Stochastik — eine interdisziplinäre Wissenschaft. Teil I [1]

L. Rüschendorf
Institut für Mathematische Stochastik,
Universität Freiburg [2]

Im folgenden sollen Probleme und Methoden der Stochastik anhand einiger Beispiele erläutert werden. Ich werde dabei nicht so sehr auf die innermathematische Entwicklung dieses Gebietes eingehen, sondern die Aspekte betonen, die von interdisziplinärem Charakter sind. Die Stochastik hat sich von Anfang an durch Fragestellungen aus und in den unterschiedlichsten Gebieten entwickelt. Die frühen hervorragenden Stochastiker waren zu einem großen Teil Theologen, Juristen, Astronomen, Genetiker, Agronomen, Physiker, Psychologen, Geometer und Geologen. Die mathematisch orientierten Stochastiker haben dann das Gebiet vereinheitlicht und die wesentlichen Grundlagen und Systematik geschaffen. Ausdruck des interdisziplinären Charakters der Stochastik ist auch das Ende 1994 gegründete Freiburger Zentrum für Datenanalyse und Modellbildung (FDM), das zur Zeit aus Mitgliedern der Biometrie, Physik und mathematischen Stochastik besteht und mit weiteren Forschungsrichtungen der Universität, Wirtschaft und Industrie kooperiert. Ein zentrales datenanalytisches Motiv für die Bedeutung der Stochastik in unterschiedlichen Anwendungsgebieten kommt in folgender Schlagzeile eines Artikels über das FDM gut zum Ausdruck (aus der Badischen Zeitung vom 10.11.1995):

Am Zentrum für Datenanalyse und Modellbildung der Freiburger Universität arbeiten Statistiker. Physiker und Mathematiker zusammen

Die Wissenschaftler als Lotsen in der Datenflut

In Deutschland ist die Universitätsforschung auf dem Gebiet Stochastik bis auf wenige Ausnahmen separat in den verschiedenen Fakultäten organisiert; die mathematische Stochastik ist im wesentlichen an allen mathematischen Fakultäten integraler Bestandteil der Mathematikausbildung. In vielen anderen, insbesondere den angelsächsischen Ländern, ist die Stochastik in all ihren Verzweigungen typischerweise in großen Statistics Departments gemeinsam organisiert. Eine Konsequenz dieser Situation ist das Bestehen einer Reihe von statistischen Gesellschaften in Deutschland wie z.B. Deutsche Statistische Gesellschaft (DSTG), Biometrische Gesellschaft, Gesellschaft der Städtestatistiker, Fachgruppe Stochastik der DMV (in der die mathematisch orientierten Stochastiker organisiert sind). Einen Eindruck von der Vielzahl der beteiligten Gebiete und Methoden gibt die folgende unvollständige Übersicht:

---
[1] Festvortrag zur Verleihung des Gödecke-Forschungspreises am 17.11.1995 in Freiburg.
[2] Eckerstraße 1, D-79104 Freiburg /u/st*/o*/R*/P*/stochastik-I

	Agriculture	Authorship	Artificial Intelligence	Biology	Business & Stock Market	Chemical Engineering	Clinical Trials	Credit Scoring	Commn. Th. & Cryptology	Econometrics	Engineering	Law	Manufacturing	Market Research	Medical Diagnosis	Meteorology	Oper. Res.	Pattern Recognition	Philosophy of Science	Population Genetics	Psychology	Seismology	Soc., Polit. Sc., Govt.	Statist. Mechanics
Assay				●																				
Bayes. Wt. of Evidence[a]		●	●					●	●			●			●				●					
Contingency Tables				●											●								●	
Decision Theory[a]					●														●					
Design	●										●								●					
EDA & Descriptive Statistics	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Entropy & Cross-Entropy Max.									●									●	●					●
Markov chains & Renewal Theory																	●							
Multivariate (cont.)				●	●																●		●	
Nonparametrics[b]	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●
Path Analysis																				●			●	
Quality Control													●											
Reg., ANOVA, Linear Models	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●
Reliability													●											
Response Surfaces						●																		
Sequential Analysis							●						●											
Survey Sampling[a]														●									●	
Time Series & Stochastic Processes					●					●						●						●		

Einige Anwendungsgebiete von statistischen Methoden
(aus: Kotz & Johnson: Encyclopedia of statistical science.)

Gliederung:

I. Historische Anmerkungen

II. Zufall und Wahrscheinlichkeit

III. Stochastische Modellbildung

Im separaten zweiten Teil dieser Darstellung soll dann detaillierter auf folgende Anwendungsbeispiele eingegangen werden:

A) Stochastik in Algorithmen und Optimierung

B) Klassifikation und räumliche Modelle

C) Finanzmathematik und stochastische Prozesse

D) Biometrie und Expertensysteme

E) Emissionstomographie

Ziel ist es, daß deutlich wird, wo die Ursachen für stochastische Analysen liegen, mit welchen Problemen und Methoden sich Stochastiker befassen und darzustellen

1.) die Vielfalt der Probleme und Anwendungen stochastischer Methoden und Modelle

2.) das Potential für wissenschaftliche Erklärungen

I. Historische Anmerkungen

Stochastik ist die Wissenschaft, die sich mit zufallsabhängigen Phänomenen beschäftigt. Das Wort stammt aus dem Griechischen

$\sigma\tau\omega\chi\alpha\zeta\varepsilon\sigma\theta\alpha\iota \quad \sim$ mutmaßen, vermuten

Stochastische Phänomene sind solche, deren Determiniertheit nicht festzustellen ist, die aus ihren bedingenden Ursachen oder wegen ihrer Komplexität nicht eindeutig beschreibbar sind. Insofern liefert der Zufall keine kausale Erklärung von Ereignissen (im deterministischen Sinne), sondern verweist zumindest in Teilen auf 'noch' nicht verstandene Anteile oder auf (prinzipiell) nicht erfassbare Quellen. So wurde Zufall in Form von Los-Orakeln schon in frühen Geschichtsepochen zur Erforschung des göttlichen Willens verwendet.

Es hat in der Wissenschaftsgeschichte und insbesondere in der Entwicklung der neueren Physik grundlegende philosophische und wissenschaftstheoretische Auseinandersetzungen in großer Fülle gegeben, inwiefern **Zufall** eine konstituierende, denknotwendige Größe in der Beschreibung der Gesetzmäßigkeiten der Mikrowelt der Atome durch die Quantentheorie ist. Man denke hier etwa an die Quanten-/Wellennatur des Lichtes, das Plancksche Wirkungsquantum oder die Heisenbergsche Unschärferelation. Diese Diskussion ist bis heute nicht abgeklärt. Eindrucksvoll hat sich aber herausgestellt, daß die **Stochastik**, bestehend aus **Wahrscheinlichkeitstheorie**, **Statistik** und deren **philosophischer Aspekte**, mit Erfolg in diejenigen Sphären menschlichen Denkens eingedrungen ist, die auf den ersten Blick der wissenschaftlichen Strenge und Gesetzmäßigkeit entbehren und sich nur schwer dem mathematischen Formalismus unterziehen zu wollen scheinen. Die folgende Schlagzeile eines Zeitungsberichtes von H. Witting über die Aufgabe der mathematischen Statistik bringt diesen Gedanken einprägsam zum Ausdruck.

