Zur Gruppierung mehrvariabler Beobachtungen

Vorbemerkung

Der vorliegende Beitrag ist Teil einer größeren Studie, die seit den Jahren 1970/71 ausgeführt wird, ohne daß es bisher zu einer Veröffentlichung gekommen wäre (1). Deshalb soll hier ein methodisches (quantitatives) Problem herausgegriffen und vorab dargestellt werden. Ich bin Prof. G. Styan, Dept of Mathematics, McGill University, für seine Anregungen und Hilfe zu großem Dank verpflichtet.

Zur Problemlage

Zum Studium der Unterscheidbarkeit einzelner Baumarten (2) auf Farbluftbildern wurden unter anderem Densitometer-Messungen an Farbnegativ-Transparenten durchgeführt, und zwar an einer Stichprobe, die 148 Einzelbäume umfaßte. Diese Objekte gliedern sich in 15 Gruppen, nämlich die gewählten 15 Laubbaumarten (3). Die Artbestimmung jedes Baumes erfolgte durch Bodenkontrolle. Die Ektachrome-Luftbilder im Maßstab 1 : 5000 wurden im Juni 1969 aufgenommen. Es sollte untersucht werden, ob sich die 15 Baumarten in ihrer Farbe (auf dem Negativ-Transparent) zuverlässig unterscheiden lassen. Ziel des Projektes ist eine Klassifikationsprozedur für Baumarten, aufgrund von Farbmessungen auf Luftbildern.

Meßresultate

Die Densitometer-Messungen erfolgten mit einem Macbeth TD-404 (4), und zwar resultierten für jedes Objekt vier Meßwerte, nämlich die Dichte je im Rot-, Grün- und Blaubereich sowie über den gesamten Spektralbereich 400–700 μm. Damit ergibt sich ein vierdimensionales Modell mit 148 Punkten, die 15 Gruppen angehören sollen. Diese Hypothese wurde statistisch überprüft.

Varianzanalyse

Die Nullhypothese einer Varianzanlyse (5) behauptet, daß alle g beteiligten Gruppen im Modell Stichproben einer einzigen Grundgesamtheit sind, sich also statistisch nicht unterscheiden. Wird die Nullhypothese verworfen, so bedeutet dies hingegen nicht direkt, daß g Grundgesamtheiten vorhanden sind, sondern lediglich, daß es sich um mehr als eine Population handelt. Zum Testen der Nullhypothese werden die Gruppen- und Zwischengruppendistanzen (genauer: die Distanzquadratsummen [6]) berechnet; daraus ergibt sich als Kriterium Wilk's L-Wert (7), definiert als

$$L = \frac{|W|}{|W+B|} = \frac{|W|}{|T|}$$

wobei W = Gruppendistanz = innere Distanzquadratsumme
B = Zwischengruppendistanz = äußere Distanzquadratsumme
T = Gesamtdistanz = gesamte Distanzquadratsumme

Die Matrizen W, B und T sind die natürlichen mehrvariablen Erweiterungen der eindimensionalen Distanzquadratsummen einer Stichprobe:
$S = \sum (x_i - \bar{x})^2$.
Die Signifikanz des resultierenden L-Wertes wird mittels einer F- oder einer χ^2-Approximation (8) getestet, wobei gilt:

$$V = -\ln L \cdot m \qquad V \frown \chi^2_{\nu_1}$$

$$R = \frac{1 - L^{1/s}}{L^{1/s}} \cdot \frac{\nu_2}{\nu_1} \qquad R \frown F^{\nu_1}_{\nu_2}$$

Dabei sind n = Anzahl Beobachtungen (hier: 148)
p = Anzahl Variablen (hier: 4)
g = Anzahl Gruppen (hier: 15)
q = g – 1
m = n – 1 – (p+g)/2
$s = \sqrt{(p^2 q^2 - 4) / (p^2 + q^2 - 5)}$
$\nu_1 = p\,q$
$\nu_2 = m\,s - p\,q/2 + 1$

Im vorliegenden Fall der 148 Messungen an Baumkronen ergibt sich ein L-Wert von 0.345, entsprechend $\chi^2_{(56)} = 146.2$ und $F_{(56,508)} = 2.85$. Diese Resultate sind für $\alpha = 0.1\%$ signifikant, d. h. die Nullhypothese muß (mit einer sehr kleinen Irrtumswahrscheinlichkeit) verworfen werden. Die 148 Messungen entstammen also nicht einer einzigen Grundgesamtheit.

