On Secondary Data Integration

Sahyun Hong and Clayton V. Deutsch

A longstanding problem in geostatistics is the integration of multiple secondary data in the construction of high resolution models. In the presence of multiple secondary data, a workflow for modeling would involve: (1) integrate secondary data into unit of the primary variable, then (2) integrate secondary derived estimates with the well data. Results from the first integration step will be a probability map/cube or estimate of the continuous variable. These intermediate results are integrated with well data through spatial modeling. Variants of kriging are well established techniques to account for spatial variability. A large variety of methods are used for integrating the secondary data. Probability combination schemes have received much attention recently for this. The main challenge with these probability combination schemes (PCS) is fair consideration of redundant secondary data. Several combination models to meet this challenge are reviewed. Their limitations and possible applications are described. As an alternative to combining probability, direct modeling the multivariate distribution between secondary and primary variables is advanced. In this method, joint distribution is modeled in a nonparametric way and it is refined under marginal constraints. A simple and fast algorithm is developed to impose marginal conditions on initial joint distributions. The effectiveness of the discussed methods is demonstrated through some examples with related references.

Introduction

Building numerical reservoir models is an intermediate but essential step for reservoir management. Numerical models are usually used to plan new wells, calculate overall hydrocarbon reserves and to predict the reservoir performance in a flow simulator. Accurate reservoir models will lead to accurate predictions of reservoir performance and improve reservoir management decisions. Thus, constructing numerical geologic models is an important step in reservoir management. Accurate reservoir modeling, however, is difficult to achieve given sparse well data; the reservoir properties such as facies, porosities, permeabilities and fluid saturations are typically sampled at few well locations. Due to this sparse knowledge the built reservoir models are poorly constrained away from well locations, which leads to considerable uncertainty in the spatial distribution of reservoir properties.

Secondary data helps reduce uncertainty in a numerical model. Diverse auxiliary data sources are commonly available in petroleum applications. These data include well test data, seismic data, analog outcrops, and conceptual geological interpretations. Each data source carries information on the reservoir properties at different scales and with varying levels of precision. The main purpose of utilizing these secondary data is to provide accurate models and to reduce the uncertainty in the reservoir performance predictions. The idea is to integrate various data in a consistent manner to characterize reservoir properties of interest. All of the information must be integrated into comprehensive and consistent representation of the reservoir.

Primary and Secondary Data

Data used for reservoir modeling is divided into two types: primary and secondary data. Direct measurements of reservoir properties are denoted as primary or hard data, while data that provide indirect measurement are denoted as secondary or soft data. Primary data is a direct measurement of the properties being predicted, but they are sparsely distributed over the reservoir. Well logs and core data are treated as the primary data. Secondary data are usually acquired from geophysical surveys, geologic interpretations and/or previously simulated variables being related to the primary variable; thus, they have wide coverage. Seismic data is an important source of secondary information. Several seismic attributes are extracted from a raw data and they could be good indicators of reservoir properties being predicted with varying degree of relation. In addition to quantitative seismic data, geologic maps derived from geologist or analogue information are valuable data that should be accounted for during reservoir modeling.

The uncertainty in the constructed model will decrease with additional data sources, however, it is not easy to reconcile various data because they have varying scale and spatial coverage. Secondary data that typically have larger scale than the modeling scale. Multiple secondary are first integrated together, and secondary derived estimates are then combined with small scaled primary data. When integrating secondary data first, it is crucial to properly account for data redundancy that represents how closely secondary data are related with respect to the primary data. Data redundancy may not fully explained by correlation coefficient or covariance among data and it arises from data interaction that is usually better described in the nonlinear manner.