Gewißheit in der Ungewißheit

Von Hermann Witting

Die Stochastik ist als Teil der **wissenschaftliche Methode (Rationalität)** in fast alle wissenschaftlichen Bereiche in Natur-, Geistes- und Sozialwissenschaften eingedrungen und hat sich aus diesen Fragestellungen entwickelt. Wesentliche Elemente und Qualitäten der wissenschaftlichen Methode (Rationalität) sind:

1.) Interesse und Bedeutung des Forschungsgegenstandes
2.) Beobachtung, Beschreibung und Klassifikation der Objekte
3.) Wiederholbarkeit der Experimente
4.) Verwendung genügend genauer Begriffe und Notationen
    (nicht zu einengende Begriffe in der Exploration)
5.) Feststellung von Korrelation zwischen Phänomenen
    Diese Feststellung führt zu vermuteten Beziehungen (Induktion) oder zu kausalen Erklärungen
6.) Vorhersage und Kontrolle basierend auf Theorien, Konstruktion neuer Theorien
    Kontrolle und Vorhersage stützen eine Theorie, Theorie dient der Erklärung
7.) Anwendung operationaler Regeln auf axiomatische Theorien
    (z.B. auf undefinierte Terme in der Quantenmechanik)
8.) Ableitung von Theorien und/oder Hypothesen und deren Überprüfung in Experiment und Beobachtung
    (prädikativ oder explanatorisch)
9.) Maß für Objektivität
    (Ausschluß von Wunschdenken)
10.) Vereinheitlichung des Gebietes

Die Wahrscheinlichkeitsrechnung ist entstanden aus der Diskussion von Glücksspielen im 17. Jahrhundert (Cardano, Huygens, Pascal, Fermat). Als zentrale Quelle gilt ein Briefwechsel aus dem Jahre 1654 von dem Mathematiker und Philosophen Blaise Pascal mit dem Mathematiker und Juristen Pierre Fermat zu mathematischen und moralischen Fragen des Grafen Antoine de Méré.

Pascal (1623-1662)

Fermat (1601-1665)

Huygens (1629-1695)

Bernoulli (1654-1705)

Huygens veröffentlichte in seinen 'De ratiociniis in ludo aleae' eine Sammlung von 14 Problemen mitsamt Lösungen zu Würfelspielen.

Das erste systematische Lehrbuch der Stochastik stammt von Jakob Bernoulli, die 'Ars conjectandi' aus dem Jahre 1713. Im Eingang dieses Werkes werden die von Huygens stammenden Probleme zur kombinatorischen Wahrscheinlichkeit aufgegriffen. Im dritten Teil seiner Schrift behandelt Bernoulli Einzelprobleme der Wahrscheinlichkeitsrechnung, bei deren Lösung er seine Ergebnisse aus der Kombinatorik (Teil 2) verwenden kann. Der wichtigste Teil der 'Ars conjectandi' ist der Teil 4. Hier führt er die 'aleae geometria', die Pascalsche Mathematik des Glücksspiels, aus der Enge der Spielsalons heraus. Schon der Titel des ganzen Werkes von Bernoulli macht diese Zielsetzung deutlich. Im Titel der Schriften von Cardano und Huygens, aber auch in dem der von Pascal geplanten Arbeit ist der Bezug zum Glücksspiel da. Bei Bernoulli aber geht es um eine allgemeine 'Kunst der Vermutung'[3] :

> "Irgend ein Ding vermuten heißt seine Wahrscheinlichkeit messen. Deshalb bezeichnen wir soviel als *Vermutungs- oder Mutmaßungskunst* (Ars conjectandi sive stochastice) die Kunst, so genau wie möglich die Wahrscheinlichkeit der Dinge zu messen und zwar zu dem Zwecke, daß wir bei unseren Urteilen und Handlungen stets das auswählen und befolgen können, was uns besser, trefflicher, sicherer oder ratsamer erscheint. Darin allein beruht die ganze Weisheit der Philosophen und die ganze Klugheit des Staatsmannes."

Moivre erweitert Bernoullis Darstellung. Er beweist in seinem Buch 'Doctrine de chance' den ersten zentralen Grenzwertsatz. 1812 publiziert Laplace die 'Théorie analytique des probabilités'. Über dieses grundlegende analytische Werk schreibt H. Suter in seiner 1875 erschienenen 'Geschichte der mathematischen Wissenschaften':

> "Dasselbe zerfällt in drei Haupttheile, eine Einleitung, die Laplace auch als selbstständiges Werk herausgegeben hat, unter dem Titel: *Essai philosophique sur les probabilités*, ein erstes Buch: *Du calcul des fonctions génératrices*, und ein zweites Buch: *Théorie général des probabilités*.

[3]Aus Meschkowski: Problemgeschichte der Mathematik II.

Die Einleitung verdient ihren Titel im vollsten Maasse; sie ist vielleicht eines der tiefsten und gedankenreichsten Werke, die auf dem Gebiete der Mathematik veröffentlicht worden sind. Die Entwicklung der ganzen Theorie der Wahrscheinlichkeiten ohne das Hülfsmittel der Analysis, die Anwendung derselben auf die interessantesten Fragen des polititschen und socialen Lebens, kurz, die feine Behandlungsweise des so reichen und alle Wissenszweige berührenden Stoffes, machen diese Einleitung zu einer wahrhaft philosophischen Abhandlung. Schade, dass die heutigen Philosophen das Wesen der mathematischen Wissenschaften nicht mehr humanistisch genug finden! Sie könnten aus dem Studium solcher Werke die Möglichkeit einer innigen Berührung rein abstracter Grundsätze und Theorien mit den mannigfaltigsten Verhältnissen und Actionen des Lebens auf's Schönste erkennen. — Die Hauptcapitel der Einleitung sind diejenigen, welche über die Anwendung der Wahrscheinlichkeitsrechnung auf die Naturphilosophie, auf Wahlen und Beschlüsse von Versammlungen, auf richterliche Urtheile, auf Mortalität, Lebensversicherungen und Rentenanstalten, über die Täuschungen in der Schätzung der Wahrscheinlichkeit und über die Geschichte dieses Calcüls handeln. Dieser letztere Abschnitt, obgleich sehr kurz, gibt ein äusserst prägnantes Bild der historischen Entwicklung der Wahrscheinlichkeitsrechnung von Pascal bis auf Laplace."

Weitere wichtige Stationen der Entwicklung sind die Gesetze seltener Zahlen von Poisson (1781-1840) sowie die Einführung der Normalverteilung von Gauß (1777-1855) in Zusammenhang mit der Entwicklung der Fehlerausgleichsrechnung.

26. bis 29. März 1996

Freiberger Stochastik-Tage

Fachgruppe Stochastik

Die moderne axiomatische Wahrscheinlichkeitstheorie entsteht seit Mitte des 19. Jahrhunderts. Ein entscheidender Schritt ist die von Kolmogorov 1933 vorgenommene Axiomatisierung durch die Maßtheorie.

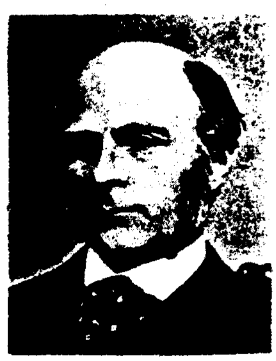
Die **Statistik** hat ihren Ursprung im Staatswesen; italienisch Statista = Staatsmann.

Die ersten wichtigeren Beiträge stammen von den Engländern Petty und Grant. Sie nannten ihre Forschungsrichtung 'Politische Arithmetik'. Ziel war die Gewinnung von Erkenntnissen und Informationen aus Daten über die Bevölkerung, Handel, Ökonomie und im Staatswesen.

