Für das Supremum $\alpha^{max}_{i,k}$ gilt dann:

$$\bar{x}^{min}_{i,k} = x^o_{i,1,k} \cdot e^{-\alpha^{max}_{i,k} \cdot t_c} \qquad (2.6)$$

mit $t_c >> 0$. $\bar{x}^{min}_{i,k}$ ist die in der Zeit $t = t_c$ abgebaute, minimal zu erwartende, mittlere Konzentration der i.-Verbindung des k.-Öltyps. Typisch müßte t_c einem Zeitraum von mehreren Wochen, Monaten oder sogar Jahren entsprechen. Die Halbachsen lassen sich einfach aus der Beziehung berechnen:

$$(r_s)_{i,k} = (x^o{}_{i,1,k} - \bar{x}^{min}_{i,k}) . \qquad (2.7)$$

3. Mathematisches Modell für quantitative Kontaminierungsbestimmung

Nehmen wir einmal an, daß man über ein aus mehreren Vergleichsproben- und zugehörigen Alterungsproben-Mustern bestehendes Klassifizierungsmodell, mit etwa $20 \leqq n_{Klasse} \leqq 150$, verfügt. n_{Klasse} ist, nach wie vor, die Anzahl der im vollständigen Klassifizierungsmodell vertretenen bekannten Öltypen. Zu diesem Datenbestand fügt man etwa 15 unbekannte Probenmuster hinzu, wie etwa in Abb. 1, und wendet das im ARTHUR befindliche Unterprogramm KARLOV (engl.: Karhunen-Loéve expansion) an, um sich eine mit den unbekannten Mustern ausgestattete Zeichnung (Eigenvektor-Projektion) des Klassifizierungsmodells in 2 Dimensionen zu verschaffen. Man stellt erfreulicherweise fest, daß die überwiegende Mehrheit der unbekannten Muster als Punkte innerhalb bzw. in unmittelbarer Nähe der verschiedenen Klassen zugeordneten Punktewolken (engl.: clusters) abgebildet wird. Jedoch findet man einige wenige, relativ weit von den Punktewolken entfernte unbekannte Punkte. Wie läßt sich das erklären? Die aus der Atmosphäre stammenden Einwirkungen hält man für vernachlässigbar klein. Darüber hinaus, weil man sich geeignete Mustermerkmale (vgl. FRIOCOURT et al., 1983) ausgesucht hat, darf man die Effekte von Flora und Fauna als ausgeschaltet betrachten. Da man von der im Abschnitt 2 erzählten Wasserproben-Situation ausgeht und aus stichhaltigen Gründen glaubt, daß es sich nicht um ausgelassene Vergleichsproben handelt, gelangt man zu der Schlußfolgerung: Die »Ausreißer« lassen sich nur mehrdeutig klassifizieren. Anschließend ruft man das Unterprogramm KNN und HIER auf, um sich davon zu überzeugen, daß es sich bei den »Ausreißern« um zwei oder mehrere Vergleichsproben-Kandidaten handelt.

Physikalisch gesehen darf man in einem solchen Falle den Verdacht äußern, daß die Wasserprobe aus einer oder mehreren Quellen stammt, welche aus zwei oder mehreren gemischten Ölsorten bestehen. Die Frage stellt sich, wie kann man eine solche Probe sowohl quantitativ als auch qualitativ klassifizieren? Die Antwort liegt wohl nahe, falls man die Ausreißer-Merkmale y_i (bzw. y) als Linearkombinationen verschiedener Vergleichsproben-Merkmale $\bar{x}_{i,k}$ bilden kann. Diese Beziehungen lassen sich als ein multilineares Regressionsmodell mit stochastischen Fehlern (engl.: perturbations oder errors) in der Form einer Matrix-Gleichung (vgl. DUDA et al., 1973, und GOLDBERGER, 1963) schreiben:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{bmatrix} = \begin{bmatrix} \bar{x}_{11} & \bar{x}_{12} & \bar{x}_{13} & \cdots & \bar{x}_{1n} \\ \bar{x}_{21} & \bar{x}_{22} & \bar{x}_{23} & \cdots & \bar{x}_{2n} \\ \bar{x}_{31} & \bar{x}_{32} & \bar{x}_{33} & \cdots & \bar{x}_{3n} \\ \cdot & \cdot & \cdot & \cdots & \\ \cdot & \cdot & \cdot & \cdots & \\ \cdot & \cdot & \cdot & \cdots & \\ \bar{x}_{m1} & \bar{x}_{m2} & \bar{x}_{m3} & \cdots & \bar{x}_{mn} \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \cdot \\ \cdot \\ \beta_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \cdot \\ \cdot \\ e_m \end{bmatrix} \qquad (3.1)$$

