EAGE

P20

Integration of Disparate Data with Logratios of Conditional Probabilities

S. Hong* (University of Alberta), J.M. Ortiz (University of Chile) & C.V. Deutsch (University of Alberta)

SUMMARY

Combining conditional probabilities based on logratios is considered and compared with other integration models. Each data set is at first treated separately and merged with considering data source redundancies. Permanence of ratios and tau-model are presented as a way of merging conditional probabilities and their results are compared to those of logratio model. In logratios model, redundancy factors are iteratively optimized in order to improve the integrated estimate. Measure of goodness is used to quantitatively evaluate the models of integration and logratio model gives the best experimental results in terms of local uncertainty and closeness to true facies.

Introduction

A challenge of petroleum reservoir characterization is the integration of data of different type, scale, and precision. Typically local well logs and global seismic data are considered for building geostatistical models of the reservoir and these disparate data are used to account for a particular variable at every location. Geologic and/or geophysical sources to be used for building the models are fairly correlated to each other since they are informing about the same variable of interest.

One approach in geostatistics considers the integration of these sources of information under the full data independence assumption. This kind of independence assumption enables ones to simplify the data integration process, however, this independence assumption permits the final integrated probability to be above 1.0, which is not allowed in probabilities.

The conditional independence assumption is a second simplified way which is equivalent to the permanence of ratios model proposed by Journel (2002) (Journel 2002). However, as the number of correlated secondary variables increases, the fundamental assumption gets untenable which carries undesirable results. In probabilistic data integration, the main concern is to quantify redundancy of data sources since accounting for a data redundancy is critical and ignoring it can lead to bias and inconsistencies (Krishnan 2004).

The goal of this research is to describe new probabilistic integration methods considering data redundancy among correlated secondary variables and to compare other approaches with test examples.

Models of Integration

We have tested a categorical variable estimation with primary local well data and secondary continuous data sets. Let us define the following notations:

- Categories: $A_1,\ldots A_k$, k is the number of categories
- $IK(\mathbf{u}:A_k)$ is an indicator kriged probability of A_k at location $\mathbf{u}$
- Secondary variables: $z_1(\mathbf{u}),\ldots,z_m(\mathbf{u})$, m is the number of secondary variables
- Target posteriori probability of A_k at location $\mathbf{u}$: $\hat{P}(A_k)$

Tau-model expresses the target posteriori probability of A_k:

$$\hat{P}(A_k) = \frac{IK(\mathbf{u};A_k)}{P(A_k)}\left(\frac{P(A_k \mid z_1(\mathbf{u}))}{P(A_k)}\right)^{\tau_1}\cdots\left(\frac{P(A_k \mid z_m(\mathbf{u}))}{P(A_k)}\right)^{\tau_m} P(A_k)\cdot C$$

where τ_1 is redundancy factor between primary and first secondary variable

τ_m is redundancy factor between primary and m_{th} secondary variable

C is a common factor $P(A_k)P(z_1(\mathbf{u}))^{\tau 1}P(z_2(\mathbf{u}))^{\tau 2}\cdots P(z_m(\mathbf{u})^{\tau m}/P(A_k,z_1(\mathbf{u}),\ldots,z_m(\mathbf{u}))$

Tau value is a redundancy factor in tau-model and it can be interpreted as following (Journel 2002; Krishnan 2004):

- permanence of ratios when $\tau = 1.0 \Leftrightarrow$ conditional independence assumption
- secondary information is amplifying when τ is greater than 1.0 and influence of secondary information is decreasing when τ is less than 1.0
- τ is sequence-dependent weight

Linear data correlation can be used to obtain appropriate τ value such as:

$$\tau = 1.0 - \text{correlation(primary,secondary)}$$

or when amplifying the secondary information,

$$\tau = 1.0 + \text{correlation(primary,secondary)}$$

All τ values must be decided explicitly to use tau-model, however, we still have the problem of as how to get the τ value optimally. We propose a new technique that generalizes the tau-model. The new method introduces $\lambda_1,\ldots,\lambda_{m+1}$ to describe the data source redundancy such as:

$$\frac{\hat{P}(A_k)}{P(A_k)} = \left(\frac{IK(\mathbf{u};A_k)}{P(A_k)}\right)^{\lambda_1}\left(\frac{P(A_k \mid z_1(\mathbf{u}))}{P(A_k)}\right)^{\lambda_2}\cdots\left(\frac{P(A_k \mid z_m(\mathbf{u}))}{P(A_k)}\right)^{\lambda_{m+1}}\cdot C$$

where C is a common factor $P(A_k)^{\lambda 1}P(z_1(\mathbf{u}))^{\lambda 2}P(z_2(\mathbf{u}))^{\lambda 3}\cdots P(z_m(\mathbf{u})^{\lambda m+1}/P(A_k,z_1(\mathbf{u}),\ldots,z_m(\mathbf{u}))$

