Klassifikation von Ausreißern in Kompositionsdaten

K. Gerald van den Boogaart, Institut für Stochastik

Robuste Statistik / Ausreißer / Kompositionsdaten

Kompositionsdaten sind Daten über die Anteile verschiedener Komponenten in einer Probe. Oft erkennt man solche Daten daran, dass sich die Mengenangaben in einer Zeile zu 100% addieren. In der Statistik ist allgemein bekannt, dass solche Daten nicht mit den klassischen multivariaten statistischen Verfahren ausgewertet werden dürfen, da dann Artefakte des Messprozesses das Ergebnis maßgeblich beeinflussen können (z.B. ob eine negative oder positive Korrelation vorliegt). Für solche Kompositionsdaten gibt es eine spezielle statistische Methodik, die auf die Arbeiten von John Aitchison zurückgeht und die in den letzten Jahren durch die Arbeiten von Vera Pawlosky-Glahn mit dem Prinzip des „Arbeitens in Koordinaten“ endlich auch für Anwender zugänglich geworden ist. Dieses Prinzip sagt, dass man mit einer isometrischen log-ratio Transformation (ilr) transformierte Kompositionsdaten theoretisch genauso behandeln kann wie einen üblichen multivariaten Datensatz. In den letzten Jahren sind daher viele neue und leicht anwendbare statistische Methoden für Kompositionsdaten entwickelt und in Software verfügbar gemacht worden. Wir waren dabei z.B. maßgeblich an der Entwicklung eines entsprechenden Paketes für das Statistikpaket R beteiligt.

Die aktuelle Forschungsarbeit zur Kompositionsdatenanalyse in Zusammenarbeit mit Matevz Bren, Maribu, Slovenien, und Raimon Tolosana Delgada, Barcelona, Spanien im Jahre 2009 bezog sich darauf, dass sich Ausreißer und Messfehler in Kompositionsdaten nun völlig anders auswirken als in klassischen multivariaten Datensätzen, und das Prinzip des Arbeitens in den Koordinaten hier nicht unmittelbar anwendbar ist. So ändert z.B. ein einziger Messfehler alle Prozentanteile, so dass eventuell Komponenten als Ausreißer erkannt werden, die keine sind. Allerdings können Ausreißer mittels der auf die ilr-transformierten Daten angewendeten klassischen multivariaten Ausreißererkennungsverfahren weiter erkannt werden.

Konkret wurde eine Systematik verschiedener möglicher Ursachen von Ausreißern entwickelt: Einzelne Messfehler, mehrere Messfehler, zufällige extreme Beobachtung, multimodale Verteilung, schwere Verteilungsschwänze, Verunreinigung.

Zu dieser Systematik wurden explorative statistische Verfahren entwickelt, die es erlauben, einzelnen Ausreißern und Ausreißergruppen mögliche Ursachen zuzuordnen, gleichartige Ausreißer zu gruppieren und Aussagen über die Fehlerhäufigkeit im Datensatz zu treffen, wenn Ausreißer und zufällige extreme Beobachtungen in der gleichen Größenordnung liegen. Dadurch können nun Ausreißer in Kompositionsdatensätzen analysiert, klassifiziert und verschiedenen Ursachen zugeordnet werden.

K.Gerald van den Boogaart, Raimon Tolosana-Delgado, Matevz Bren (2009) Classification of outliers in compositional data, submitted to Computers and Geosciences, under revision