Die Matrix besteht aus n Spalten, mit $n = \tilde{n}_{Muster}$, und m Zeilen, mit $m = n_{Merkmal}$. $\tilde{n}_{Muster}$ weist hier auf eine geeignete Teilmenge aller Vergleichsprobenmuster hin. Die Spalten werden auch als Spaltenvektoren, die Zeilen als Zeilenvektoren bezeichnet. Demzufolge stellen die Spaltenvektoren n aus dem Klassifizierungsmodell befindliche (nicht-gealterte) Vergleichsprobenmuster dar. Der auf der linken Seite der Gleichung stehende Spaltenvektor y besteht aus Feldproben-Merkmalen, die einem Ölgemisch mit den entsprechenden prozentuellen Anteilen β_j zugeordnet werden. Der auf der rechten Seite stehende Spaltenvektor e beinhaltet überwiegend stochastische Fehler des Spaltenvektors y. Die stochastischen Fehler sind auf Alterungsprozesse in Ölen zurückzuführen, die den Vergleichsproben zuzuordnen sind. In verkürzter Form kann man ja Gleichung (3.1) auch so schreiben:

$$y = X \cdot \beta + e . \qquad (3.2)$$

Man geht davon aus, daß die Anzahl der Matrixzeilen größer oder gleich der Anzahl der Matrixspalten, also $m \geqq n$ ist. Ferner wird vorausgesetzt, daß die Produktmatrix $(X^T \cdot X)$ nicht-singular (det $(X^T \cdot X) \neq 0$) ist. Da die β_j Bruchteile eines Gemisches beschreiben, gilt die Beziehung:

$$\sum_{j=1}^{n} \beta_j = 1 . \qquad (3.3)$$

Analog zur Gleichung (2.1) gilt für statistisch-zeitlich verteilte (kleine) Abweichungen $\delta x_{i,k}$:

$$\bar{x}_{i,k} = x_{i,k} - \delta x_{i,k} . \qquad (3.4)$$

$\bar{x}_{i,k}$ ist nach wie vor das als Mittelwert im normierten Merkmalsraum aufzufassende Ausgangsmerkmal einer nicht-gealterten Vergleichsprobe. $x_{i,k}$ und $\delta x_{i,k}$ sind jeweils das entsprechende gealterte Merkmal und die dafür notwendige (stochastische) Änderung. Ähnlich läßt sich der Spaltenvektor y in zwei Summanden aufteilen, nämlich:

$$y_i = \bar{y}_i + \delta y_i . \qquad (3.5)$$

$\bar{y}_i$ spielt die parallele Rolle zum Merkmal $\bar{x}_{i,k}$ als Merkmal einer nicht-gealterten Ölgemischprobe. $\bar{y}_i$ und $\bar{x}_{i,k}$ liegen als bekannte Meßgrößen, also als Feld- und Vergleichsproben vor. Aus der Gleichung (3.1) bzw. der Gleichung (3.2) sowie aus den Beziehungen (3.4) und (3.5) erhält man die Matrixgleichungen:

$$\bar{y} = \bar{X} \cdot \beta \qquad (3.6)$$

$$\text{und } y = X \cdot \beta = (\bar{X} + E_x) \cdot \beta = \bar{X} \cdot \beta + e . \qquad (3.7)$$

Damit sind die stochastischen Abweichungen $e_i = \delta y_i$ als Linearkombinationen der stochastischen Abweichungen $\delta x_{i,k}$ zu verstehen:

$$e = E_x \cdot \beta . \qquad (3.8)$$

Unter den oben aufgeführten Voraussetzungen erhält man eine Lösung zur Gleichung (3.1) nach der Methode der kleinsten quadratischen Abweichungen (vgl. GOLDBERGER, 1963):

$$\hat{\beta} = (X^T \cdot X)^{-1} \cdot X^T \cdot y . \qquad (3.9)$$

Hier ist $\hat{\beta}$ als Schätzvektorfunktion für β bzw. als optimale Lösung im Sinne der oben erwähnten Methode der kleinsten quadratischen Abweichungen zu verstehen. X^T ist die zur Matrix X transponierte Matrix, deren Zeilenvektoren die Spaltenvektoren von X und umgekehrt, deren Spaltenvektoren die Zeilenvektoren von X, sind. Bei der nicht-singularen Produktmatrix $(X^T \cdot X)$ ist der Rang dieser Produktmatrix und damit auch der Matrix X gleich n. Falls der Rang von X weniger als n, also $Rg(X) < n$, ist, müßte man eine verallgemeinerte Matrix-Inversion anwenden. Hierzu wende man statt $(X^T \cdot X)^{-1} \cdot X^T$ die LANCZOS-Matrix-Inversion (oder eine andere geeignete verallgemeinerte Matrix-Inversion) an:

$$H_L = V_p \cdot \Lambda^{-1}_p \cdot U^T_p , \qquad (3.10)$$

$$\text{dabei gilt } X = U_p \cdot \Lambda_p \cdot V^T_p . \qquad (3.11)$$

Die Zerlegung (3.11) heißt Singuläre-Werte-Zerlegung von X (vgl. STOER u. BULIRISCH, 1978, und JACKSON, 1972). Die Matrizen U_p und V_p beinhalten der Matrix X zugehörige Eigenvektoren. Λ_p ist eine Diagonal-Matrix, deren Diagonalelemente entsprechende Eigenwerte sind, ferner ist $p = Rg(X)$.