Taking the logarithm of the above equation we obtain,

$$log\left(\frac{\hat{P}(A_k)}{P(A_k)}\right) = \lambda_1 \log\left(\frac{IK(\mathbf{u};A_k)}{P(A_k)}\right) + \lambda_2 \log\left(\frac{P(A_k \mid z_1(\mathbf{u}))}{P(A_k)}\right) + \cdots + \lambda_{m+1} \log\left(\frac{P(A_k \mid z_m(\mathbf{u}))}{P(A_k)}\right) + \log(C)$$

Arbitrary λ values falling in [0,2] are to be tested with primary samples iteratively until they satisfy a given objective function. We used the least square error (to be minimized) and Markov-Bayes calibration factor B (to be maximized) as objective function such as:

$$Obj_1 = \overline{(1.0 - \hat{P}_{True})^2}$$

$$Obj_2 = \overline{B_k} \quad \text{for all k}$$

$$\text{where, } B_k = E\{\hat{P}(A_k(u)) \mid I(u;k) = 1\} - E\{\hat{P}(A_k(u)) \mid I(u;k) = 0\}$$

Case Study

Four data sets are prepared to estimate categorical variables: one primary local well data and three continuous secondary data that are transformed into Gaussian space individually (shown in Figure-1). Secondary variables are sampled exhaustively. Areas where noted as Facies 1 have higher secondary data values and areas noted as Facies 2 have lower secondary data values. The linear correlation coefficient among all secondary variables is 0.8 on average.

Facies estimation only using primary well data is performed by indicator kriging and secondary data is calibrated to generate probability map of each facies separately. Figure-2 represents the process of generating a probability map from both secondary data and primary well samples. Histograms of collocated secondary data values are built and the proportion of belonging to each facies is assigned to the grid node based on the built histograms. Lower part of Figure-2 is an example of proportion map of Facies 1 using the first secondary data.

For the evaluation of facies estimate we adapted a quantitative measure of goodness; classical entropy defined by Shannon in information theory and closeness to true facies. Closeness to true facies is defined by (Deutsch 1998):

Closeness to True facies: $C = \overline{\frac{C_k - P(A_k)}{P(A_k)}}$ for all k, where $C_k = E\{\hat{P}(u;A_k) \mid true\ facies = k\}$

Data integration models, permanence of ratios, tau-model (amplifying and decreasing the influence of secondary information) and logratio model, are applied for the same data sets and compared in terms of this goodness measure. Table-1 and -2 shows evaluation results both for each facies and overall. Tau model is tested with two different schemes: amplifying secondary information with $\tau > 1$ and decreasing influence of secondary information with $\tau < 1.0$. Any integrated model provides better estimate results than indicator kriged estimate only

using primary data. In tau-model, the case of amplifying the secondary information was better than the case of decreasing the influence of secondary. Logratio model shows the best performance among integration models and improves closeness and entropy measures at least by 11%.

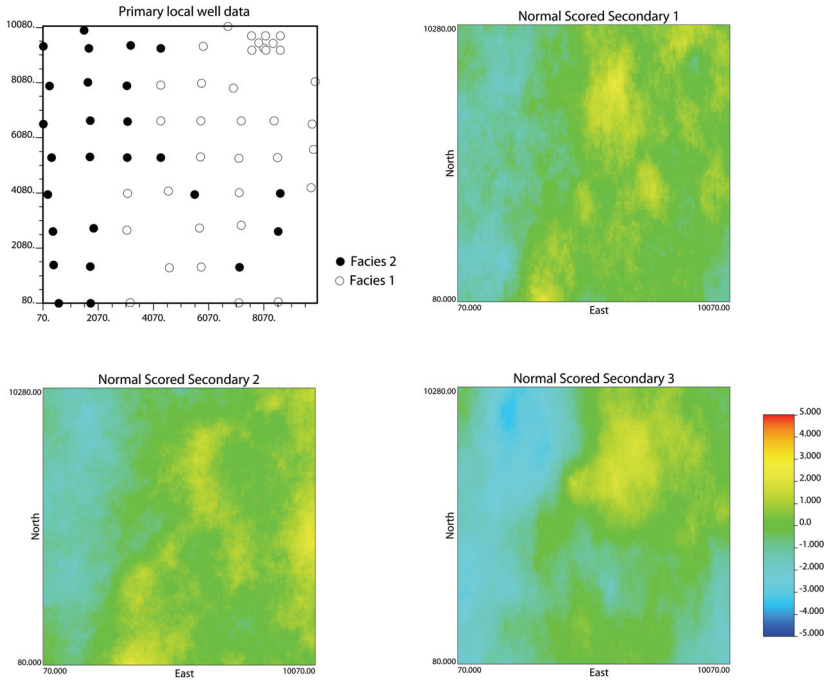

Figure-1: Primary and secondary data sets used in the study. All secondary variables are normal scored values after univariate Gaussian transformation.