Grant (1620-1674) und **Petty** (1623-1687) beschäftigen sich auch mit dem merkwürdigen Phänomen, daß das Geschlechterverhältnis von Neugeborenen sich stabilisiert trotz individueller Zufälligkeit, sowie mit Sterberaten und -tafeln. **Halley** (1656-1748) entwickelte Sterbetafeln in Breslau zum Verkauf von Leibrenten. **Gauß** (1777-1855) und **Bessel** (1784-1848) untersuchen geodätische und astronomische Daten, die mit Fehlern behaftet sind. Gauß war Leiter der Sternwarte in Göttingen. Mit Hilfe der von ihm entwickelten Fehlerausgleichsrechnung entdeckte er den Planeten Ceres im November 1801 wieder, nachdem dieser in der Silvesternacht 1800/1801 erstmalig entdeckt, dann bis Mitte Februar von Astronomen beobachtet, jedoch danach aus den Augen verloren wurde.

Die moderne Entwicklung der Statistik geschieht seit etwa 1890 und ist verbunden mit Forschern wie Quetelet, Galton, Pearson, Fisher, Neyman, und Wald.

K. Pearson (1857-1936)



Galton (1822-1911)

Fisher (1890-1962)

Neyman (1894-1981)

Wald (1902-1950)

Zu Beginn standen empirische Studien, Methoden der explorativen Datenanalyse (EDA), multidimensionale Skalierung (MDS), Faktorenanalyse (Psychometrie), Seriation (Archäologie) und die Bestimmung und Darstellung von Korrelationen im Vordergrund. In der Physik hatte die statistische Mechanik von Maxwell, Boltzmann und Gibbs ihre großen Erfolge. Bose, Einstein und Dirac entwickelten die ersten Modelle der Quantenstatistik. Snow (1848) führte epidemiologische Untersuchungen über Verlauf und Ausbreitung der Cholera durch. Hiermit konnten erstmalig Infektionen als Ursache der Erkrankung nachgewiesen werden. Aus Problemen der Agrarwissenschaft bildete sich die für viele Anwendungen besonders wichtige Versuchsplanung heraus. Eine Reihe von führenden Statistikern hatten Lehrstühle in der Genetik inne. An Fragen der Weiterentwicklung und Evaluierung der Mendelschen Gesetze bildete sich das statistische Instrumentarium heraus.

II. Zufall und Wahrscheinlichkeit

Wahrscheinlichkeitsmaße beschreiben zufällige Ereignisse, aber der Begriff der **Wahrscheinlichkeit** ist nicht identisch mit dem Begriff des **Zufalls**.

In einem Münzwurfexperiment werde das Ergebnis eines n-maligen Wurfes einer Münze mit einer Folge aus den Symbolen 0 und 1 beschrieben, $0 \mathrel{\hat{=}}$ Wappen, $1 \mathrel{\hat{=}}$ Zahl. Dieses Experiment ist nicht deterministisch beschreibbar.

Betrachtet man etwa die folgenden drei Ergebnisfolgen:

$$
\begin{aligned}
x_1 &= 111111111111111111111111\\
x_2 &= 100110011001100110011001\\
x_3 &= 101101000110101100011110
\end{aligned}
$$

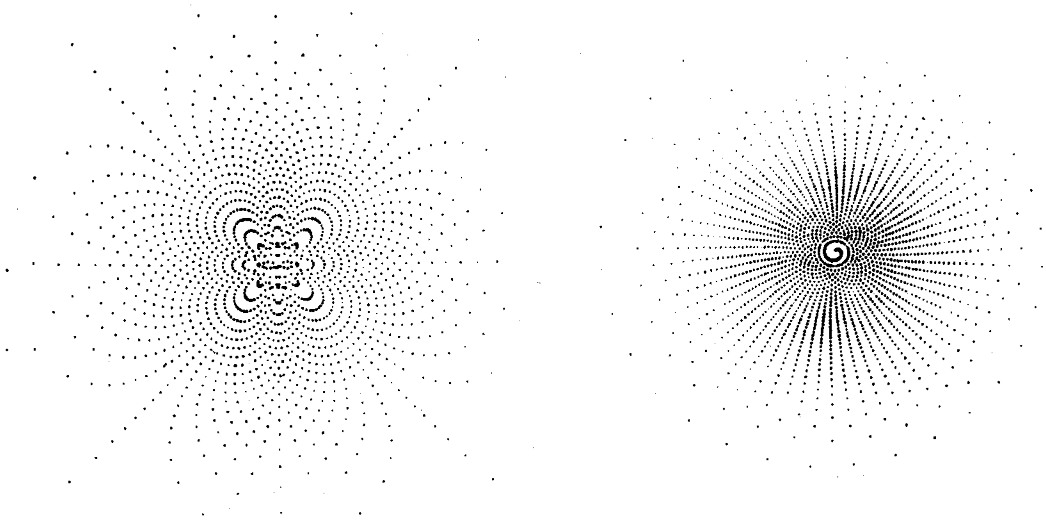
So sind offensichtlich x_1 und x_2 nicht zufällig hergestellt. x_3 erscheint dagegen zufälliger; die Folge ist schwieriger zu beschreiben. Bei genauerem Studium stellt sich heraus, daß es sehr schwer ist, sich lange zufällige Folgen auszudenken. Trotz des unterschiedlichen Grades an Zufälligkeit sind alle 3 Folgen in einem echten Münzwurfexperiment gleich wahrscheinlich. Offenbar gilt nämlich, daß die Wahrscheinlichkeit gegeben ist durch die äußerst kleine Größe

$$P(x_1) = P(x_2) = P(x_3) = \frac{1}{2^{24}}.$$

Damit kommt man zu folgendem Vergleich der Begriffe, dem die Entropie als Maß für die Information noch beigefügt ist:

Wahrscheinlichkeit	—	Entropie	—	Zufall
quantitative Erfassung	—	Information	—	Beschreibungskomplexität

Zufall wird in Simulationen deterministisch erzeugt durch iterative Anwendung von einfachen Rechenverfahren. Es gibt wohlbekannte Probleme dieser Konstruktionsverfahren. Es entstehen Muster in den Daten.



Plot von Paaren (X,Y) des polaren Algorithmus $X_i = (65X_{i-1} + 1) \bmod 2048$ und des Box-Muller-Algorithmus.

Die Entdeckung von unerwünschten Strukturen in Zufallszahlengeneratoren ist eine schwierige Aufgabe. Ein ähnliches Problem ensteht in der **Datenanalyse.** Man betrachte etwa 500 Befragungsdaten $x_1, \ldots, x_{500}$, jede in 20 Dimensionen. Man möchte in den Daten Muster, Strukturen entdecken. Auf dem Bildschirm sind aber nur ein oder zwei Projektionen darstellbar. Die Frage ist dann, wie sich eine 'grand Tour' durch die Projektionen erzeugen läßt, die geeignet ist, solche Muster zu finden. Das zur Zeit effektivste Verfahren hierfür basiert auf zufälligen Matrizenprodukten (Reflexionen). Dieses Verfahren ist auch in der Kryptographie von Bedeutung.

Ein wichtiger Aspekt der Stochastik ist die Aufgabe der **Entscheidung unter Unsicherheit und Wahrscheinlichkeit.** Mit dem Begriff der Wahrscheinlichkeit wird ein bedeutendes Mittel zur Analyse von Entscheidungssituationen unter Unsicherheit gegeben.

