International Journal of Coal Geology 112 (2013) 2–13

Contents lists available at SciVerse ScienceDirect

International Journal of Coal Geology

journal homepage: www.elsevier.com/locate/ijcoalgeo

Review article

Geostatistics: A toolkit for data analysis, spatial prediction and risk management in the coal industry

R. Mohan Srivastava *

Benchmark Six Inc., #1100-120 Eglinton Avenue East, Toronto, Ontario, Canada M4P 1E2

ARTICLE INFO

Article history:
Received 28 September 2012
Received in revised form 15 January 2013
Accepted 15 January 2013
Available online 29 January 2013

Keywords:
Geostatistics
Spatial variation
Variogram
Spatial estimation
Kriging
Spatial simulation
Sequential Gaussian simulation
Risk analysis

ABSTRACT

An overview of the geostatistical toolkit is presented, from data analysis through estimation and simulation, with a focus on problems that typically arise in the assessment and development of coal deposits. Geostatistical procedures for the data analysis are described, leading to a discussion of the importance of spatial variation and the variogram. The most common geostatistical estimation procedure, ordinary kriging, is presented as an improvement to inverse-distance methods; two ways are presented of understanding kriging without recourse to the underlying mathematics. Estimation and simulation are compared and contrasted, and the benefits of a family of equally likely scenarios are covered. The paper concludes with brief summaries of the 16 additional papers in the International Journal of Coal Geology's Special Issue on Geostatistics, and provides two indexes to guide the reader to papers according to the problems they address and according to the tools they use.

Contents

1. Introduction

Coal is currently the largest source of fuel for the generation of electricity and, with hundreds of years of identified reserves, coal will likely remain the dominant energy fuel for decades. The development and production of these huge, untapped reserves, however, is going to encounter greater technical and economic difficulties: increased variability in coal quality, greater structural complexity, and more deleterious components. Geostatistics offers tools that are becoming increasingly valuable for building the earth models needed by the coal industry in order to assess and manage risk in the increasingly difficult projects that remain to be developed and mined.

This overview presents a quick tour of the geostatistical landscape to support the papers that appear in this International Journal of Coal Geology (IJCG) Special Issue on Geostatistics. It begins where a typical project

* Tel.: +1 416 322 2857; fax: +1 416 322 5075.
E-mail address: MoSrivastava@benchmarksix.com.

0166-5162/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.coal.2013.01.011

begins … data analysis … discussing how models of spatial variation are developed. This is followed by a discussion of resource/reserve estimation that includes a presentation of kriging, the work-horse of geostatistical estimation. Resource/reserve classification is then discussed, with a look at geostatistical methods for identifying the locations of “measured”, “indicated” and “inferred” resources in a deposit. Conditional simulation, the spatial version of Monte Carlo procedures, is introduced as an earth modeling method that is well suited for studies of blending, as well as studies that involve risk assessment.

The final section of this overview takes a look at the 16 papers that have been selected for this volume, organizing them into thematic groups so that the reader can develop a sense for where they'll find material with relevance to specific problems and issues.

2. Spatial variation

In the discussion that follows, and in the papers in this Special Issue, it will quickly become apparent that the cornerstone of geostatistics is spatial variation. The main focus of geostatistical data analysis is understanding and describing the spatial patterns in variables like coal quality and thickness. A key parameter that distinguishes geostatistical estimation from other types of estimation, such as inverse-distance interpolation, is a semi-variogram model (often just called a “variogram model” for short) that controls the weights given to nearby data. The feature of geostatistical simulation that sets it apart from other methods for building earth models is that it explicitly aims to correctly portray spatial variation. From start to finish through a geostatistical study, it is spatial variation that structures our thinking.

