

CURSO DE ANÁLISIS ESTADÍSTICO PARA DATOS COMPOSICIONALES
Organizado por el Departamento de Matemáticas de la Universidad de los Andes
A impartirse en el Instituto Colombiano del Petróleo
Piedecuesta, Santander, marzo 20 a marzo 23 del 2007

Expositores:

Dra. Vera Pawlowsky-Glhan, Universidad de Girona

Dr. Juan José Egozcue, Universidad Politécnica de Cataluña

Coordinador

Dr. René Meziat, Universidad de los Andes

Motivación

Datos composicionales son aquellos que sólo contienen información relativa. Suelen presentarse en forma cerrada, como tantos por uno o por cien, partes por millón, concentraciones químicas, y aparecen frecuentemente en las aplicaciones de geología, biología, medicina, química, ingeniería, econometría y ciencias sociales. Su análisis estadístico presenta dificultades debido a la correlación espuria (Pearson 1897). Desde que J. Aitchison (1982) introdujo la representación de los datos en el simplex y las técnicas basadas en log-cocientes, se han desarrollado técnicas estadísticas apropiadas, basadas en la idea que el simplex tiene una geometría propia. Los últimos avances en el análisis corresponden a la representación de estos datos en coordenadas ortogonales en dicha geometría, lo que permite trabajar en un marco coherente y amplio con las técnicas estadísticas estándar.

Objetivo

Brindar a los participantes una introducción a los conceptos del análisis de datos composicionales.

Metodología

El curso se realiza en cuatro días consecutivos con 4 horas lectivas cada uno. Cada sesión de 4 horas consiste en dos presentaciones conceptuales seguidas de la exposición de casos prácticos. Se prevén recesos después de cada sesión.

Pre-requisitos de los asistentes

Se espera que los asistentes tengan una formación universitaria en Probabilidad, Estadística y Álgebra Lineal. Es conveniente estar familiarizado con el uso de hojas de cálculo y programas comerciales para problemas de estadística general. El curso se presenta en español pero la mayor parte del material está en inglés. Dirigido a: ingenieros civiles, ingenieros de petróleo, ingenieros de minas, estadísticos, geólogos, físicos y matemáticos interesados en herramientas de estadística para la toma de datos composicionales en geología que tengan los prerequisites mencionados.

Programa del Curso

Día 1 Presentación, Historia e Introducción

R. Meziat. Presentación del curso, su contenido, su metodología y los expositores. 30 minutos.

J.J. Egozcue. Escala de datos. Espacio muestral. Variables aleatorias, densidad de probabilidad, momentos de distribuciones de probabilidad. La distribución normal bi-variada. 60 minutos. Receso de 30 minutos.

R. Meziat. Espacios vectoriales y espacios Euclídeos. Dimensión, bases y bases ortogonales. Coordenadas, cambios de coordenadas y matrices de cambios de coordenadas. Producto escalar, norma y distancia en espacios vectoriales. Subespacios lineales, proyecciones y complemento ortogonal. 60 minutos. Receso de 10 minutos.

V. Pawlowsky. Introducción a los datos composicionales. Historia. Estructura de Espacio Euclídeo. 50 minutos.

Día 2 Geometría de los Datos Composicionales (CoDa)

V. Pawlowsky. Explicación sobre la instalación del programa Codapack. 30 minutos.

V. Pawlowsky. Prácticas sobre operaciones básicas con CoDa en el simplex y técnicas de representación e interpretación de CoDa en el ternario. 60 minutos. Receso de 30 minutos.

J.J. Egozcue. Bases ortogonales en el simplex. Representación en coordenadas. Distribución Normal. 50 minutos. Receso de 10 minutos.

J.J. Egozcue. Prácticas de los conceptos. SBP, base y representación simplex-coordenadas. 60 minutos.

Día 3 Análisis Exploratorio de CoDa

R. Meziat. Descomposición en Valores Singulares (SVD). 50 minutos. Receso de 10 minutos.

V. Pawlowsky. Centro y variabilidad. Matrices de variación. El biplot para CoDa. Dendrograma de balances. Uso conjunto de biplot y dendrograma de balances. 60 minutos. Receso de 30 minutos.

V. Pawlowsky. Prácticas de aplicación de los conceptos: biplot, dendrogramas, uso de las coordenadas y matrices de variación. 90 minutos.

Día 4 Regresión en el Símplex y Aplicaciones

R. Meziat. Regresión lineal por mínimos cuadrados. Pruebas de hipótesis en estadística. 50 minutos. Receso de 10 minutos.

J.J. Egozcue. Procesos en el símplex. Regresión lineal en el símplex. 60 minutos. Receso de 30 minutos.