Dr. Martin Boesch, Bruggwiesenweg 20a, 9000 St. Gallen.

Klassifikationsanalyse

Auf die anschließende Klassifikationsanalyse (9) soll hier nicht im Detail eingegangen werden. Sie zeigt jedenfalls, daß von den 148 Testdaten nur 42 (d. h. 28%) der korrekten Gruppe zugeordnet werden können. Dieses Ergebnis ist dahingehend zu interpretieren, daß zwar wohl mehr als eine echte Gruppe, keinesfalls aber deren fünfzehn vorliegen. Im vierdimensionalen Raum überlappen sich die einzelnen Gruppen zum Teil beträchtlich. Damit erhebt sich die Frage, wieviele echte Gruppen vorliegen, und wie sie gefunden werden können.

«Pooling»-Analyse

Für die Lösung dieses Problemes, unechte «Gruppen» zu echten zusammenzufassen, gibt es keine Standardprozedur (10). Die bekannten Distanzgruppierungsmethoden gehen von Einzelpunkten aus, die nicht a priori einer Gruppe angehören. Ein solches Zusammenfassen erscheint aber als sinnvoll, indem danach wirklich distinkte Gruppen vorliegen, die sich auch als Grundlage für weitere Klassifikationen (das eigentliche Endziel des Projektes) eignen. Zur Lösung des Problems wird die folgende Prozedur (11) vorgeschlagen.

Ausgangspunkt der Zusammenfassung ist die Idee, daß schrittweise diejenigen zwei unechten Gruppen zusammengefaßt werden sollen, die sich aufgrund einer paarweisen Varianzanalyse als am ähnlichsten (bei allen möglichen Paarungen) erweisen. Dabei soll als Ähnlichkeitskriterium der oben definierte χ^2-Wert (12), dem die Varianzen innerhalb und zwischen den Gruppen zugrunde liegen, verwendet werden.

Der Zusammenfassungsprozeß kommt dann zum Abbruch, wenn die Ähnlichkeit zwischen den verbleibenden Gruppen (deren Zahl ja bei jedem Schritt um eine reduziert wurde) ein vorgegebenes Maß unterschreitet. Es liegen dann Gruppen vor, die als echt angesprochen werden können, indem sie je eine Grundgesamtheit repräsentieren. Selbst dann muß aber immer noch damit gerechnet werden, daß die Klassifikation der Testdaten nicht vollkommen korrekt erfolgen wird, treten doch auch bei echten Gruppen immer noch Überlappungen auf.

2	9.3													
3	15.1	6.4												
4	12.3	1.8	5.6											
5	5.8	22.0	12.6	14.1										
6	8.3	7.5	7.9	6.0	14.4									
7	2.2	7.2	5.7	6.6	5.7	7.9								
8	8.3	16.4	10.5	12.6	1.7	10.9	5.1							
9	1.3	11.4	18.4	17.6	8.6	12.2	4.9	12.0						
10	6.5	2.6	5.4	1.3	7.4	3.4	3.0	6.7	10.3					
11	7.5	4.1	11.1	9.4	14.0	3.3	7.0	12.0	10.6	4.4				
12	10.3	18.8	9.8	13.7	3.0	10.1	8.7	4.6	14.5	8.1	11.4			
13	4.6	6.7	16.0	3.2	16.6	4.0	8.2	16.4	7.0	6.3	2.2	12.8		
14	3.3	16.8	17.6	14.6	18.6	10.0	8.6	16.9	5.6	12.8	12.6	10.9	8.3	
15	3.3	14.8	17.5	18.4	7.4	13.9	3.3	9.1	3.9	10.3	15.3	11.8	7.0	3.2
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Abb. 1: Die χ^2-Werte für paarweise Gruppierung bei 15 Ausgangsgruppen

Abb. 1 zeigt die ursprünglichen 105 χ^2-Werte, die den Paarungen zwischen je zwei Gruppen entsprechen. Es ist ersichtlich, daß einige dieser Werte sehr klein sind, was auf große Ähnlichkeit der beiden beteiligten Gruppen schließen läßt.