Beim Glücksspiel waren die Wahrscheinlichkeiten *a priori* gegeben, etwa durch die Voraussetzung, daß für alle Würfe mit einem Würfel die Chancen (für eine 1, 2, ..., oder 6) gleich seien. Bernoulli sieht nun die Möglichkeit, Wahrscheinlichkeiten *a posteriori* (aus statistischen Daten) zu bestimmen. Das gibt der Wahrscheinlichkeitsrechnung ganz neue Perspektiven für ihre Anwendung. Wichtig ist an dieser Stelle das Bernoullische *Gesetz der großen Zahlen*, über das er sagt:

> "Man muß ... in Betracht ziehen, woran vielleicht niemand bisher gedacht hat. Es bleibt nämlich noch zu untersuchen, ob durch Vermehrung der Beobachtungen ständig auch die Wahrscheinlichkeit dafür wächst, daß die Zahl der günstigen zur

Zahl der ungünstigen Beobachtungen das wahre Verhältnis erreicht, und zwar in dem Maße, daß diese Wahrscheinlichkeit schließlich jeden beliebigen Grad der Gewißheit übertrifft."

Anhand von einigen Beispielen soll diese Problematik erläutert werden.

a) Münzwurf und Prognosen 10.000 Würfe werden mit einer fairen Münze mit 'Erfolgs'wahrscheinlichkeit $p = \frac{1}{2}$ für Wappen $\widehat{=}$ 1 und Zahl $\widehat{=}$ 0 durchgeführt. Als Ergebnis erhält man einen Vektor $x = (x_1, \ldots, x_{10.000})$ aus Nullen und Einsen. $S = \sum_{i=1}^{10.000} x_i$ ist die Anzahl der Wappen. Ohne Stochastik läßt sich nur die offensichtliche Aussage $0 \leq S \leq 10.000$ treffen. Mit den Regeln der Wahrscheinlichkeitsrechnung ergibt sich jedoch präziser, daß gilt

$$\begin{aligned} P(4.900 \leq S \leq 5.100) &\approx 0,95 \\ P(4.850 \leq S \leq 5.150) &\approx 0,99. \end{aligned}$$

Mit großer Wahrscheinlichkeit liegt also die Anzahl der Wappen in der Nähe von 5.000. Dieses Ergebnis ist grundlegend für die **Möglichkeit der Prognosen**. Es erklärt auch die oben angesprochene Stabilisierung der Geschlechterverhältnisse bei Neugeborenen, denn mit der Setzung (0 $\widehat{=}$ Mädchen, 1 $\widehat{=}$ Knabe) liefert dieses Ergebnis mit großer Wahrscheinlichkeit, daß in großen Populationen vom Umfang n die relative Häufigkeit von Knaben ungefähr $\frac{1}{2}$ ist,

$$\frac{S}{n} \approx \frac{1}{2}.$$

Weitere Interpretationen (1 $\widehat{=}$ Erfolg, 0 $\widehat{=}$ Mißerfolg) sind von Bedeutung bei der Untersuchung der Wirksamkeit von Medikamenten, beim Vergleich von Behandlungsmethoden, etc.

Es werde z.B. durch eine Marktumfrage für ein neues Produkt festgestellt, daß das Marktsegment an potentiellen Kunden den Umfang 10.000 hat. Mit der Wahrscheinlichkeit 1/10 kaufen potentielle Kunden aus diesem Segment das Produkt. Wieviele Exemplare des Produktes sollte die Herstellerfirma sinnvollerweise herstellen bzw. vorrätig halten? Die obige Größe S entspricht nun den positiven Kaufentscheidungen. Aus der Beziehung $P(S \geq 1.090) \approx 0,0013$ ergibt sich, daß etwa 1.090 vorrätig gehaltene Produkte ausreichen, um mit großer Wahrscheinlichkeit den Bedarf abzudecken. Auf demselben Prinzip basieren auch Stichprobenverfahren, die mit großem Erfolg z.B. bei Wahlprognosen verwendet werden.

b) Situationsanalysen Es gibt eine Reihe von Regeln dafür, sich in der Welt des Zufalls sicher zu bewegen. Ein wichtiges derartiges Hilfsmittel ist die **Bayessche**

Regel. Gegeben seien zwei Typen von Ereignissen: $A, B_i, \ \ 1 \leq i \leq n$. Dann gilt für die bedingten Wahrscheinlichkeiten $P(A|B_i), P(B_i|A)$ die Beziehung:

$$P(B_i|A) \ = \ P(A|B_i)\frac{P(B_i)}{\sum_{k=1}^{n} P(B_k)P(A|B_k)}$$

Wenn die bedingten Wahrscheinlichkeiten von einem Typ bekannt sind, dann kann man die von dem anderen Typ aus obiger Formel ermitteln.

Eine Anwendung: Ein Patient wird mit einem Test auf eine Erkrankung hin untersucht.

		Patient			Test
K	$\widehat{=}$	krank	p	$\widehat{=}$	Test positiv
G	$\widehat{=}$	gesund	n	$\widehat{=}$	Test negativ

Wir nehmen an, der Test sei zuverlässig im folgenden Sinn. Wenn der Patient krank ist, dann ist der Test mit großer Warhscheinlichkeit positiv, bei gesundem Patienten mit großer Wahrscheinlichkeit negativ.

$$\begin{aligned} P(p|K) &= 0,94 \\ P(n|G) &= 0,96 \end{aligned}$$

Darüberhinaus sei die Häufigkeit der Erkrankung bekannt: $P(K) \ = \ \frac{1}{130}$. Vom Standpunkt des Patienten überraschend ist nun die erfreuliche Einsicht, die sich aus der obigen Bayesschen Regel ergibt:

$$P(K|p) \ = \ 0,10.$$

Trotz positiven Ergebnisses des Tests ist die Wahrscheinlichkeit einer Erkrankung sehr gering. Dieses Ergebnis läßt sich qualitativ auch mit gesundem Menschenverstand erzielen. Die Stochastik hat in diesem Zusammenhang also die Funktion, solche Analysen quantitativ und sicher durchzuführen.

Weitere Anwendungen der Bayesschen Formel betreffen in ähnlicher Weise die Kommunikationstheorie, Übertragungskanäle, Bildverarbeitung und viele andere. Vor etwa 10 Jahren wurde mit diesem Verfahren, angewandt auf das Problem der Autorenidentifikation, ein neuentdecktes Werk von Shakespeare identifiziert.

c) Challenger-Katastrophe (eine statistische Analyse, vgl. JASA 84, 945-957)