Aus den Beziehungen (3.6), (3.7) und (3.9) geht hervor, daß man nicht nur die geschätzten Gemischgewichte $\hat{\beta}_j$, sondern auch die stochastischen Fehler $\hat{e}_i$ der Meßgrößen y_i und damit das geschätzte nicht-gealterte Ölgemischmerkmal $\hat{\bar{y}}_i$ berechnen kann:

$$\hat{e} = y - \hat{\bar{y}} \qquad (3.12)$$

$$\text{mit } \hat{\bar{y}} = \bar{X} \cdot \hat{\beta} . \qquad (3.13)$$

Nehmen wir nun an, daß eine größere Menge eines nichthomogenen Ölgemisches vorliegt. Dazu sind wir wohl zu der Annahme berechtigt, daß die sämtlichen Gemischkomponenten (x_j) (ungefähr) gleich alt sind. An relativ weit auseinanderliegenden Standorten werden die n Proben (y_i) genommen, die durch die folgenden Matrixgleichungen dargestellt werden:

$$\begin{aligned} y_1 &= \bar{X} \cdot \beta_1 + e_1 \\ y_2 &= \bar{X} \cdot \beta_2 + e_2 \\ y_3 &= \bar{X} \cdot \beta_3 + e_3 \\ &\cdot \quad \cdot \quad \cdot \\ &\cdot \quad \cdot \\ &\cdot \quad \cdot \\ y_n &= \bar{X} \cdot \beta_n + e_n . \end{aligned} \qquad (3.14)$$

Aus den Beziehungen (3.8), (3.9), (3.12) und (3.13) stellt man ein zweites Matrixgleichungs-System auf:

$$\begin{aligned} \hat{e}_1 &= E_x \cdot \hat{\beta}_1 \\ \hat{e}_2 &= E_x \cdot \hat{\beta}_2 \\ &\cdot \\ &\cdot \\ &\cdot \\ \hat{e}_n &= E_x \cdot \hat{\beta}_n, \end{aligned} \qquad (3.15)$$

und bildet aus den Spaltenvektoren $\hat{e}_l$ und $\hat{\beta}_l$ zwei neue Matrizen:

$$\hat{E}_y = [\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n] \qquad (3.16)$$

$$\hat{B} = [\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n]. \qquad (3.17)$$

Die Matrizen E_x und $\hat{E}_y$ haben beide m Zeilen und n Spalten, und werden dementsprechend als (m x n)-Matrizen bezeichnet. $\hat{B}$ ist eine (n x n)-Matrix und weist die folgende Beziehung zu E_x und $\hat{E}_y$ auf:

$$\hat{E}_y = E_x \cdot \hat{B}. \qquad (3.18)$$

Wenn $\hat{B}$ nicht-singular ist, folgt durch Nachmultiplikation die Beziehung:

$$\hat{E}_x = \hat{E}_y \cdot \hat{B}^{-1}. \qquad (3.19)$$

Da $\hat{E}_x$ Schätzwerte für die entsprechenden Alterungsabweichungen in X enthält, ist man nach Beziehung (3.7) in der Lage, Schätzwerte für die gealterten Vergleichs-Öle zu berechnen:

$$\hat{X} = \bar{X} + \hat{E}_x. \qquad (3.20)$$

Schließlich möchte ich eine ergänzende Bemerkung zu der am Anfang dieses Abschnitts geäußerten Annahme bezüglich der zwischen den Größen y_i und $\bar{x}_{i,k}$ bestehenden Linearitäten machen.

Falls diese Voraussetzung nicht erfüllt wird – welches man zunächst einmal wohl nicht zu befürchten hat – müßten die Größen $\bar{x}_{i,k}$ durch geeignete nicht-lineare Funktionen $\bar{f}_{i,k} = \bar{f}(\bar{x}_{i,k})$ ersetzt werden.

4. Empirisches Vorgehen: Bestimmung und Verifizierung des Klassifizierungsmodells

Nun wollen wir diejenigen Aspekte aus den Abschnitten 1. bis 3. herausgreifen und anwenden, die zu einem verifizierbaren Klassifizierungsmodell führen. Zunächst wird vorausgesetzt, daß man über eine aus Vergleichsproben- und Feldprobenmustern bestehende Datenbasis, aus der man beliebige Teilmengen (also sogenannte Lern- und Testdatensätze) bilden kann, verfügt. Das angewandte Kriterium für die Klassenzugehörigkeit wird aus den Alterungsprodukten der Vergleichsproben hergeleitet. Praktisch könnte es so aussehen, daß man die Ausgangsdatei in etwa drei Teile aufstellt. Der erste Teil besteht aus den Vergleichsprobenmustern mit ihren Alterungsprobenmustern, der zweite Teil aus Wasserprobenmustern und der dritte Teil aus Vogelprobenmustern. Damit die Meßgrößen verschiedener Muster miteinander vergleichbar werden, muß man als nächsten Schritt die sämtlichen in den Mustern enthaltenen Daten normieren (dafür wende ich ein Programm namens NORM an). Falls man den Verdacht hegt, daß die Klassen nicht (hinreichend) getrennt sind, kann man solche im ARTHUR-Programm befindlichen Unterprogramme wie KARLOV, WEIGHT, SELECT, KNN, HIER und MULTI anwenden. Diese Programme ermöglichen die Feststellung und eventuell die Beseitigung der falschen Klassenzuordnungen von Standardprobenmustern. Zu diesem Zweck bieten WEIGHT und SELECT die Möglichkeit zum Aussortieren der Redundanz-Merkmale (Merkmal-Reduktion) an.