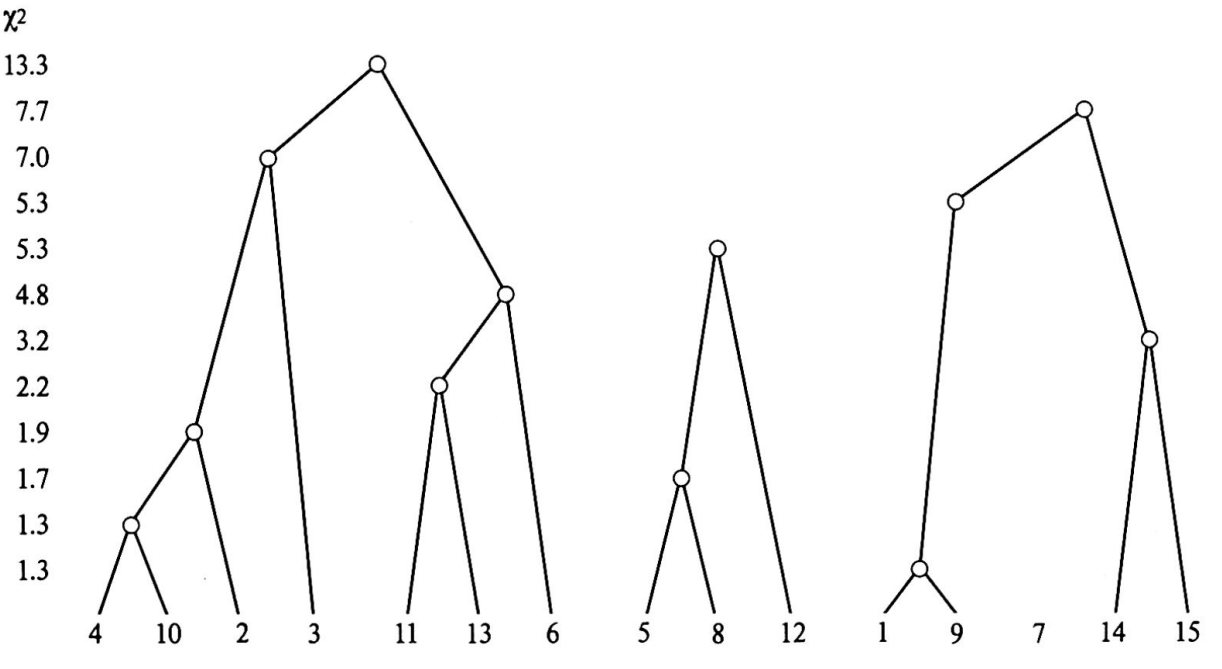
Die Zusammenfassung im ersten Schritt betrifft nun die zwei Gruppen Nr. 1 und Nr. 9, welche den tiefsten χ^2-Wert aufweisen. Damit liegen 14 Gruppen vor; die W- und B-Matrizen müssen z. T. neu berechnet werden, was Veränderungen in der L- und entsprechend in der χ^2-Matrix ergibt. Nach diesem Durchlauf beginnt der Zusammenfassungsprozeß von neuem.

Die Zusammenfassung zu neuen Gruppen wird fortgesetzt, bis nurmehr einige wenige Gruppen vorliegen. Auf jeder dieser Ebenen muß entschieden werden, ob der Prozeß abgebrochen werden soll. Dieser Entscheid ist ziemlich willkürlich, da ein eindeutiges Abbruchkriterium fehlt. Am naheliegendsten ist es, wiederum die χ^2-Werte dafür heranzuziehen. Im vorliegenden Fall sieht dies so aus:

Gruppen	Kleinster χ^2-Wert	Zugehöriger kritischer Wert α	Korrektheit der Klassifikationsanalyse
5	7.69	10.4%	51%
4	13.25	1.01%	56%
3	30.60	$< 0.01\%$	69%