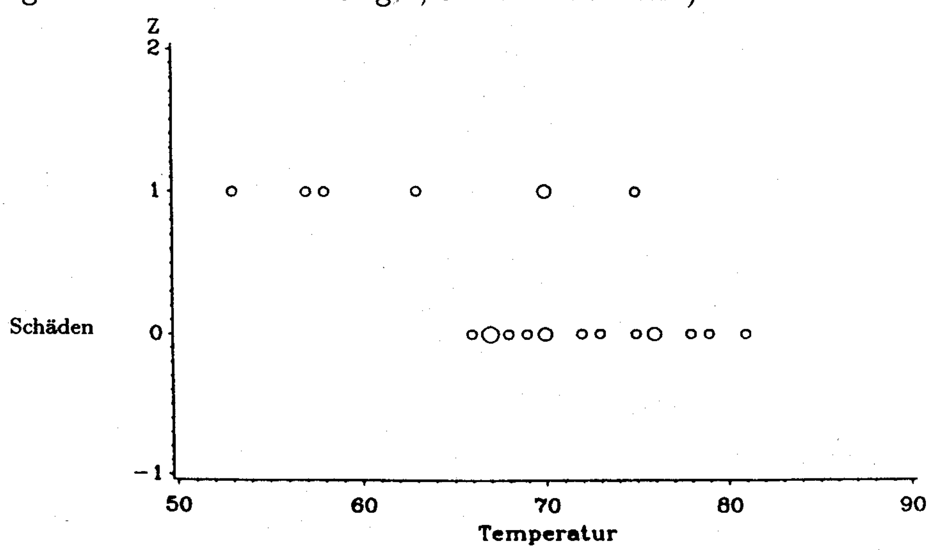
Am Morgen des 28. Januar 1986 explodierte die Raumfähre Challenger mit 7 Astronauten an Bord wegen Materialermüdung eines Dichtungsringes an den Triebwerken. Die Raumfähre verfügte über insgesamt 6 solcher Ringe. In der Nacht zum 28. Januar hatten die Experten des Triebwerke-Herstellers Morton Thiokol mit den

Experten der NASA und des Marshall Space Centre eine mehrstündige Konferenz, die sich wesentlich auf die Wettervorhersage von niedrigen 31° Fahrenheit Außentemperatur für die Startzeit der Raumfähre am nächsten Morgen und den Effekt auf die Zuverlässigkeit der Dichtungsringe konzentrierte. Die folgenden Daten spielten eine wichtige Rolle. Sie geben die Flüge an, bei denen solche Materialermüdung aufgetreten war, sowie die jeweiligen Außentemperaturen.

Flug-Nr.	Datum	Temperatur (in F)
STS-2	12.11.81	70°
41-B	03.02.84	57°
41-C	06.04.84	63°
41-D	30.08.84	70°
51-C	24.01.85	53°
61-A	30.10.85	75°
61-C	12.01.86	58°

Die Experten kamen zu dem Schluß, daß auf Grund der Daten kein Einfluß der Außentemperatur auf die Zuverlässigkeit der Ringe nachzuweisen sei. Trotz einiger Widersprüche empfahl Thiokol der NASA den planmäßigen Start des Space Shuttle.

In einer von Präsident Reagan eingesetzten Untersuchungskommission wurde festgestellt, daß ein Fehler in der Analyse der Materialermüdungsdaten war, daß die Flugdaten ohne technische Probleme aus der Diskussion des Einflusses der Außentemperatur herausgenommen wurden. Der vollständige Datensatz ist wie folgt (1 $\hat{=}$ Versagen mindestens eines Ringes, 0 $\hat{=}$ kein Problem):



Scatterplot der O-Ringe-Daten; größere 'Kreise' bedeuten zwei bzw. drei identische Beobachtungen.

Wie aus dieser Tabelle ersichtlich ist, finden die problematischen Flüge eher bei niedrigen Temperaturen statt. Eine routinemäßige statistische Analyse bestätigt diese Bedenken. Es sollte jedem Studenten der Statistik klar sein, daß die problematischen Flugdaten allein nicht 'suffizient' für das obige Problem sind.

Ein geeignetes Modell (für binäre Daten) ist das **Logit Modell.** Sei $\alpha(x)$ die Wahrscheinlichkeit für das Versagen mindestens eines Dichtungsringes bei Temperatur x und setze

$$\log \frac{\alpha(x)}{1-\alpha(x)} = \beta_0 + x\beta_1, \quad \text{äquivalent} \quad \alpha(x) = \frac{\exp(\beta_0 + x\beta_1)}{1+\exp(\beta_0 + x\beta_1)},$$

Dann erhält man aus der Maximum-Likelihood-Schätzmethode als Schätzwerte für die obigen Parameter β_0, β_1 zusammen mit den Varianzen $\widehat{\sigma}_0$ und $\widehat{\sigma}_1$

$$\begin{aligned} \widehat{\beta}_0 &= 15,04 & \widehat{\beta}_1 &= -0,23 \\ \widehat{\sigma}_0 &= 7,3 & \widehat{\sigma}_1 &= 0,10. \end{aligned}$$

Dieses liefert bei der Temperatur x den Schätzwert $\widehat{\alpha}(x)$, so z.B. bei der Temperatur 31° $\widehat{\alpha}(31) = 0,99$ oder bei der Temperatur 70° $\widehat{\alpha}(70) = 0,23$. Die Indikation für das Versagen eines Dichtungsringes bei der Temperatur von 31° Fahrenheit ist also überdeutlich. Die Ungenauigkeit der Schätzung ist hoch, aber die Schätzwerte sind sicher ein deutliches Indiz zur Vorsicht.

III. Stochastische Modellbildung

Schon am Beispiel der Challenger-Daten läßt sich die Bedeutung von stochastischen Modellen erläutern. Das Modell enthält bedeutungsvolle Parameter, die sich aus den vorliegenden Daten schätzen lassen. Daraus lassen sich dann Konsequenzen ableiten, wie z.B. die Unbedenklichkeit eines Medikamentes, die Zuverlässigkeit eines Sicherheitssystems, etc. Auf passender Modellbildung basieren im wesentlichen die Fortschritte in den Anwendungen statistischer Verfahren in den verschiedenen Anwendungsgebieten. Bei statistischen Analysen können die verwendeten Modelle jedoch auch einen pragmatischen Charakter haben. Genauer lassen sich im wesentlichen drei Typen statistischer Analysen unterscheiden, nämlich

1.) Explorative Statistik (EDA). Die EDA verwendet nur in geringem Umfang Modellannahmen. Sie ist ein wichtiges Hilfsmittel zur Hypothesenbildung.

2.) Prädiktive Statistik. Ziel ist die Vorhersage und Klassifikation von verrauschten oder nicht beobachteten Daten, z.B. die räumliche Interpolation einer Funktion $y(x)$ bei beobachteten $y(x_1), \ldots, y(x_n)$ oder die Bildrekonstruktion aus einem verrauschten Bild.

Modelle spiegeln nur einige Eigenschaften der wirklichen Situation wider. Sie sind aus pragmatischen Gründen vorgenommen und dienen zur Verbesserung der Vorhersage.

3.) Konfirmative Statistik. Hier spiegeln Modelle die Realität wider. Die Parameter der Modelle reflektieren wichtige Eigenheiten der Phänomene. Die Untersuchungen haben einen hohen Grad von Objektivität und Verbindlichkeit.

Einige Beispiele sollen den Einfluß der Modellbildung erläutern.

a) Extremwertmodelle Die Extremwertstatistik befaßt sich mit der Analyse des Verhaltens von extremen Beobachtungen, Maxima oder Minima in einer Versuchsserie $x = (x_1, \ldots, x_n)$, also $M_n = \max(x_1, \ldots, x_n)$, $m_n = \min(x_1, \ldots, x_n)$.

Unter gewissen Bedingungen läßt sich dann M_n durch drei Typen von Verteilungen, die Extremwertverteilungen, beschreiben, die bis auf einen Skalen- und Verschiebungsparameter eindeutig bestimmt sind.Ein relevantes Anwendungsgebiet ist die Modellierung von Hochwasserständen bei Fluten. Ziel ist die Bestimmung der notwendigen Höhe von Dämmen zum Schutz gegen extreme Fluten. Relevante Fragen sind dann etwa:

1.) Wie hoch ist die Wahrscheinlichkeit, daß ein bestehender Deich in den nächsten 10 Jahren überspült wird? Veränderung des Flutverhaltens, z.B. Abschmelzprozesse bei Polareis?