Bevor man einen Klassifizierungslauf zur Identifizierung unbekannter Feldproben durchführt, möchte man die Zuverlässigkeit des Klassifizierungsmodells, die in die nach Alterungsprodukten klassifizierten normierten Vergleichsprobenmuster, schätzen bzw. verifizieren. Dafür richtet man eine speziell für das Programm ARTHUR aufgebaute Datei mit einem sogenannten Lerndatensatz (engl.: training set) und einem Testdatensatz (engl.: test/evaluation set) ein (vgl. RANGER, 1983, und VESPERMANN, 1981). Der Lerndatensatz beinhaltet einen Teil bzw. die Mehrzahl der Vergleichsprobenmuster, der Testdatensatz besteht üblicherweise aus einigen wenigen (restlichen), beliebig gewählten Vergleichsprobenmustern. Auf diese Art und Weise läßt sich auf die Güte des Klassifizierungsmodells schließen. Zur Klassifizierung von Wasser- und Vogelprobenmustern werden diese Muster schließlich in den Testdatensatz eingetragen, und solche Unterprogramme wie KNN oder/aber BAYES bei einem ARTHUR-Programmlauf aufgerufen.

Bisher gingen wir davon aus, daß die Alterungsprobenmuster verschiedener der gealterten Vergleichsprobenmuster zugeordneten Klassen vorliegen. Sollte dies nicht der Fall sein, schlage ich ein empirisches aus dem Abschnitt 2. abgeleitetes Verfahren zur Berechnung »synthetischer« Alterungsprobenmuster vor. Offensichtlich sind die für ein solches Verfahren erforderlichen Bezugsdaten eben die vorhandenen Vergleichsprobenmuster:

$$x_{i,k}^{o''} = x_{i,1,k}^{o''} \qquad (4.1)$$

mit $i = 1,2,3, \ldots, n_{Merkmal}$ und $k = 1,2,3, \ldots, n_{Klasse}$, wobei n_{Klasse} ja die ursprüngliche Anzahl der Vergleichsproben ist. Entsprechend der Gleichung (2.2) bzw. (2.3) wird ein (im räumlichen Sinne) »mittleres« Alterungsprodukt $\bar{x}''_{i,k}$ »erzeugt«:

$$\bar{x}''_{i,k} = x_{i,k}^{o''} \cdot h_{i,k} \qquad (4.2)$$

mit $h_{i,k} = (1 - w_{i,k})$. Hier spielt $h_{i,k}$ die Rolle des Faktors

$$e^{-f_{i,k}(t)}$$ und es ist:

$$h_{i,k} = e^{-\alpha_{i,k} \cdot t} = \bar{x}''_{i,k} \,/\, x_{i,1,k}^{o''} \leqq 1 \qquad (4.3)$$

$$h_{i,k}^{min} = e^{-\alpha_{i,k}^{max} \cdot t_c} = \bar{x}_{i,k}^{min''} \,/\, x_{i,1,k}^{o''} < 1 \qquad (4.4)$$

bzw. $h_{i,k}^{min} = (1 - w_{i,k}^{max})$ mit $0 \leqq w_{i,k}^{max} << 1$.

(Die rechts oben an den Größen geschriebenen Doppelstriche »''« weisen darauf hin, daß es sich um nicht-normierte Merkmale handelt.) Die empirische Größe $w_{i,k}$ heißt empirisches Alterungsgewicht. Bei Verbindungen (bzw. Merkmalen), die auf weniger Stabilitäten hinweisen, könnte man $0{,}015 \leqq w_i \leqq 0{,}030$ setzen, und sonst bei stabileren Merkmalen $w_i < 0{,}015$. Entsprechend der Gleichung (2.1) für statistisch-räumliche Fluktuationen mit der Gauss-Verteilung gilt:

$$x''_{i,j,k} = \bar{x}''_{i,k} + \delta x''_{i,j,k} \qquad (4.5)$$

$$\text{oder } x''_{i,j,k} = \bar{x}''_{i,k} + r_{i,j,k} \qquad (4.6)$$

mit $j = 2,3,4, \ldots, n_{Muster}$. Hier ist $r_{i,j,k}$ eine Zufallsvaribale bzw. ein Zufallszahlengenerator (engl.: random number generator) mit der Gauss-Verteilung (auch Normalverteilung genannt). Ferner hat $r_{i,j,k}$ die Standardabweichung $\sigma_{i,k}$ und den Mittelwert $\mu_{i,k}$. Für diese statistischen Parameter gelten die Beziehungen:

$$\sigma_{i,k} = s_F \cdot w_{i,k} \cdot x_{i,k}^{o''} \qquad (4.7)$$

$$\text{und } \mu_{i,k} = 0 \qquad (4.8)$$

wobei s_F eine Funktion von $w_{i,k}$, also $s_F = s_F(w_{i,k})$, und $0 \leqq s_F \leqq 0{,}667$. Bei $w_{i,k} \leqq 0{,}030$ ist $s_F \leqq 0{,}333$ zu setzen; s_F wird Fluktuations-Streuungskoeffizient genannt. Dementsprechend ergeben sich die Schätzgrößen für das Maximum und das Minimum der erzeugten Alterungsmerkmale:

$$\hat{x}_{i,k}^{max} = \bar{x}''_{i,k} + 1{,}5 \cdot \sigma_{i,k} \qquad (4.9)$$

$$\hat{x}_{i,k}^{min} = \bar{x}''_{i,k} - 1{,}5 \cdot \sigma_{i,k} \qquad (4.10)$$