Bei fünf verbleibenden Gruppen muß also mit einer Irrtumswahrscheinlichkeit von über 10% gerechnet werden, gegenüber 1% bei vier Gruppen und weit unter 0.01% bei drei Gruppen. Der Entscheid wird auf vier verbleibende Gruppen fallen. Nun kann aber auch die Korrektheitsrate (13) einer Klassifikationsanalyse der Testdaten als Abbruchkriterium herangezogen werden. Dabei ergibt sich, daß der Schritt von fünf auf vier Gruppen nur eine Verbesserung von 51 auf 56% bringt, währenddem eine weitere Zusammenfassung zu drei Gruppen diese Rate der korrekten Klassifikation der Testdaten auf 69% steigen läßt. Stehen also eher theoretisch-statistische Momente im Vordergrund

Abb. 2: Dendrogramme des Zusammenfassungsprozesses.
Die Baumarten-Nummern beziehen sich auf die Liste in Anmerkung 3.



der Überlegungen, so dürfte eine Zusammenfassung in vier Gruppen gegeben sein, währenddem beim Überwiegen des Klassifikationsproblems eine Reduktion auf lediglich drei Gruppen angezeigt wäre.
Abb. 2 zeigt das Dendrogramm des Zusammenfassungsprozesses in unserem konkreten Fall. Die Baumarten-Nummern beziehen sich auf die Liste in Anmerkung 3. Ferner sind die χ^2-Werte angegeben, welche die Ähnlichkeit der jeweils zusammengefaßten Gruppen charakterisieren. Dabei sind die folgenden Kritischen Werte zu beachten:

α(%)	50	20	10	5	1	0.1
χ^2 (4)	3.36	5.99	7.78	9.49	13.28	18.47

Interpretation der Ergebnisse

Obschon (wie eingangs erwähnt) hier v. a. die methodischen Aspekte des Problems erörtert werden sollen, sei doch kurz auf die inhaltliche Bedeutung der Ergebnisse eingegangen. Die «Farben» der Baumkronen, gegeben als vierdimensionale Meßpunkte für jedes Objekt, sind der Ausgangspunkt der Untersuchung. Die ersten Schritte (Varianzanalyse, Klassifikationsanalyse) zeigen, daß zwar Farbunterschiede zwischen den einzelnen Baumarten bestehen, daß sie aber nicht ausreichen, um alle fünfzehn Arten voneinander zu unterscheiden.
In der «Pooling»-Analyse wird dann versucht, Baumarten so zusammenzufassen, daß distinkte Gruppen entstehen. Dabei ist wiederum die Farbe das alleinige Kriterium. Es muß also erwartet werden, daß sich zwar die resultierenden Gruppen von Baumarten bezüglich Farbe unterscheiden, daß aber in jeder anderen Hinsicht (nämlich nach ökologischen, morphologischen, phänologischen oder gar taxonomischen Gesichtspunkten) die neue Gliederung nicht sinnvoll ist. Wie der Liste der 15 Baumarten, geordnet nach der vorgeschlagenen Gliederung, zu entnehmen ist (Abb. 3), hat sich diese Annahme bestätigt. Anderseits zeigt eine Bestimmung des Farbtones nach dem Munsell Code (14), daß sich die Gruppen tatsächlich hinsichtlich Farbe der Baumkronen unterscheiden. Damit ist eine zusätzliche Entscheidungshilfe geboten bei der Bestimmung der Baumarten ab Farbluftbildern.

Abb. 3: Die neuen Gruppen. - Farbwerte nach Munsell Code

Baumarten	5 Gruppen	4 Gruppen	3 Gruppen
Basswood Yellow birch White ash Slippery elm	5.0 R	5.0 R	5.0 R
Silver maple White elm Red maple	5.0 R	5.0 R	
Manitoba maple Black willow Black cherry	2.5 R	2.5 R	2.5 R
White birch Balsam poplar Trembling aspen	10.0 RP	10.0 RP	10.0 RP
Sugar maple Beech	7.5 RP		