2.) Wie hoch sollten die Deiche sein, damit in den nächsten 10 Jahren mit Wahrscheinlichkeit $\leq \frac{1}{10.000}$ das Hochwasser die Dämme nicht überspült?

Empirische Daten für die erste Frage liegen (wenn überhaupt) nur in geringer Anzahl vor, so daß eine direkte empirische Schätzung der Überschreitungswahrscheinlichkeit ausgeschlossen erscheint. Mit der Modellierung durch Extremwertverteilungen lassen sich jedoch die Parameter sinnvoll schätzen. Mit diesen Schätzwerten können begründete Prognosen gegeben werden. Solche Modelle werden insbesondere in den Niederlanden in umfangreichem Maße für obige Fragestellungen untersucht.

Weiteres Anwendungsgebiet dieser Modelle ist

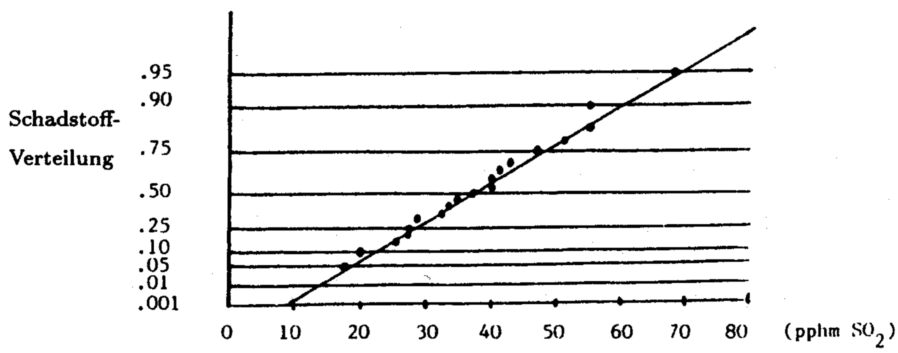
a1) Bestimmung des radioaktiven Zerfalles. Berechne M_n die Zeit bis zum vollständigen Zerfall von n Isotopen. Insbesondere nach der Tschernobylkatastrophe (26. April 1986) wurden solche Berechnungen vielfach durchgeführt. Es ist bekannt, daß die Lebensdauer von Isotopen exponentiell verteilt ist. Für 137 g Cäsium Cs^{137} erhält man hieraus für den Erwartungsert der vollständigen Zerfallszeit: $E\, M_n \approx 2.960$ Jahre. (Die Halbwertszeit $h = \frac{\log 2}{\lambda}$ beträgt bei Cs^{137} 37 Jahre.) Die Streuung ist $\sigma \approx 68$ Jahre, so daß die erwartete Zerfallszeit bis auf etwa $\pm$ 150 Jahre zutreffend ist.

a2) Schwefeldioxyd-Belastung der Luft. Die folgenden Daten über die maximale Schwefeldioxyd-Belastung der Luft stammen aus Long Beach über den Zeitraum von 1956 – 1974.

Jahr	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965
jährl. Max.	47	41	68	32	27	43	20	27	25	18
jährl. Ø	4,0	3,0	3,4	2,1	1,9	1,9	1,5	1,3	1,4	2,6

Jahr	1966	1967	1968	1969	1970	1971	1972	1973	1974
jährl. Max.	33	40	51	55	40	55	37	28	34
jährl. Ø	3,0	2,5	3,1	2,5	2,4	2,5	2,5	1,9	1,7

Angepaßt wird in diesem Fall ein Extremwertmodell mit der Extremwertverteilung $G(x) = e^{-e^{-(\sigma x+\mu)}}$, $x \in \mathbb{R}^1$. Mit der Methode der kleinsten Quadrate werden Skalenparameter der Verteilung geschätzt. Damit ergibt sich eine gute Anpassung des Modells an die maximalen Schadstoffkonzentrationen. Mit diesem Modell lassen sich Vorhersagen treffen, Trends und Änderungen des Konzentrationsverhaltens ermitteln.



Fit durch Extremwertmodell
(aus Pfeifer: Einführung in die Extremwertstatistik.)

Weitere Anwendungsgebiete der Extremwertmodelle sind in **Materialbeständigkeitsuntersuchungen**, (z.B. bei der Zugfestigkeit von Blechen in der Autoindustrie) und in der **Zuverlässigkeitsanalyse komplexer Systeme** (z.B. Sicherheitsberechnungen von Kraftwerken) zu finden.

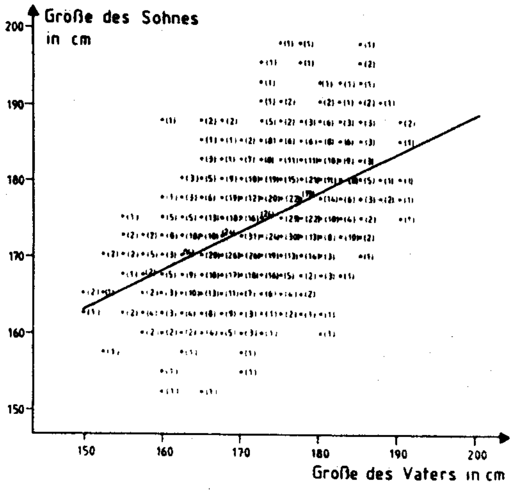
b) Regressionsprobleme Dieses sind **die** klassischen und wahrscheinlich am häufigsten angewendeten statistischen Modelle. Sie sind von der Form:

$$y_i = f(x_i) + \varepsilon_i, \quad f \in \mathcal{F}.$$

Ein wichtiger Spezialfall sind die linearen Modelle, bei denen die Funktionen linear sind. Die Bedeutung der Variablen ist:

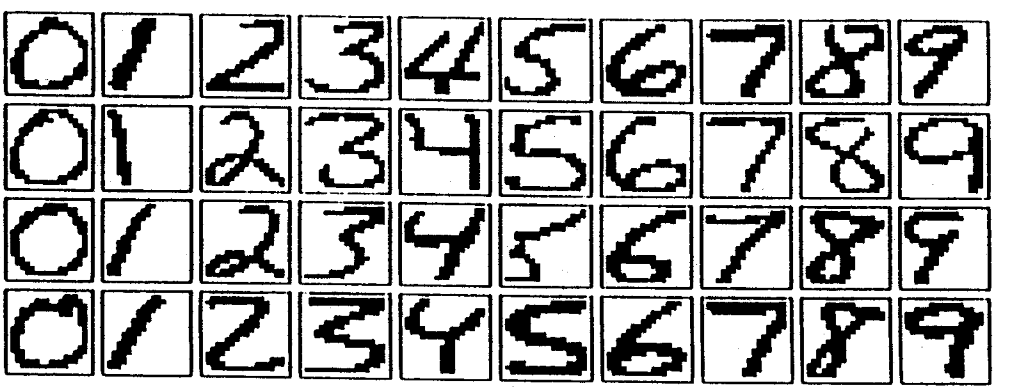
$y_i \mathrel{\widehat{=}}$ Beobachtungsgröße, $x_i \mathrel{\widehat{=}}$ Einflußgrößen, $\varepsilon_i \mathrel{\widehat{=}}$ Zufallsfehler.

