recovery
energy
environment
economy
2011 CSPG CSEG CWLS Convention

Multivariate Data Cleaning and Classification

Petro Babak*, Enzo Insalaco, Patrick Henriquel and Olena Babak
Total E&P Canada Ltd.
petro.babak@external.total.com; pbabak@ucalgary.ca
*Presenter; Correspondence address: Total E&P Canada Ltd.
2900, 240 – 4th Ave SW, Calgary, AB; Email: petro.babak@external.total.com; pbabak@ucalgary.ca; Phone: 403-538-8859

Abstract

Geological data sets often contain some errors. These errors might be linked to different issues like measurement devise limitations, measurement recording glitches or simply interpretation inconsistencies related to change in interpreter and/or interpretation concept. However, irrelevant of the reason for erroneous data, every geomodeler, geologist and/or engineer have to deal with data problem issues as otherwise he or she would waste days drawing wrong conclusions because the errors have not been first identified and excluded.

In the literature one can find several methods for data cleaning, the most common being statistical methods (standard deviation, range, or clustering algorithms), data transformation, and duplicate elimination. Not all of the approaches are equally useful and effective. In this talk we present and further develop a validated multivariate cleaning methodology based on the Mahalanobis distance approach. This method is widely used cluster analysis and classification techniques; application to geological data is novel. After the methodology is explained in detail it is illustrated using the core well data from the Joslyn Project. In particular, it is shown on the Jolsyn example how to identify and analyze errors in datasets containing many variables, including bitumen grade and particle size distributions (psd's) characterizing rock granulometry and specified by 22 bin sizes. Analysis is done on a by-facies basis. Furthermore, to avoid losing valuable information, for example, misclassified psd, geologically driven classification is developed. This classification incorporates multivariate geological information and allows assigning mislabeled data to the best suitable facies group or class. The proposed classification not only makes geological sense, but also makes further geological analysis more consistent and straightforward.

Acknowledgements

The authors would like to thank Total E&P Canada Ltd. and its partners, Suncor, Occidental Petroleum and Inpex for support of this work.