AUSTRIAN JOURNAL OF STATISTICS
Volume 38 (2009), Number 4, 287–291

Book Reviews

Christian F. G. SCHENDERA (2010). **Clusteranalyse mit SPSS. Mit Faktorenanalyse**. Oldenbourg Verlag, München, 435 Seiten, ISBN 978-3-486-58691-6. (€ 49.80)

In diesem Buch werden in drei Kapiteln vier unterschiedliche Methoden der multivariaten Statistik und deren Anwendung in SPSS beschrieben: Clusteranalyse, Faktorenanalyse, Hauptkomponentenanalyse und Diskriminanzanalyse. Die Methoden werden in Textform beschrieben, die zugehörigen mathematischen Formeln sind im Anhang angeführt. Auf eine algorithmische oder formelhafte Beschreibung der Methoden wird zumeist verzichtet.

Im Folgenden werden diese drei Kapitel einzeln besprochen, bevor eine generelle Zusammenfassung versucht jenen Personenkreis zu beschreiben, für den dieses Buch eine Bereicherung sein könnte.

Clusteranalyse: Das Erfreuliche an diesem Kapitel ist, dass sehr viel Platz verwendet wird, den Anwender darauf hinzuweisen, dass es nicht genügt, in SPSS ein paar schnelle Klicks zu tätigen und dann den Output zu interpretieren. Es wird anschaulich darauf hingewiesen, dass die Standardisierung von Variablen, das Skalenniveau der Variablen, die Auswahl der Variablen und die Methodenauswahl großen Einfluss auf das Result haben. Speziell die Anwendbarkeit von Methoden bei verschiedenen Skalenniveaus der Daten ist übersichtlich und höchst informativ dargestellt.

Leider wird in diesem Kapitel einer der wichtigsten Schritte ausgespart: die Transformation der Daten (bei stetiger Skalierung der Variablen). Bei „extrem“schief verteilten Variablen werden Gruppen in den Daten eher (nur) nach einer Transformation der Daten gefunden. D.h. auch wenn für manche der Verfahren keine Verteilungsannahme per se formuliert ist, so werden elliptische, approximativ bi- oder multimodale Clusterstrukturen in der Regel besser erkannt (siehe z.B. Templ, Filzmoser, und Reimann, 2008). In einem Beispiel zur Clusterung wurden außerdem Kompositionsdaten (Ausgaben in Urlaub, Sport, Kfz, etc.) verwendet ohne eine geeignete Transformation (Aitchison, 1986) zu verwenden.

Ganz entscheidend in der Clusterananlyse ist auch die Auswahl von Variablen. Das wird im Buch erwähnt, leider werden aber hier keine Strategien aufgezeigt dies praktisch – fern von fachstatistischen Gründen – durchzuführen. Weiters wird die Bedeutung der Distanzmatrix als Input vieler Clusterverfahren erwähnt und die Auswirkung bei Verwendung von unterschiedlichen Metriken angedeutet. Hier offenbart sich eine Schwachstelle in SPSS: die Prozedur CLUSTER bietet laut Buch keine Möglichkeit die Metrik zur Berechnung der Distanzmatrix anzugeben, sondern es können nur abgespeicherte Resultate, erzeugt mit der Prozedur PROXIMITY, geladen werden.

Das Buch weist darauf hin, dass es wichtig ist *das* bestmögliche Resultat zu erhalten und es werden Strategien aufgezeigt, wie dieses „bestmögliche“Resultat zu erreichen ist. Der Autor rät im Allgemeinen von einer explorativen Suche ab. Dies stellt jedoch eine unnötige Einschränkung dar und gerade die Clusteranalyse als eine Methode des unüberwachten Lernens sollte in explorativer Weise eingesetzt werden. Jedes einzelne (verschiedene) Resultat kann einen Erkenntnisgewinn bringen (Templ et al., 2008).

Viel Platz wird Methoden für die hierarchische Clusterung, partitionierenden Clusterverfahren (im Wesentlichen k-means) und einem Zwei-Schritt-Verfahren (für die Clusterung von Variablen mit unterschiedlichen Skalenniveaus) eingeräumt und weitere Methoden werden kurz erwähnt. Moderne, oft bessere Methoden, wie z.B. modellbasierte Clusterung (Fraley und Raftery, 1998, 2002), werden nicht angeführt – wohl aus einem guten Grund: sie sind nicht in SPSS implementiert. Ebenso wird die Existenz von Fuzzy-Clustering (siehe z.B. Höppner, Klawonn, und Kruse, 1999) verschwiegen.

Ausgespart bleibt auch fast gänzlich die Beurteilung der optimalen Clusteranzahl, wichtiger Eingangsparameter für viele Verfahren. Die Beurteilung über einen Scatterplot (Anzahl von Clustern gegen Gütemass) wird gezeigt, jedoch die (vielen möglichen) Gütemasse zur Beurteilung des Clusterergebnisses und zur Feststellung der optimalen Clusteranzahl werden weder beschrieben noch zitiert.

