Z. geol. Wiss., Berlin 40 (2012) 4/5: 235 – 244, 2 Abb., 1 Tab.

Compositional data analysis: one hundred years of debate

Analyse von Kompositionsdaten: 100 Jahre Forschungsgeschichte

El análisis de datos composicionales: cien años de debate

VERA PAWLOWSKY-GLAHN, Girona; JUAN JOSÉ EGOZCUE, Barcelona

Key words: closure problem, compositional data, simplex, Aitchison geometry

Abstract

Back in 1897, Karl PEARSON published a paper, which title began with the words "On a form of spurious correlation...". He was the first to point out the dangers which may befall the analyst when using conventional statistical methods with compositional data, known at that time as closed data. Many have been the scientists that have tried to understand, explain, and solve the closure problem since then, especially among geologists. But it was not until the 1980's that a solution was proposed, which has developed in a completely new methodology. This approach was the log-ratio approach, put forward by John AITCHISON. In this contribution, we summarize the problems and the state-of-the-art of the new developments.

Zusammenfassung

Karl PEARSON hat als erster im Jahr 1897 auf das Problem der Scheinkorrelationen aufmerksam gemacht, die bei der Anwendung herkömmlicher statistischer Methoden auf Kompositionsdaten – auch unter dem Begriff „geschlossene Systeme" bekannt – auftreten. Im vorliegenden Beitrag wird die Forschungsgeschichte der Analyse von Kompositionsdaten behandelt, wobei vier Phasen genauer betrachtet werden. In der ersten Phase von 1897 bis 1960 wurden vor allem die Fallgruben bei der Anwendung der klassischen multivariaten Statistik untersucht. In der zweiten Phase von 1960 bis 1980 führte der Petrologe Felix Chase das Konzept des negativen Bias ein – als „closure problem", weil die Restriktion auf konstante Summe (100% oder 1) der Variablenwerte negative Korrelationen erzwingt. Der Nachweis der Fehlerhaftigkeit der Anwendung multivariater Standardmethoden auf Kompositionsdaten war Ziel dieser Untersuchungen, die aber keine Lösung brachten. Erst 1980 entwickelte John AITCHISON den Log-Ratio-Ansatz; das ist die dritte Phase der Entwicklung. Ausgangspunkt seiner Überlegungen war die Tatsache, dass Kompositionsdaten nur Information über relative Werte der Variablen liefern, nicht absolute Werte! Da Log-Ratios mathematisch einfacher zu handhaben sind als Ratios (Brüche) und die Log-Ratio-Transformation eine bijektive (eineindeutige) Abbildung auf den reellen euklidischen Raum darstellt, konnten damit die Standardmethoden der multivariaten Statistik (ohne die Restriktion auf konstante Summen) auf die transformierten Daten angewendet werden. Die Ergebnisse dieser statistischen Auswertung können in den Simplex rücktransformiert werden und liefern damit Aussagen für die Kompositionsdaten. Dieser Log-Ratio Ansatz führte zu einer völlig neuen Methodik und Geometrie.

Die vierte Phase ist gekennzeichnet von der Erkenntnis, dass die internen Simplexoperationen Perturbation und Potenzierung, sowie die Einführung einer Distanz im Simplex (durch J. AITCHISON) einen euklidischen Raum definieren. Das führt dazu, im Simplex zu bleiben, indem man Kompositionen in Koordinaten darstellt und ihre Interpretation aus den Beziehungen ihrer Lage im Simplex ableitet. Der Probenraum von (Zufalls-) Kompositionen ist somit der Simplex mit Simplex-Metrik und Simplex-Maß, die verschieden von der euklidischen Metrik und dem Lebesgue-Maß im Reellen sind. Der Effekt dieser neuen Geometrie wird in den Abb. 1 und 2 besonders deutlich, wo Beispieldaten in einem Dreiecksdiagramm und als Koordinaten (Tab. 1) dargestellt sind. Ein Beispiel zeigt, wie eine Orthonormalbasis und die dazugehörigen Koordinaten durch den Einsatz einer sequentiellen binären Partition (SBP) erzeugt werden, die einfach zu definieren und zu interpretieren ist.

Resumen

En 1897, Karl Pearson publicó un artículo cuyo título empezaba con la palabras “Una forma de correlación espúrea...”. Fue el primero en señalar los riesgos en los que puede incurrir un analista al usar métodos estadísticos convencionales con datos composicionales, conocidos entonces como datas clausurados. Muchos científicos han intentado desde entonces entender, explicar, y resolver el problema de la clausura, en particular, muchos geólogos. Pero no fue hasta los 1980’s que se propuso una solución, solución que se ha desarrollado en una metodología completamente nueva. Este enfoque, basado en log-cocientes, fue propuesto por John Aitchison. En la presente contribución resumimos los problemas y el estado del arte de los nuevos desarrollos.

$$D$$

$$i=1$$