Falls man lediglich Zugang zu einem Zufallszahlengenerator mit der gleichförmigen bzw. Rechteck-Verteilung der Zahlen im Intervall $0 < u_{i,l,k} < 1$ hat, kann man die erzeugten Zahlen als Wertepaare, also als $u_{i,l_1,k}$ und $u_{i,l_2,k}$ in die normalverteilten Werte $r_{i,j,k}$ durch die folgenden Beziehungen transformieren:

$$v_{i,j,k} = (-2 \cdot \ln u_{i,l_1,k})^{1/2} \cdot \tilde{g}(u_{i,l_2,k}) \qquad (4.11)$$

$$\text{mit } \tilde{g}(u_{i,l_2,k}) = \cos(2 \cdot \pi \cdot u_{i,l_2,k}),$$

$$\text{und } r_{i,j,k} = \sigma_{i,k} \cdot v_{i,j,k} + \mu_{i,k} \qquad (4.12)$$

(Vgl. KNUTH, 1969). Bei einem Zufallszahlengenerator mit der Gauss-Verteilung, der unmittelbar die Zahlen $v_{i,j,k}$ mit der Standardabweichung $\sigma'_{i,k} = 1$ und dem Mittelwert $\mu'_{i,k} = 0$ erzeugt, braucht man sonst offensichtlich nur die Beziehung (4.12) (zum Erzeugen der künstlichen Alterungsprobenmuster habe ich ein Prototypprogramm namens SYNPAT [engl.: synthetic pattern] erstellt).

Nun werden nach geeigneter Wahl der empirischen Alterungsgewichte $w_{i,k}$ (und s_F) bezüglich der nicht-normierten ursprünglichen Vergleichsprobenmuster die synthetischen Alterungsmuster erzeugt. Bevor die oben geschilderten Mustererkennungs-Methoden verwendet werden können, müssen die Daten zunächst einmal normiert werden. Wir wählen eine gewichtete Normierung der Form:

$$x_{i,j,k} = \tilde{w}_i \cdot (x''_{i,j,k} - \bar{x}''_{i,k}) / g_{j,k} \qquad (4.13)$$

$$\text{mit } g_{j,k} = \left[(n/(n-1)) \cdot (m'_2)_{j,k}\right]^{1/2},$$

dabei ist
i der Merkmalsindex
mit $i = 1,2,3 \ldots, n_{Merkmal}$,
j der Musterindex
mit $j = 1,2,3, \ldots, n_{Muster}$,
k der Klassenindex mit
$k = 1,2,3, \ldots, n_{Klasse}$,
n die Anzahl der Merkmale im Muster,
$n = n_{Merkmal}$,
$\bar{x}''_{j,k}$ der Mittelwert der Meßdaten (bzw. der Merkmalwerte) in dem j.-Muster der k.-Klasse mit

$$\bar{x}''_{j,k} = (1/n) \cdot \sum_{i=1}^{n} x''_{i,j,k} \qquad (4.14)$$

$(m'_2)_{j,k}$ das zweite gewichtete Moment mit

$$(m'_2)_{j,k} = (1/n) \cdot \sum_{i=1}^{n} \tilde{w}_i^2 \cdot (x''_{i,j,k} - \bar{x}''_{j,k})^2, \qquad (4.15)$$

$\tilde{w}_i$ das Normierungsgewicht des i.-Merkmals bei sämtlichen Mustern und sämtlichen Klassen mit $0 \leqq \tilde{w}_i \leqq 1$.

Die somit transformierten Muster (x_i) kann man als Stichproben mit der konstanten Varianz $\sigma_x^2 = 1$ und dem konstanten Mittelwert $\mu_x = 0$ betrachten.

5. Schlußbemerkungen

Gegenüber den Klassifizierungsmodellen, die lediglich die aus Rohölen stammenden Vergleichsprobenmuster enthalten, können diejenigen problematisch sein, in denen auch die aus Produktölen stammenden Muster vorkommen. Produktöle sind teilweise deshalb schwierig zu klassifizieren, weil sie einerseits den Rohölen, von denen sie abstammen, recht ähnlich sind und andererseits untereinander große Ähnlichkeiten aufweisen können. Eine wichtige Rolle bei der richtigen Klassifizierung in der Mustererkennung spielt das Verhältnis $n_{Muster}/n_{Merkmal}$. Bei dem Verhältnis $n_{Muster}/n_{Merkmal} = 2$ schätzen DUEWER und Mitarbeiter die Wahrscheinlichkeit einer richtigen Klassifizierung auf etwa 0,50. Bei einer solchen Wahrscheinlichkeit könnte das Modell ja soviel falsche wie richtige Identifizierungen bringen. Deshalb empfiehlt DUEWER ein Verhältnis $n_{Muster}/n_{Merkmal} \geqq 4$ (vgl. DUEWER et al., 1975a, auch ZIEGLER, 1984).

Die Alterungsprodukte lassen sich auf verschiedene Art und Weise herstellen.