Andere Unterkapitel des Buches, wie z.B. das Auffinden von Ausreißern mit der Prozedur ANOMALY werden unzureichend beschrieben, andere wieder entsprechen nicht dem Stand der Technik, z.B. das Unterkapitel über das visuelle Auffinden von Clustern (siehe z.B. (Cook und Swayne, 2007)). Nicht nur hier werden einschlägige Bücher und Publikationen nicht zitiert (wie z.B. Kaufman und Rousseeuw, 1990; Everitt, Landau, und Leese, 2001).

Faktorenanalyse: Erfreulich ist, dass in diesem Buch nicht über *die* Faktorenanalyse gesprochen wird, sondern hervorgestrichen wird, dass viele unterschiedliche Methoden für die Faktorenanalyse ausgewählt werden können. Leider wird hier der alte Fehler begangen und es wird wieder einmal die (modellfreie) Hauptkomponentenanalyse als einer der Verfahren der (modellhaften) Faktorenanalyse bezeichnet, was sie nicht ist und wo auch, wie vom Autor angeführt, die Anwendungsgebiete – trotz vieler Ähnlichkeiten – oft nicht identisch sind.

Der Autor erhebt in diesem Buch Anspruch, nicht nur die Anwendung in SPSS, sondern die Faktorenanalyse selbst zu erklären. Aus didaktischen Gründen wäre eine Visualisierung der Grundideen anhand von zweidimensionalen Beispielen (wie liegen Faktoren; zeigen der unterschiedlichen Rotation) vonnöten.

In diesem Kapitel wird nicht auf den wichtigsten Inputparameter für die Faktoren- und Hauptkomponentenanalyse eingegangen: der Kovarianz/Korrelationsmatrix und deren Schätzung. Speziell Methoden der Faktorenanalyse können von Ausreißern sehr stark beeinflusst werden (siehe z.B. (Pison, Rousseeuw, Filzmoser, und Croux, 2003)). Deshalb sei dem Anwender eine robuste Schätzung ans Herz gelegt von deren Existenz und Vorteilen im Buch nichts erwähnt wird. Auch die wichtigste Methode zur Interpretation der Ergebnisse wurde nicht erwähnt – es fehlt der Biplot (Gabriel, 1971), welcher den Zusammenhang zwischen Beobachtungen, Variablen und Faktoren/Komponenten zeigt.

Diskriminanzanalyse: Im vorigen Kapitel hat der Autor gut herausgestrichen, dass es unterschiedliche Methoden der Faktorenanalyse gibt. In diesem Kapitel wird dies ausgespart und es wird nur die (bayesianische) lineare Diskriminanzregel (kurz) vorgestellt. Vor allem die Diskriminanzregel von Fisher und die quadratische Diskriminanzanalyse wird der kundige Nutzer vermissen.

Zusammenfassung: *„Die Auswahl (Anm.: von Methoden) sollte in Absprache mit einem erfahrenen Methodiker/Statistiker erfolgen“*. Dieser Ausspruch im Vorwort des Buches gilt beim Leser – sofern er nicht Methodiker/Statistiker ist – möglicherweise auch nach der Lektüre des Buches. Das Zielpublikum ist der Nicht-Statistiker, welcher die Daten mit hochkomplexen Methoden der multivariaten Statistik in SPSS analysieren sollte. Diesem werden oft in leicht verständlichen Sätzen die Grundzüge von Methoden erklärt und er wird – speziell im Kapitel der Clusteranalyse – behutsam an die Aufgabenstellung herangeführt. Dieses Zielpublikum könnte aber auch mit einem Handbuch der SPSS Prozeduren und Fachbüchern (wie z.B. mit dem oft zitierten Bacher, 2001) zufriedenen gestellt sein.

Dem Buch merkt man leider an, dass kein professionelles Textprogramm verwendet wurde. Einigen Tabellen und Grafiken täte es gut, wenn sie mit einer referenzierbaren Nummerierung und Über- oder Unterschrift ausgestattet wären, und nicht alle Grafiken sind qualitativ hochwertig. Zum Beispiel werden einzelne Gruppen von Beobachtungen in Scatterplots mit kaum unterscheidbaren Farben und gleichen Symbolen dargestellt. Einige Grafiken erinnern an Nadeldrucker, dies liegt aber an dem teilweise archaisch wirkenden grafischen Output von SPSS.

Der wissenschaftliche Input und die Beschreibung von modernen Methoden war und kann das Ziel eines solchen Buches nicht sein. Jedoch werden einige sehr wichtige Standardmethoden in der Clusteranalyse und in der Diskriminazanalyse weder erwähnt noch zitiert. Auch werden wichtige Referenzen zu Standardwerken und Publikationen in diesem Bereichen vermisst, da fast ausschließlich Bücher im Bereich der sozialwissenschaftlichen Methodenlehre im Buch zitiert werden.

Literatur

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman & Hall.

Bacher, J. (2001). *Clusteranalyse. Anwendungsorientierte Einführung* (2. ed.). München: Oldenbourg Verlag.