Regressionsprobleme sind typische inverse Probleme. Das Gebiet der Versuchsplanung ist eng mit diesen Modellen verknüpft. Ein klassisches Beispiel ist die Regression der Größe von Erbsen nach der Größe der eingesetzten Samen, oder auch die Regression der Körpergröße von Söhnen nach der Größe ihrer Väter. (Galton-Pearson)



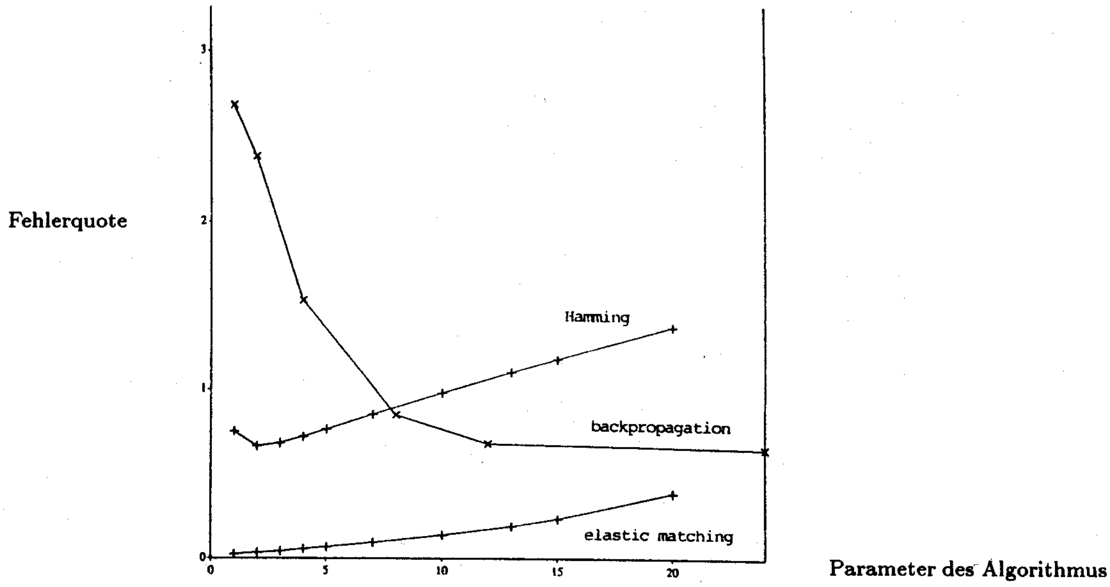
Galton-Pearson-Regression: Zusammenhang zwischen der Größe von Vätern (x) und ihren Söhnen (y), $y = 85,6742 + 0,561x$

c) Klassifikationsprobleme Regressionsmodelle finden Anwendung auch bei Klassifikationsproblemen. Einige wichtige Beispiele sind etwa die Ziffern- und Handschriftenerkennung.



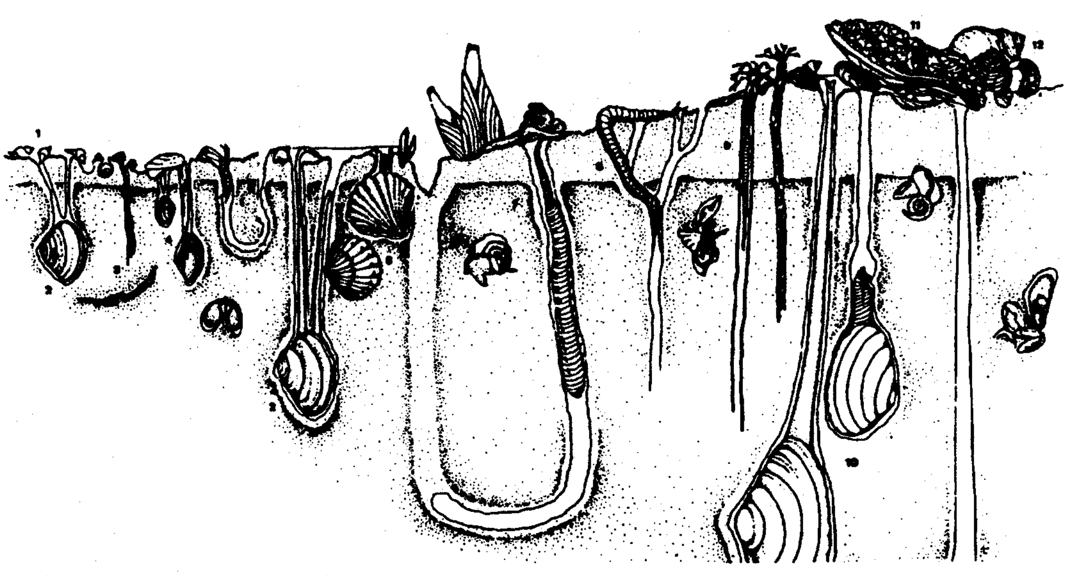
handgeschriebene Zahlen

Es gibt eine Fülle von Verfahren für dieses Problem wie z.B. die nearest neighbour Methode (NN), neuronale Netze, wavelets, projection pursuit und Deformationsdistanzen.



Vergleich von 3 Algorithmen, optimale Fehlerquote $\approx 1\%$,
(aus: Geman, Bienenstock, Doursat, in: Neural Networks Nr. 1, 1992)

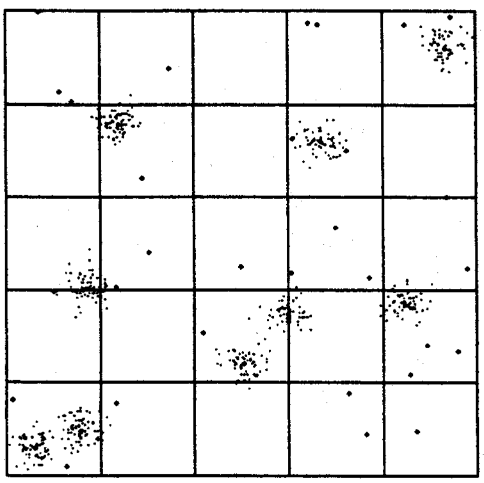
Ein umfangreicher Vergleich dieser Algorithmen finden im STATLOG-Projekt im Rahmen des ESPRIT-Programmes statt, an Hand von Beispielen wie: Shuttle Control und Stabilität, Landsat Daten, Kreditmanagement, Autosilhouettenerkennung, Buchstabenerkennung, DNA-Sequencen, Diabetes, Energieversorgung und Personenidentifikation.