Solche Verfahren sollen die in freien Gewässern herrschenden Prozesse physikalisch-chemisch-biologischer Natur simulieren. DUEWER und Mitarbeiter beschreiben einige Varianten zur Herstellung von Alterungsproben im Labor (vgl. DUEWER et. al., 1975b). Zur empirischen Beschreibung der Alterungsprozesse wäre es wünschenswert, umfangreiche, in situ sowie in vitro angelegte Versuche durchzuführen. Derartige Versuche dienten einerseits dem Zweck der Klassifizierung und andererseits der Verifizierung (bzw. der begründeten Ablehnung) solcher theoretischen Modellgrößen wie des Abbaukoeffizienten $\alpha_{i,k}^{max}$. Diese Endergebnisse könnten auch dadurch erzielt werden, daß man über eine längere Zeit hinreichende statistische Daten systematisch erfaßt und sinnvoll und sorgfältig auswertet. In der Tat ist es das Ziel dieses Beitrags, zu versuchen, die Voraussetzungen für eine sinnvolle Datenauswertung der Ölprobenmuster zu beschreiben. In diesem Sinne bin ich der Meinung, daß die sorgfältige Erzeugung mathematisch simulierter Daten auch sinnvoll ist. Empirische Ansätze zu dem im Abschnitt 2 geschilderten Alterungsmodell sind in der Literatur zu finden. E. MERIAN teilt z.B. mit, daß die im Wasser befindliche Verbindung Benzol schnell verdunste und eine Halbwertzeit von etwa 37 Minuten bei der Zimmertemperatur habe (Vgl. MERIAN, 1983).

Zur sinnvollen Anwendung des Verfahrens für die synthetische Erzeugung der Alterungsprobenmuster gehört die geeignete Auswahl der empirischen Altersgruppengewichte $w_{i,k}$. Dafür möchte ich drei Wertebereich-Klassen vorschlagen. Eine mögliche Einteilung sieht dann in etwa so aus:

1. Der Skeptiker-Schätzbereich:
$0 \leqq w_{i,k} \leqq w_{i,k}^{max}$, mit $w_{i,k}^{max} \approx 10^{(i_s)}$, wobei i_s eine ganzzahlige Zehnerpotenz, die der Größenordnung des Meßfehlers entspricht, ist. Z.B. wäre bei einem absoluten geschätzten Meßfehler von 0,005 $i_s = -3$.

2. Der Optimisten-Schätzbereich:
$w_{i,k}^{min} \leqq w_{i,k} \leqq w_{i,k}^{max}$, wobei $w_{i,k}^{min} \geqq 0$ ist. $w_{i,k}^{max}$ wird stufenweise solange erhöht, daß die Trennung der Klassen erhalten bleibt.

3. Der empirische Schätzbereich:
$w_{i,k}^{min} \leqq w_{i,k} \leqq w_{i,k}^{max}$, wobei $w_{i,k}^{min} \geqq 0$ ist. $w_{i,k}^{min}$ und $w_{i,k}^{max}$ werden auf die empirischen Dichten der Klassen (Punktewolken) eingestellt.

Der Skeptiker-Schätzbereich stellt eine Unter- bzw. konservative Schätzung der Alterungen dar und wäre zum ersten Klassifizierungsmodell als beginnender Bereich geeignet. Solange Kontaminierungseffekte im wesentlichen auszuschließen sind, könnte man stufenweise vom Bereich der Klasse 1 zum Bereich der Klasse 2 umsteigen. Nachdem die Musterklassen im Klassifizierungsmodell stärker mit Feldprobenmuster belegt werden, kann man sinnvoll versuchen, empirische Schätzwerte für $w_{i,k}^{min}$ und $w_{i,k}^{max}$ zu bestimmen.

Die übermäßigen Aufblähungen der Grenzwerte $w_{i,k}^{max}$ bringen das Risiko mit sich, daß signifikante Kontaminierungs- bzw. Gemischkomponenten im Klassifizierungsmodell verwischt werden. Wenn das Klassifizierungsmodell mit relativ wenigen Feldprobenmuster belegt ist, kann man zunächst davon ausgehen, es handele sich um einen Gemischfall, vorausgesetzt, daß zumindest ein Feldprobenmuster ungefähr gleichweit von verschiedenen Musterklassen (Punktewolken) entfernt ist. (Natürlich muß in einem solchen Falle das Klassifizierungsmodell sämtliche relevante Klassen enthalten.) Anscheinend sind die Voraussetzungen zur optimalen Bestimmung eines Ölgemisches in einem freien Gewässer erst dann gegeben, wenn das Klassifizierungsmodell mit (sehr) vielen Feldproben belegt ist.

In Abschnitt 2 wurde behauptet, daß die Wasserproben-Situation »geradezu optimale« Bedingungen für die Klassifizierung von Feldproben anbietet. Diese günstigen Bedingungen kann man als die (starken) räumlich-zeitlichen Einschränkungen bei den Schiffs- und Wasserprobenahmen beschreiben. Bis auf flüchtige Verbindungen (und möglicherweise Kontaminierung durch andere Kohlenwasserstoffe) bestehen eine Schiffsprobe und die Wasserproben aus dem gleichen Öl, praktisch gleichen Alters. Für solche Vergleiche ist ein Klassifizierungsmodell mit geringfügig gealterten Alterungsproben, also mit dem Skeptiker-Schätzbereich, geeignet. Unter diesen Bedingungen fallen die möglichen kontaminierten Wasserprobenmuster auch stärker auf und stellen als Gemische größere Signifikanzen dar.