The reason that geostatisticians are curious about the spatial variation is that it plays an important role in many key aspects of the assessment of a coal deposit:

- Knowing something about the direction of maximum continuity is very useful when we are trying to make predictions of thickness and coal quality. Between the available drill holes, we are going to have to interpolate using the surrounding data, giving each one of the nearby samples a weight that reflects its importance to the estimate. If the geological evolution of the deposit has created a grain, a tendency for similar values to line up in a particular direction, then we can improve the reliability of our estimates by taking this into account when we interpolate.
- Knowing something about the noise in the data, or in particular sub-groups of data, allows us to make good choices about how much we should rely on any particular data point. If we have noisy unreliable data from old drill holes as well as more reliable data from new core holes, we would like to be able to lean more heavily on the reliable data, and less heavily on the not-so-reliable data. Our understanding of spatial variation will assist us in making good choices about how best to weight a combination of noisy and reliable data.
- Knowing something about spatial variation is helpful when trying to predict run-of-mine variability in a project that may require blending. If we know that significant changes in coal quality are possible over short distances, then we know that an open-pit operation with only one operating face may experience large daily fluctuations in the quality of coal shipped from the mine. If we have documented the possibility of large short-scale fluctuations, and we know that the project cannot tolerate large fluctuations, then we know that we need to consider some kind of blending … maybe the project needs several active operating faces to enable in-pit blending; maybe it needs engineered blending piles.
- Since our confidence in any particular estimate depends on how much the nearby data fluctuate, and how close they are, knowledge of the pattern of spatial variation is valuable for quantifying the uncertainty in our estimates. Risk analysis, and the ability to make decisions in the face of uncertainty, are both improved when we have an ability to establish reliable confidence intervals for all of our estimates.
- Knowing something about the spatial variation is helpful when classifying resources into the “measured”, “indicated” and “inferred” categories required by governmental regulations. If, for example, a particular estimate is based on many data that lie within the range that we can comfortably do geologic correlations, we have good reason to regard this as more reliable than another estimate that is based on data that all lie beyond the range over which we can comfortably correlate data.

3. Data analysis

3.1. Classical statistics

Many of the tools of geostatistical data analysis used in the papers in this volume will already be familiar to the reader because they are the graphical and numerical summaries used in classical statistics.

Fig. 1 shows an example of the graphical summaries commonly used for univariate data analysis along with the usual numerical summaries; these examples, as with many of the ones used in this overview, are from a coal deposit in the Powder River Basin in Wyoming. The histogram records the number (or percentage) of data values in each class, providing a visual sense for the range of values, their center and their spread. Statistics anchor this visual sense with specific information: the mean and median are measures of the center of the distribution; the variance and standard deviation are measures of the spread of the distribution.

The cumulative probability plot shows the chance (from 0 to 1) of a data value being lower than any given value on the x-axis. Often, this is presented with the y-axis, the probability axis, transformed in such a way that the data will plot as a straight line if they follow the classical bell-shaped distribution known as the Normal (or Gaussian) distribution.

A boxplot is a compact graphical summary that spares us the detail of the histogram, focusing our attention instead on a handful of key characteristics. The box in the middle of the diagram goes from the 25th percentile to the 75th percentile, i.e. it spans half the data; the arms that stick out of the box go to the minimum and maximum. The bar in the middle of the box records the location of the median, and the dot records the location of the mean. The compact nature of the boxplot makes it an excellent format for comparing distributions; with a group of side-by-side boxplots, we can readily judge if distributions from different populations are similar.

Fig. 2 shows an example of a scatterplot, the graphical summary most commonly used when we are analyzing more than one variable at a time. The statistic most commonly used to describe a bivariate relationship is the correlation coefficient, a number that lies between -1 and $+1$, and that measures how close the cloud of points comes to falling on a straight line.

The straight line most commonly used as a point of visual reference for a scatterplot is the conventional regression line. In addition to providing a mathematical equation that serves as a kind of summary of the cloud, the regression line is also a statistically optimal predictor.

When statisticians use the word “optimal” to describe a prediction, they almost always are referring to a prediction that minimizes the squared error. Whenever we try to predict something, we know that our prediction is likely wrong; we would be unimaginably lucky to hit the nail right on the head, so we accept that there is going to be an error, i.e. the difference between our prediction and the true value. “Least squares” estimates and “best linear unbiased estimates (BLUE)” are terms used in statistics for estimates produced by a procedure that explicitly aims to minimize the squared error. The classical regression line is the line that minimizes the sum of the squared differences between all the points in the cloud and the line (with the difference calculated vertically from the point to the line).