Cook, D., und Swayne, D. F. (2007). *Interactive and Dynamic Graphics for Data Analysis With R and GGobi*. New York: Springer Verlag.

Everitt, B., Landau, S., und Leese, M. (2001). *Cluster Analysis* (3. ed.). Oxford University Press.

Fraley, C., und Raftery, A. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, *41*, 578-588.

Fraley, C., und Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*, 611-631.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*, 453-467.

Höppner, F., Klawonn, F., und Kruse, R. (1999). *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*. New York: John Wiley and Sons.

Kaufman, L., und Rousseeuw, P. J. (1990). *Finding Groups in Data*. New York: John Wiley and Sons.

Pison, G., Rousseeuw, P. J., Filzmoser, P., und Croux, C. (2003). Robust factor analysis. *Journal of Multivariate Analysis*, *84*, 145-172.

Templ, M., Filzmoser, P., und Reimann, C. (2008). Cluster analysis applied to regional geochemical data: Problems and possibilities. *Applied Geochemistry*, *23*, 2198-2213.

Matthias Templ
Institut für Statistik und Wahrscheinlichkeitstheorie
Technische Universität Wien
sowie
Register, Klassifikation und Methodik
Statistik Austria

Karlheinz ZWERENZ (2009). **Statistik: Einführung in die computergestützte Datenanalyse** (4. überarbeitete Auflage). Oldenbourg Verlag, München, 434 Seiten, ISBN 978-3-486-59112-5. (€ 36.80)

Der Autor möchte mit seinem vorliegenden Buch Studierenden und Anwendern der Statistik eine Einführung in die grundlegenden statistischen Methoden geben. Dies gelingt zum Teil gut, da viele (kleine) Beispiele die Sachverhalte klar darstellen, dennoch fehlen oft Details, die dem generellen Verständnis noch weiterhelfen würden. Exemplarisch kann man dazu die Definition des Medians betrachten, die nur für Stichprobenumfänge kleiner 100 angegeben wird. Ein Hauptaugenmerk des Buches – wie der Untertitel nahelegt – liegt darin, neben den Verfahren auch deren Anwendung mithilfe der Programme Microsoft Excel und SPSS vorzustellen. Dazu werden an passenden Stellen Hinweise zu den benötigten Funktionen der Software angegeben. Die Wahl der bereits 2001 erschienenen Version Microsoft Excel 2002 sollte allerdings überdacht werden.

Inhaltlich ist das Buch in fünf Teile gegliedert, wobei sich der erste Teil mit den Grundlagen der Statistik beschäftigt. Hier wird neben einer kurzen Motivation zur Verwendung von Software auch auf einen üblichen Projektablauf und auf Grundbegriffe eingegangen. Teil 2 befasst sich mit eindimensionalen deskriptiven Statistiken. Dabei werden einfache grafische Darstellungen und die gängigsten numerischen Kenngrößen zu Lage, Streuung, Schiefe u.Ä. eingeführt. Ein ausführlicher Abschnitt widmet sich zusätzlich noch Indexzahlen. In Teil 3 folgen die Erweiterung für zweidimensionale deskriptive Statistiken – vor allem im Sinne einer Zusammenhangsanalyse – und grundlegende Eigenschaften von Zeitreihen. Teil 4 widmet sich schließlich den Grundlagen der Wahrscheinlichkeitstheorie. Dabei werden angefangen vom Zufallsexperiment und Laplace Wahrscheinlichkeiten über Zufallsvariablen und deren Verteilungen bis zu Grenzwertsätzen innerhalb von etwa 80 Seiten die wichtigsten Aspekte knapp vorgestellt. Im fünften und letzten Teil des Buches werden Grundlagen der induktiven Statistik behandelt. Dazu stellt der Autor Punkt- und Intervallschätzungen und Hypothesentests vor und liefert auch konkrete Testverfahren.

Das gesamte Buch ist als Lehrbuch konzipiert und keinesfalls als reine Methodensammlung. Dies wird auch dadurch bestätigt, dass sich auf der Homepage des Autors ein Link zu einem E-Learning Programm befindet, das im Wesentlichen analog zum vorliegenden Buch aufgebaut ist. Positiv erwähnt sei an dieser Stelle, dass in allen Abschnitten auf ein spezielles Datenbeispiel – das sogenannte Masterprojekt – eingegangen wird. Dem Leser sollte es damit aufgrund der Vertrautheit mit den Daten leichter fallen, die neuen Methoden zu verstehen. Andererseits finden sich teilweise seitenlange Tabellen zur Erklärung von Größen, die dem Verständnis wohl nicht förderlich sind. Zusammenfassend gibt das vorliegende Buch eine umfassende und einfache Einführung in elementare statistische Verfahren und deren Anwendung in Microsoft Excel und SPSS, wobei der interessierte Leser wohl in Detailfragen auf zusätzliche Literatur zurückgreifen wird.

Johannes Schauer
Institut für Statistik
Technische Universität Graz