Nun kehren wir zu der Vogelproben-Situation zurück. Allgemein gültig kann man behaupten, daß die oben erwähnten Einschränkungen weniger (stark) ausgeprägt sind. Diese Behauptung läßt sich teilweise durch das Verhalten der Vögel, teilweise durch Wind- und Seegangsverhältnisse und nicht zuletzt, durch die Probenahmen selbst erklären. Ein Beispiel für die großräumige Beweglichkeit der auf Helgoland brütenden Lummen wird von G. VAUK u. K. PIERSTORFF (1973) bestätigt. Eine Untersuchung der in der Nordsee und Deutschen Bucht auftretenden Windstärken über einen Zeitraum von 20 Jahren läßt nach VAUK und PIERSTORFF vermuten, daß »dieses treibende Öl allmählich vom Wind zusammengedrückt wird, größere ... Ölflächen bildet und schließlich an die Küsten [wo sich mehrere Vogelspezies oft aufhalten (d. Verf.)] getrieben wird«. Solche überwiegend aus dem Westen wehenden Winde verstärken demzufolge die chronische Ölverseuchung im Raum der Deutschen Bucht und an der Westküste Helgolands (vgl. VAUK und PIERSTORFF, 1973). Dazu kommt die Vermutung, daß nur ein Teil der Gefiederproben u.a.m. von frisch ausgelaufenem bzw. nicht-gealtertem Öl stammt (vgl. VAUK, 1981, 1982, 1983). Aus diesen Feststellungen bzw. Vermutungen läßt sich schließen, daß bei den Vogelproben teilweise die stärker gealterten Öle vorkommen. Die Möglichkeit, daß eine Vogelprobe durch andere Öle kontaminiert ist, kann nicht ausgeschlossen sein. Darüber hinaus teilte G. DAHLMANN mit, daß bis zum April 1984 591 Vogelproben an das Deutsche Hydrographische Institut geliefert wurden. Den chemischen Gutachten zufolge bestanden 5 Proben aus Rohölen, die restlichen Proben aus Schiffs- bzw. Pro-

duktölen (vgl. DAHLMANN, 1984). Wegen der teilweise starken Ähnlichkeiten unter den Produktölen muß man befürchten, daß die eindeutigen Klassifizierungen erschwert und zum Teil unmöglich sind. Diese Gegebenheit läßt sich durch ein Beispiel erläutern. Wenn bei einer Feldprobe V nicht nur Öltyp A vorkommt, sondern auch Öltyp B und Öltyp C, wird bei der Feldprobe V nicht nur eine Zugehörigkeit zur Klasse A, sondern auch zur Klasse B und zur Klasse C aufgrund der Ähnlichkeiten (also Abständen) zwischen Muster V und der Klasse A, B und C, durch den Einsatz des Programms KNN ermittelt. Eine solche Klassifizierung bezeichnet man als mehrdeutig; sie ist dennoch ebenso schlüssig wie eine eindeutige Klassifizierung.

6. Zusammenfassung

Theoretische und angewandte Methoden zur Klassifizierung der verölten, in den freien Gewässern der Küstenbereiche Nord-Deutschlands vorgefundenen Wasser- und Vogelproben werden in diesem Beitrag geschildert. Der physikalisch-chemisch-biologische Abbau (bzw. Alterung) sowie die Kontaminierung der sich in den freien Gewässern befindlichen Kohlenwasserstoffe werden in kurzer Form als Voraussetzung zur Anwendung des in der Mustererkennung bekannten, überwachten Lernens untersucht. Im Abschnitt 2 wird ein statistisch-mathematisches Modell für die natürlichen Abbau-Prozesse als das für die radioaktiven Zerfall-Prozesse geltende Analogon angesetzt. Ein im Abschnitt 3 mit stochastischem Fehler-Glied ausgestattetes multilineares Regressions-Modell beschreibt eine Beziehung zwischen den Abbau- und den Kontaminierungs- (bzw. Gemisch-) Prozessen. In den Abschnitten 4 und 5 wird ein empirisches Vorgehen für eine mathematische Simulierung des in der Umwelt abgebauten Öls geschildert, und eine Formel für die gewichtete Normierung angegeben. Ferner werden einige im Programm-Paket ARTHUR angebotene Methoden bzw. Programme zur Mustererkennung erwähnt. Die zwischen dem Wasserprobenahme- und dem Vogelprobenahme-Verfahren bestehenden Unterschiede, die sich als räumliche und zeitliche Korrelationen mit den etlichen Quellen- und Feldproben manifestieren, werden, im bezug auf die in der Umwelt vorkommenden Abbau- sowie die eventuellen Kontaminierungs-Prozesse, hervorgehoben.

7. Summary

Theory and applied methods for classifying oil-polluted water and bird samples found in the coastal environment of Northern Germany are discussed. The physical, chemical und biological degradation and the contamination of hydrocarbons in environmental waters are briefly explored as conditions for the application of supervised-learning in pattern recognition. Section 2. presents a stochastic model for natural degrading processes analogous to that for radioactive decay. A multilinear regression model with a stochastic error term in Section 3. links degradation to contamination (mixture) processes. Sections 4. and 5. describe an empirical approach for mathematical simulation of environmentally degraded oil and a formula for weighted normalization. Further, mention is made of methods and programs included in the program package ARTHUR for pattern recognition. Differences between water- and bird-sampling techniques in terms of spatial and temporal correlations between various field samples and source samples, including implications for environmental degradation and possible contamination, are highlighted.