Lebensraum Watt

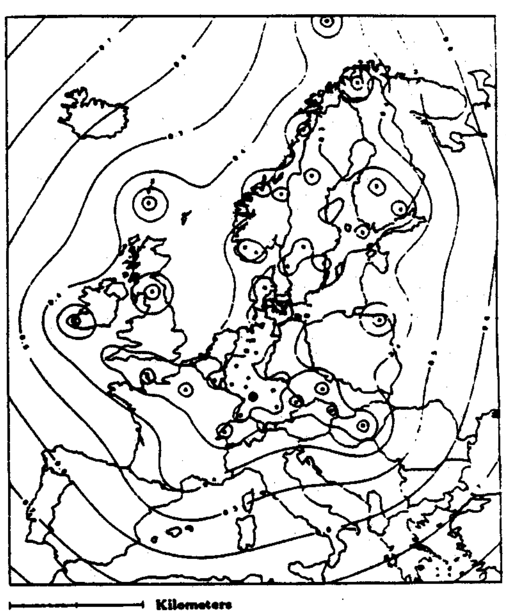
1: Wattschnecke *(Hydrobia ulvae)*
2: Pfeffermuschel *(Scrobicularia plana)*
3: Kotpillenwurm *(Heteromastus filiformis)*
4: Tellmuschel *(Macoma baltica)*
5: Schlickkrebs *(Corophium volutator)*
6: Herzmuschel *(Cerastoderma edule)*
7: Pierwurm *(Arenicola marina)*
8: Seeringelwurm *(Nereis diversicolor)*
9: Bäumchen-Röhrenwurm *(Lanice conchilega)*
10: Sandklaffmuschel *(Mya arenaria)*
11: Miesmuschel *(Mytilus edulis)*
12: Gemeine Strandschnecke *(Littorina littorea)*

Im ELAWAT-Projekt aus der Landschaftsökologie werden räumliche Regressionsmodelle (Straussmodelle) zur Bestimmung von räumlichen Entwicklungsmustern von Populationen und zur Klassifikation und Ausbreitung von Schadstoffen angepaßt. Für die Ausbreitung von Schnecken, Muscheln und Würmern im Wattenmeer wird ein Punktprozessmodell (Straussmodell) angepaßt. Die geschätzten Intensitäten enthalten wichtige Informationen über den Zustand und die Veränderung der Population.



3	3	0	2	56
4	65	0	65	0
28	7	1	4	8
26	3	81	29	58
142	6	5	2	1

Straussmodell und Schätzwerte der Intensitäten von Eltern- und Tochterprozess für Bäumchen-Röhrenwurm (aus Studie D. Pfeifer (1995): Lebensraum Watt.)



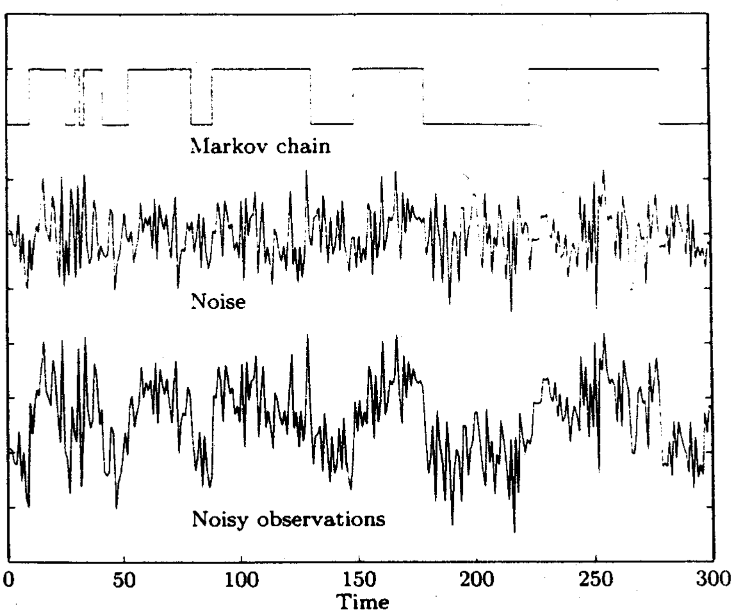
Interpolierte log-Schwefelkonzentration Januar 1984 mit geschätzten Abweichungen (aus: Hjort, Omre (1994): Topics in spatial statistics. In: Scand. J. Statistics)

Im European Monitoring Evaluation Program (EMEP) werden Luftverschmutzungsdaten (Schwefeldioxyd) in 100 Stationen aufgezeichnet. Aus den Meßergebnissen wird mit Bayesschen Glättungsverfahren, angewendet auf ein räumliches Regressionsmodell, eine Luftverschmutzungskarte erstellt. Ziel ist es, regionale Trends zu bestimmen und die Zu- bzw. Abnahmen der Luftverschmutzung zu erkennen.

d) Hidden Markovmodelle (HMM) Hidden Markovmodelle sind wie die Regressionsmodelle von der Form

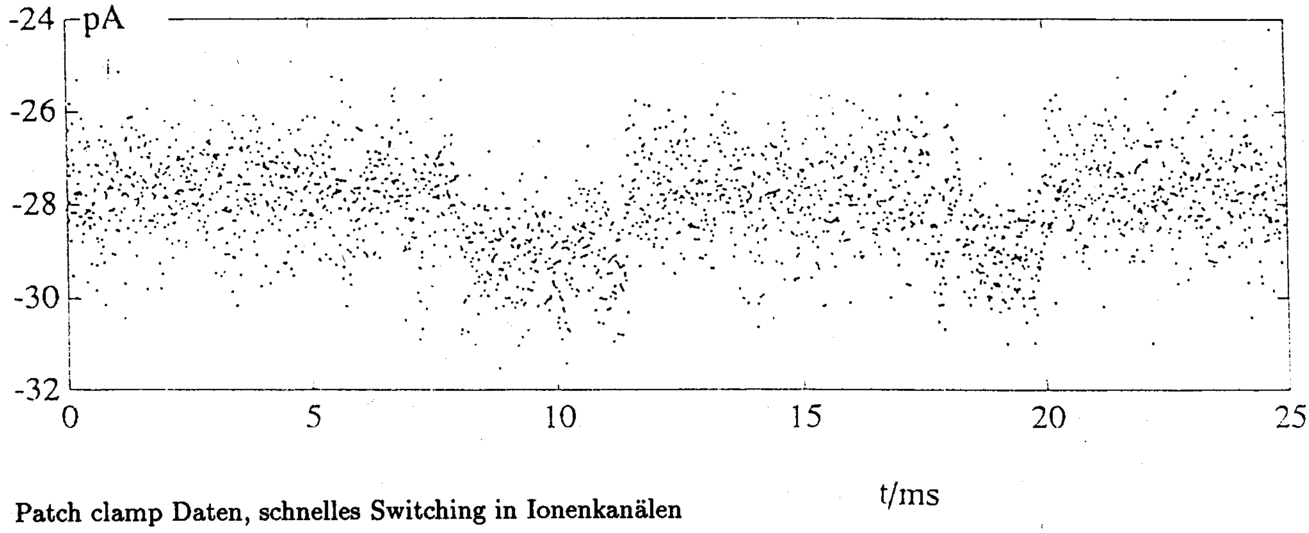
$$y_i = f(x_i) + \varepsilon_i, \qquad f \in \mathcal{F}.$$

Die (x_i) unterliegen hierbei jedoch einer nicht direkt beobachtbaren Markovschen Dynamik. Dieses ergibt einen zusätzlichen wichtigen Freiheitsgrad bei der Anpassung der Modelle. HMM sind die grundlegenden Hilfsmittel in den aktuellen Systemen zur Spracherkennung. Die Daten sind die Frequenzdaten der Sprache; die Dynamik der Phoneme wird durch eine Markovkette beschrieben.



Binäre Markovkette mit überlagertem Rauschen

Dieses Modell ist auch von Bedeutung in der Konstruktion von DNA-Abbildungen im Genomprojekt sowie in aktuellen Ansätzen in der Molekularbiologie zur Klassifikation von Ionenkanälen.



Patch clamp Daten, schnelles Switching in Ionenkanälen

Anmerkung: Den Kollegen D. Pfeifer (Hamburg), W. Hansen (Kiel), M. Schumacher, J. Timmer und insbesondere H. Witting (Freiburg) danke ich für Hinweise und die Bereitstellung von Materialien, die in diesem Artikel verwendet wurden, sowie für zahlreiche Vorschläge zur Organisation und Darstellung.