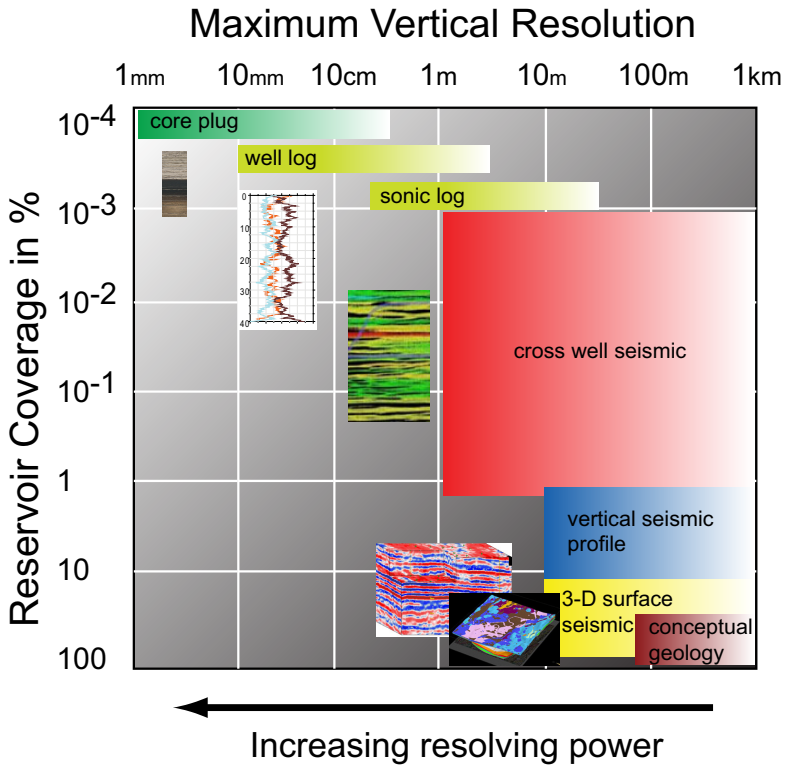

Figure-1: An illustration showing various scales and spatial coverage of data (modified from Harris and Langan, 1997).

A Workflow for Reservoir Modeling with Secondary Data

Procedure of geostatistical reservoir modeling with secondary data can be divided into two parts; secondary data are first integrated generating probability or probability distribution related to the primary variable, and then they are integrated with primary data. The sketch shown in the figure-2 demonstrates the overall workflow for reservoir modeling in the presence of multiple secondary data. Exhaustiveness is an inherent assumption on the secondary data. Qualitative map such as geologic interpretation should be converted into digitized images. Data aggregation step is aimed at reducing the number of secondary data by merging the highly correlated secondary data (Babak and Deutsch, 2007). Aggregating step should be performed when too many secondary data, e.g. more than 6, are initially prepared. Merged data will be treated as new secondary data for a subsequent integration. In the first integration step, the primary data is used only for calibrating the relation between primary and a set of secondary data. The spatial correlation of the primary data is not considered in this step. No scale different is assumed among gridded secondary data. As a result of first step, several secondary data are converted into a single probability or probability distribution term summarizing all secondary information. Relevance of the secondary data to the primary data is fully accounted for in this step. For instances, higher acoustic impedance represents lower porosity or lower proportion of porous rock, and this relation is quantified through probability distribution modeling.