8. Literatur

Seitenangaben bei Büchern weisen auf die Stelle des benutzten Zitates hin.

ANDREWS, H.C. (1972): Introduction to Mathematical Techniques in Pattern Recognition; J. Wiley and Sons, New York: 65–91

DAHLMANN, G. (1984): Protokoll: Arbeitsgespräch (am 18. Mai 1984) im »Haus der Natur« des Verein Jordsand, Wulfsdorf/Hamburg; unveröffentlicht: 1–9

DUDA, R.O. u. P.E. HART (1973): Pattern Classification and Scene Analysis; J. Wiley and Sons, New York: 151-152

DUEWER, D.L., J.R. KOSKINEN u. B.R. KOWALSKI (1975a): Documentation for ARTHUR, Version 1-8-75 (8. Januar 1975); Chemometrics Society Report Nr. 2, Department of Chemistry BG-10, University of Washington, Seattle, Washington (98195), USA.

DUEWER, D.L. u. B.R. KOWALSKI (1975b): Source Identification of Oil Spills by Pattern Recognition Analysis of Natural Elemental Composition. – Analytical Chemistry 47/9: 1573–1582

FREUND, J.E. (1962): Mathematical Statistics; Prentice Hall, Inc., Englewood Cliffs, New Jersey, USA, 5. Auflage, January 1965: 57, 209–212

FRIOCOURT, M.P., F. BERTHOU u. D. PICART (1983): Dibenzothiophene Derivatives as Organic Markers of Oil Pollution. – In: Chemistry and Analysis of Hydrocarbons in the Environment (J. ABAIGÉS, R. W. FREI u. E. MERIAN; Hrsg.); Gordon and Breach Science Publishers, New York: 125–135

GOLDBERGER, A.S. (1963): Econometric Theory; J. Wiley and Sons, New York: 156–160

JACKSON, D.D. (1972): Interpretation of Inaccurate, Insufficient and Inconsistent Data. – Geophysical Journal of the Royal Astronomical Society 28: 97–109

KILLEEN, T.J. u. Y.T. CHIEN (1976): A Probability Model for Matching Suspects with Spills ...or Did the Real Spiller Get Away; The University of Connecticut, Storrs, Connecticut: 66–72

KOWALSKI, B.R. (ca. 1973): Pattern Recognition in Chemical Research; Department of Chemistry, University of Washington, Seattle, Washington: 1–76

KNUTH, D.E. (1969): The Art of Computer Programming; Addison-Wesley Publishing Co., New York

KREISZIG, E. (1975): Statistische Methoden und ihre Anwendungen; Vandenhoek und Ruprecht, Göttingen: 92–93

LANCZOS, C. (1961): Linear Differential Operators; D. Van Nostrand Co., London: 100–201

LANDAU, L.D. u. E.M. LIFSCHITZ (1979): Lehrbuch der Theoretischen Physik, Statistische Physik, Band 5, Teil 1; Akademie-Verlag, Berlin: 316–318

MERIAN, E. (1983): The Environmental Chemistry of Volatile Aromatic Hydrocarbons. – In: Chemistry and Analysis of Hydrocarbons in the Environment; Gordon and Breach Science Publishers, New York: 167–175

RANGER, J.C. (1983): Anwendung der Mustererkennung zur Klassifizierung von Rohölen; Deutsches Hydrographisches Institut, Laboratorium-Sülldorf (Wüstland 2, 2000 Hamburg 55), Januar 1983: 1–10, »A-1« –»A-3«

SACHS, L. (1984): Angewandte Statistik, Anwendung statistischer Methoden; 6. Auflage; Springer-Verlag, Berlin: 49–89

STOER, J. u. R. BULIRISCH (1978): Einführung in die Numerische Mathematik II, 2. Auflage; Springer-Verlag, Berlin: 14–21 und 29–33

VAUK, G. (1981): Ölpestbericht Helgoland 1980. – Seevögel 2/1: 63–66

VAUK, G. (1982): Ölpestbericht Helgoland 1981. – Seevögel 3/2: 107–109

VAUK, G. (1983): Ölpestbericht Helgoland 1982. – Seevögel 4/1: 1–3

VAUK, G. u. K. PIERSTORFF (1973): Ergebnisse dreizehnjähriger Ölpestbeobachtungen auf Helgoland (1960–1972). – CORAX 4/2–3: 136–146

VESPERMANN, A. (1981): IPA – ein Programm zur interaktiven Mustererkennung. – Proceedings der Tagung der Deutschen Arbeitsgemeinschaft für Mustererkennung, Control Data GmbH, (Distrikt Nord, Mexikoring 23, 2000 Hamburg 60) Hamburg: 388–394

WEIDNER, R.T. u. R.L. SELLS (1960): Elementary Modern Physics; Allyn and Bacon, Inc., Boston, USA, 3. Auflage, August 1961: 334–337

ZIEGLER, E. (1984): Mustererkennung/Pattern Recognition. – In: Computer in der Chemie; Springer-Verlag, Berlin: 133–153

Anschrift des Verfassers:

Jon C. Ranger
Deutsches Hydrographisches Institut
Laboratorium Sülldorf
Wüstland 2
2000 Hamburg 55