Anmerkungen

1. Das Projekt Terrain Analysis System steht unter der Leitung von Prof. J. T. Parry, Dept. of Geography, McGill University. Über die Identifikation von Laubbaumarten liegen bisher nur interne Arbeitsberichte vor.
2. Die Auseinandersetzung mit der Literatur zeigt, daß das Problem der Baumartenunterscheidung (im Gegensatz zur Identifikation von landw. Kulturen, Waldgesellschaften, Böden usw.) noch wenig bearbeitet ist. Insbesondere wurden Farbluftbilder selten als Informationsquelle benutzt. Vgl. Lit. 1–12.
3. 1 White birch (Betula papyrifera)
   2 Basswood (Tilia americana)
   3 Yellow birch (Betula lutea)
   4 White ash (Fraxinus americana)
   5 Manitoba mable (Acer negundo)
   6 Silver maple (Acer saccharinum)
   7 Trembling aspen (Populus tremuloides)
   8 Black willow (Salix nigra)
   9 Balsam poplar (Populus balsamifera)
   10 Slippery elm (Ulmus rubra)
   11 White elm (Ulmus americana)
   12 Black cherry (Prunus serotina)
   13 Red maple (Acer rubrum)
   14 Sugar maple (Acer saccharum)
   15 Beech (Fagus grandifolia)
4. Macbeth Quantalog Instruments, Newsburgh/NY
5. Vgl. Lit. 15, p. 23ff u. a.
6. Vgl. Lit. 13, p. 264
7. Vgl. Lit. 15, p. 31. Der L-Wert ist auch unter der Bezeichnung Wilk's Λ (Lambda) bekannt.
8. Vgl. Lit. 16. p. 86. Eine Approximation ist deshalb nötig, weil der L-Wert selbst keiner Testverteilung folgt.
9. Vgl. Lit. 16, p. 217ff u. a. Es wurde eine klassische lineare Diskriminanz-Analyse angewendet.
10. Vgl. Lit. 13–16
11. Das entsprechende MATLAN-Programm kann beim Autor bezogen werden.
12. Dem χ^2-Kriterium ist hier der Vorzug zu geben, weil es von der Zahl der Datenpunkte unabhängig ist. Für die gesamte Matrix gilt der Freiheitsgrad 4, was Vergleiche direkt ermöglicht (d. h. ohne die Berechnung des jeweiligen kritischen Wertes). Daneben ist von berechnungsökonomischer Bedeutung, daß die χ^2-Approximation bedeutend einfacher ist als die F-Approximation.
13. D. h. der Anteil korrekt klassifizierter Punkte an der gesamten Stichprobe.
14. Vgl. Lit. 17–19. Die Umsetzung der Densitometer-Werte in Farbcode nach Munsell erfolgte durch ein vom Autor entwickeltes Computer-Programm, welches sich auf die Untersuchungen von Rib (Lit. 19) stützt.

Literatur

1. BERESIN, A. M. und VINOGRADOV, B. V.: Mikrofotografische Analyse der Abbildung wichtiger Baumarten der Taigazone auf großmaßstäbigen Luftbildern. 1961.
2. Forestry Branch, Dept. of Resources and Development: Canadian Woods, their properties and uses. Ottawa, 1951.
3. HELLER, R. C. et al.: Identification of tree species on large-scale panchromatic and color aerial photographs. US Dept of Agriculture, Forest Service, Agriculture Handbook No. 261. Washington D. C. 1964.
4. HOSTROP, B. W. und KAWAGUCHI, T.: Aerial color in forestry. in: Photogr. Eng. Bd. 37, Nr. 6, 1971.
5. JOHNSON, P. L. (Herausg.): Remote sensing in ecology. - Univ. of Georgia Press, Athens (Ga.), 1969.
6. KRUMPE, P F. et al.: The deliniation of forest cover and site parameters by multiband remote sensing. Papers 37th Annual meeting, Amer. Soc. of Photogrammetry, pp 98–122,1971.
7. NORTHROP, K. G. und JOHNSON, E. W.: Forest Cover Type Identification. in: Photogr. Eng. Bd. 36, Nr. 5, 1970.
8. PARRY, J. T. et al.: Color for Coniferous forest species. in: Photogr. Eng. Bd. 35, Nr. 7, 1969.
9. SAYN-WITTGENSTEIN, L.: Recognition of tree species on air photographs by crown characteristics. Canad. Dept. of Forestry, Technical Note No. 95, 1960.
10. SAYN-WITTGENSTEIN, L.: Phenological aids to species identification on air photographs. Canad. Dept. of Forestry, Technical Note No. 104, 1961.