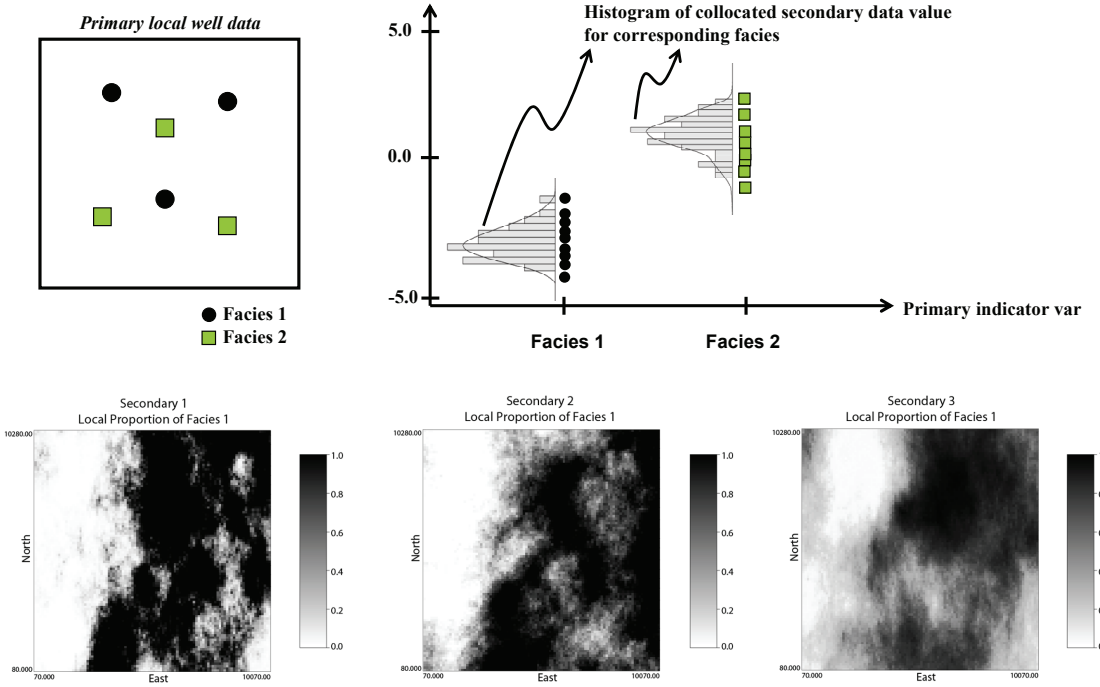

Figure-2: Process of secondary data calibration. Lower figures are examples of probability map from the secondary data set.

Table-1: Computed closeness to true facies

Closeness to True Facies						
Category	**IK**	**PR**	**Tau with 1+corr**	**Tau with 1-corr**	**Logratio-LSE***	**Logratio-B****
1	0.3871	0.6283	0.6412	0.5794	0.6228	0.6302
2	0.8261	1.1063	1.1706	0.9274	1.2884	1.2388
Overall	**0.6066**	**0.8673**	**0.9058**	**0.7534**	**0.9556**	**0.9345**

Table-2: Computed entropy

Entropy						
Category	**IK**	**PR**	**Tau with 1+corr**	**Tau with 1-corr**	**Logratio-LSE***	**Logratio-B****
1	0.2938	0.0605	0.0421	0.1549	0.0374	0.0404
2	0.2442	0.0982	0.0528	0.1668	0.0458	0.0454
Overall	**0.2730**	**0.0763**	**0.0474**	**0.1599**	**0.0416**	**0.0429**

* Least square error objective function is used as an objective function
** Markov-Bayes B calibration factor is used as an objective function

Discussion

Combining conditional probability for disparate data integration is described. Merging conditional probability does not require multivariate Gaussian or any type of parametric assumption, which is usually unattainable in general. However, choosing appropriate data redundancy factors is a critical point when one combines individual conditional probabilities.

Logratio model for integrating disparate data has been suggested and we used an iterative scheme to find optimal redundancy weights. Effectiveness of the proposed method is tested with highly correlated secondary data sets. The advantage of this algorithm is in finding optimal redundancy factor that accounts for correct primary sample data value, and in adjusting the influence of secondary information. Logratio model represents the best estimate performance having lowest entropy and highest closeness to true.

References

Journel, A. G. [2002] Combining Knowledge from Diverse Sources: An alternative to Traditional Data Independence Hypotheses. Mathematical Geology Vol. 34, No. 5

Krishnan, S. [2005] Experimental Study of Multiple-support, Multiple-point Dependence and its modeling. Geostatistics Banff 2004 Vol. 2, Springer, Dordrecht

Deutsch, C. V. [1998] A Short Note on Cross Validation of Facies Simulation Methods, Centre for Computational Geostatistics Annual Report No. 1