J.J. Egozcue. Prácticas sobre representación de procesos en el símplex y en sus correspondientes coordenadas euclídeas. Modelos estadísticos en el símplex: regresión. Un estudio de regresión en el símplex. 90 minutos.

horas	MARTES 20	MIÉRCOLES 21	JUEVES 22	VIERNES 23
2:00	R. Meziat	V. Pawlowsky	R. Meziat	R. Meziat
	J.J. Egozcue	V. Pawlowsky	R. Meziat	R. Meziat
3:00	J.J. Egozcue	V. Pawlowsky	V. Pawlowsky	J.J. Egozcue
	receso	receso	V. Pawlowsky	J.J. Egozcue
4:00	R. Meziat	J.J. Egozcue	receso	receso
	R. Meziat	J.J. Egozcue	V. Pawlowsky	J.J. Egozcue
5:00	V. Pawlowsky	J.J. Egozcue	V. Pawlowsky	J.J. Egozcue
	V. Pawlowsky	J.J. Egozcue	V. Pawlowsky	J.J. Egozcue

Material que se aporta con el curso

Las transparencias utilizadas por los expositores en forma impresa. Archivo en *pdf* de una versión reciente del texto titulado: *Lecture notes on compositional data analysis*, de V. Pawlowsky-Glahn, J.J. Egozcue y R. Tolosana-Delgado. Algunos artículos de investigación recientes de los expositores. Los archivos correspondientes a los datos básicos para desarrollar algunos casos prácticos. El programa CoDaPack estará disponible.

Referencias básicas

Aitchison, J. (1986): *The Statistical Analysis of Compositional Data*, Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press).

Mathematical Geology (2005), Vol. 37 (7): Special issue on compositional data analysis.

Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (Eds.) (2006): *Compositional data analysis in the geosciences, from theory to practice*, Special publication 264, Geological Society, London.

Datos del Profesorado

Dr. Vera Pawlowsky-Glahn (vera.pawlowsky@udg.es)

Dr. Juan José Egozcue (juan.jose.egozcue@upc.edu)

Dr. René J. Meziat (rmeziat@uniandes.edu.co)

Dr. Vera Pawlowsky-Glahn is a professor of the Department of Computer Science and Applied Mathematics at the University of Girona. She studied Mathematics at the University of Barcelona in Spain and obtained her PhD (*doctor rerum naturam*) from

the Free University of Berlin in Germany. Before going to Girona, she was professor at the School of Civil Engineering at the Technical University of Catalonia (UPC) in Barcelona. Her main research topic since 1982 has been the statistical analysis of compositional data. The results obtained over the years have been published in multiple articles, proceedings and a book in the Oxford University Press series *Studies in Mathematical Geology*. She has been guest editor for a special issue on this topic for *Mathematical Geology* in 2005 and has acted, together with A. Buccianti and G. Mateu-Figueras, as editor of a book on compositional data analysis published by the Geological Society, London, as special publication 264. She is the leader of a research group on this topic involving professors from different Spanish universities located in Girona, Barcelona, Murcia and Cáceres. The group organises every two years a workshop on compositional data analysis, known as CoDaWork, and their research has received regularly financial support from the Spanish Ministry for Education and Science and from the University Department of the Catalan Government. Prof. Dr. Pawlowsky-Glahn has been vice-chancellor at UPC from 1990 to 1994, head of the Department of Computer Science and Applied Mathematics at the University of Girona in 2004-05, and dean of the Graduate School of the University of Girona in 2005-06. She received in 2006 the William Christian Krumbein Medal of IAMG and has been nominated Distinguished Lecturer for 2007 by the same association.

Dr. Juan José Egozcue is full professor at the Dept. of Applied Mathematics III at the Universidad Politécnica de Cataluña (UPC) (Spain). He studied Physics, oriented to Geophysics and Meteorology, at the University of Barcelona (Spain) and obtained his PhD in the same university with a dissertation on maximum entropy spectral analysis (1982). In 1978 he got a position as a lecturer in the school of civil engineering in Barcelona (Escuela de Ingeniería de Caminos, Canales y Puertos de la UPC), teaching several topics on applied mathematics. In 1983, he started teaching Probability and Statistics. He became Associate Professor in 1985, and Full Professor in 1989, in the UPC, where he has been Chair of the Department of Applied Mathematics III and vice-rector of the UPC.

His research activities are presently centered in two lines: Estimation of natural hazards using Bayesian methods, specially applied to seismic, rainfall and ocean wave hazards; and Statistical analysis of compositional data, with special emphasis in the geometry of the sample space, the simplex.

He has been leader of several research projects. Nowadays, his main project is on vulnerability of dikes under action of severe ocean waves (granted by the Port Authority of Spain) and participates actively in a project of compositional data analysis supported by the Spanish ministry for education and science.

Dr. René Meziat is associated professor at the Mathematics Department of La Universidad de los Andes, he is currently in charge of the Research Group on Numerical Analysis and Optimization. He has a bachelor in Mathematics from La Universidad de los Andes and a PhD in Applied Mathematics from La Universidad Politécnica de Cataluña. His main research activity focuses on numerical analysis and operations research. He is author of several research papers and chair of several research grants supported by outstanding institutions.