11. STELLINGWERF, D. A.: Interpretation of tree species and mixtures on aerial photographs. in: Actes du IIe Symposium International de Photo-Interpretation, Paris 1966.
12. ZSILINSZKY, V. G.: Photographic Interpretation of tree species in Ontario. Ottawa, 1966
13. BAHRENBERG, G. und GIESE, E.: Statistische Methoden und ihre Anwendung in der Geographie. Teubner, Stuttgart 1975.
14. COOLEY, W. W. und LOHNES, P. R.: Multivariante Data Analysis. Wiley, New York 1971.
15. HOPE, K.: Methods of Multivariate Analysis. Univ. of London Press, London 1968.
16. TATSUOKA, M. M.: Multivariate Analysis. Wiley, New York 1971.
17. GOURLEY, J. et al.: Automatic technique for abstracting color descriptions from aerial photography. in: Photogr. Science & Engineering, Vol. 12, No. 1, 1968.
18. Munsell Book of Color, Serial No. 63A. Munsell Color Company, Baltimore (Md.), 1960.
19. RIB, H. T.: Color measurements. in: Manual of color aerial photography, pp. 12-24. Amer. Soc. of Photogrammetry, Falls Church (Va.), 1968.

Literaturbesprechung

THE GEOGRAPHY AND MAP DIVISION, LIBRARY OF CONGRESS: A Guide to its Collection and Services. Washington 1975. 42S. Erhältlich durch: Superintendent of Documents, Government Printing Office, Washington D. C. 20402, für $ 1.44 mit einem 25-prozentigen Zuschlag für Porto ins Ausland.
THE GEOGRAPHY AND MAP DIVISION, LIBRARY OF CONGRESS: A List of Geographical Atlases in the Library of Congress, Vol. 7. Washington 1973. 703 S. Erhältlich wie oben für $ 11.75 plus Porto.

Bei ihrer Gründung im Jahre 1800 besass die Kongressbibliothek der USA 3 Karten und 4 Atlanten. Heute sind es über 3,5 Millionen Karten, 38'000 Atlanten, 250 Globen und 500 Reliefmodelle, womit sie über die grösste Kartensammlung der Welt verfügt. Dies gilt nicht nur anzahlmässig, sondern auch in regionaler, historischer und thematischer Hinsicht. Den Schwerpunkt der Sammlung bilden jedoch erwartungsgemäss Dokumente des nordamerikanischen Kontinentes. Der 42-seitige Führer zu dieser umfangreichsten Kartenkollektion beginnt mit einem Vorwort des heutigen Abteilungschefs, Walter W. Ristow und einem kurzen Ueberblick über die abwechslungsreiche Vergangenheit der Institution. Zur Hauptsache enthält er jedoch Beschreibungen der Spezialsammlungen, die den kostbarsten Teil der gesamten Sammlung darstellen. Das letzte Kapitel vermittelt schliesslich einen Einblick in die Organisation und die Dienstleistungen der "Geography and Map Division" im Gesamtrahmen der Kongressbibliothek.
Eine solch umfangreiche Bibliothek bzw. Sammlung präsentiert sich dem Benützer naturgemäss als unüberblickbar und deshalb schwer zugänglich. Die von Philip Lee Phillips, dem ersten Chef der Kartenabteilung, im Jahre 1909 begonnene Zusammenstellung der "List of Geographical Atlases" leistet hier eine wertvolle Hilfe. Der vorliegende 7. Band von Mrs. LeGear, einer langjährigen Mitarbeiterin der Kartenabteilung, enthält über 7000 Atlanten ausschliesslich der westlichen Hemisphäre, die zwischen 1920 und 1969 in den Besitz der Kongressbibliothek kamen. Die Unterteilung erfolgt primär nach Regionen, wobei die Spezialatlanten entsprechend der Themenstellung noch weiter gruppiert worden sind. Jeder verzeichnete Atlas ist nummeriert und beschrieben; in vielen Fällen ist zusätzlich noch das Inhaltsverzeichnis abgedruckt. Ein 130-seitiges Autorenverzeichnis vervollständigt das bibliographische Werk.

H. Kishimoto