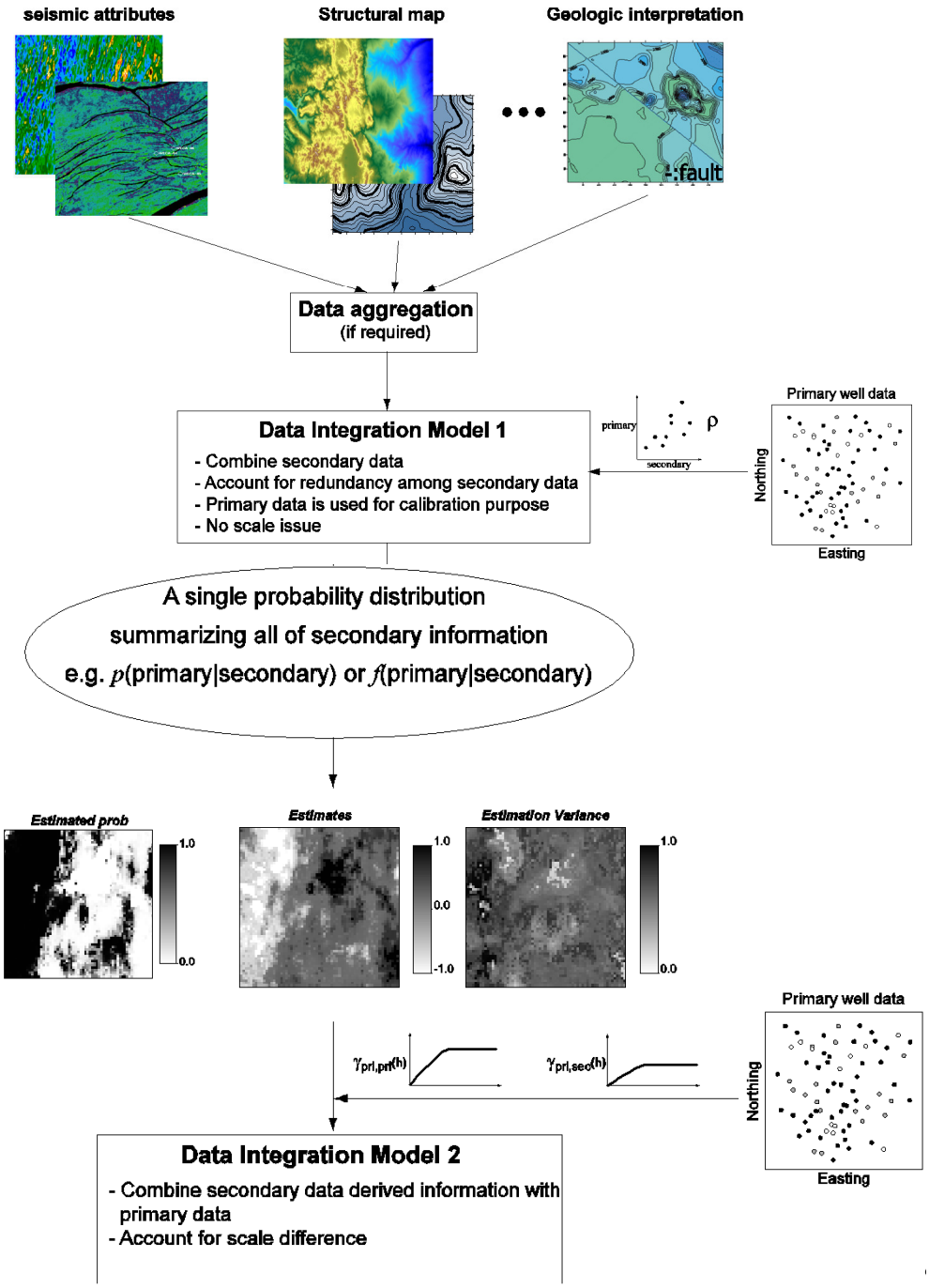

Figure-2: A workflow of geostatistical reservoir modeling in the presence of secondary data

Accounting for data redundancy is crucial in the secondary data integration. Results will be highly biased if data redundancy is not properly considered. Figure-3 illustrates the effect of incremental information. Facies probability is estimated at an unsampled location from no secondary to several secondary data. Each secondary data has a linear correlation of 0.46 to the primary variable, and secondary data themselves have linear correlations between 0.46 – 0.82. Probabilities are estimated from secondary data with and without accounting for data redundancy, and they are plotted as filled rectangles and diamonds in the figure-3. It is a simple and reasonable view to assign the global proportion at an unsampled location when no data is considered. If the related secondary data are available they would

give more confident information about an unsampled location than global proportion leading to an increase of the estimated probability. The estimated probabilities; however, unfairly become close to 1 when redundancy is simply ignored in which case the relevance of each secondary data to the primary variable is fully considered. Bias of the estimated probability is reduced when data redundancy is accounted for.

