

An introduction to two data mining techniques: Kohonen maps and text mining

Phong Nguyen, General Statistics Office (Vietnam)

Dominique Haughton, Bentley College (USA) and Toulouse School of Economics (France)

An introduction to Kohonen maps (1.25 hours)

In this presentation we will introduce participants to the Kohonen map - otherwise known as self-organizing map – methodology in the context of a comparison of the living standards of Vietnamese provinces. It is now well recognized that the living standard of a province is a multi-dimensional concept . Limitations of currently used single indices such as GDP per capita or the poverty rate, as well as composite indicators such as the Human Development Index (HDI) will be discussed. We will explain how the Kohonen map methodology makes it possible to map Vietnamese provinces on a two-dimensional grid, on the basis of a number of living standards indicators, and how to interpret the dimensions on the grid.

Specifically, participants will learn:

- How to prepare the data and construct a Kohonen map with the Kohonen Matlab toolkit
- How to interpret the resulting output, including the U-matrix and Component matrix
- How to use bootstrap methods to obtain measures (and their standard deviations) of accuracy and stability of a Kohonen map

An introduction to text mining with SAS TextMiner (1.25 hours)

This presentation will first introduce participants to text mining with SAS Text Miner in the context of an analysis of free-text responses to a questionnaire administered to 416 undergraduate students in a business university about their perceptions of unfair grading practices. Secondly we will present a case study where text mining was used to help predict missing text data in the context of a major information systems corporation in the Boston area.

Specifically participants will learn:

- How to use a text mining tool to identify terms and their frequency of occurrence in text data
- How to cluster open text responses with SAS Text Miner, using both the Expectation-Maximization (EM) algorithm and hierarchical clustering
- How to use this clustering to identify preliminary themes within the responses

Summary of the case study objectives

The text analysis case study was performed in the context of the database marketing group of a leading provider of information infrastructure. At several contact points, customers have provided their job title, but have not always filled in the field “Key Player Description”, which consists of a brief (a few words) job description. The problem here is to impute the missing brief job description for a file of about a million such customers on the basis of available job titles. We will demonstrate how this problem was addressed with SAS Text Miner, using a sample of 10,000 customers.