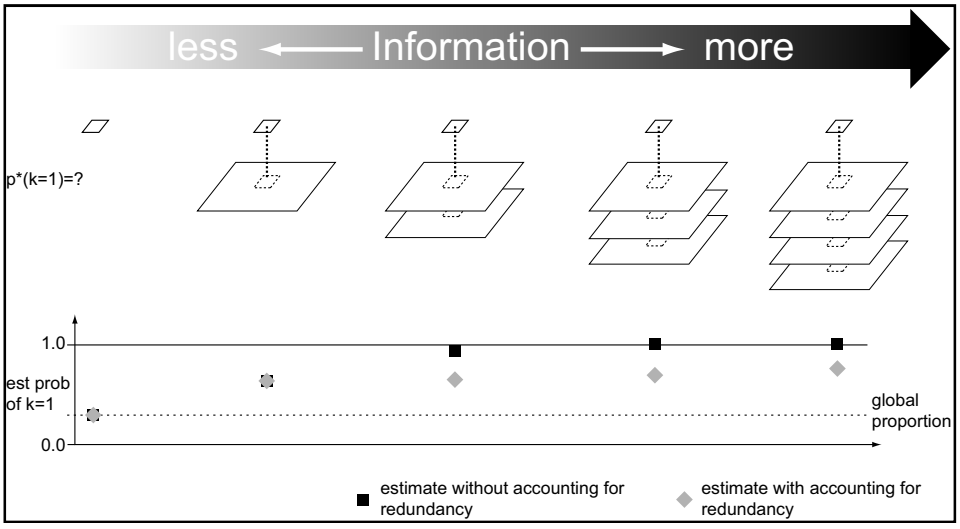

Figure-3: An example of illustrating incremental information impact on the probability estimate. Probabilities are estimated with and without accounting for data redundancy.

The second part in the overall workflow (figure-2) is to combine the primary data and the secondary data-derived probabilities or probability distribution. Spatial (cross) variability is modeled in this step. Although multiple secondary data is initially considered, a single calibrated secondary variable is used hereafter because former step integrates the multiple secondary data converting them into a single secondary-derived variable. The effort of cross variogram modeling is considerably reduced; one cross variogram modeling is necessary regardless of the number of secondary data. The secondary data themselves could be used as secondary variables for estimation and simulation without first integration step. The secondary data calibration enters through the modeling of cross variograms between the primary and secondary data. Relevance and redundancy of secondary data are implicitly modeled in the cokriging equation. Despite of its flexibility of cokriging, two step modeling process is preferred because: (1) inference of cross variogram becomes tedious in a direct use of secondary data, (2) non-linear relation among secondary data can be modeled in the secondary data integration, and (4) the integrated results themselves give useful information about the spatial variability of primary variable which potentially could be used for locating new wells.

Variants of kriging have been well established techniques for the second integration step of workflow demonstrated in the figure-2. This overview paper laid more emphasis on several methods for the first integration step. In particular, probability combination scheme and direct modeling the joint pdf scheme. For clarity, we will focus on the modeling of a categorical variable. An application to the continuous variable modeling is introduced at the end of the paper with related references.

Probability Combination

Probability combination schemes (PCS in short) have been developed independently in many research areas in order to find a consensus probability using several single source derived probabilities (Winkler, 1981; McConway, 1981; Lee et al., 1987; Benediktsson and Swain 1992, Journel, 2002). Main principle of probability combination approaches is to approximate the target probability through linking the individual probability that is computed using individual secondary data. In a geostatistical data integration context, our interest is to estimate the probability of primary variable that is jointly conditioned to secondary data and this probability will be estimated by combining individual probability that are obtained from calibrating each secondary data (not jointly conditioning). Figure-4 shows a schematic diagram of PCS.

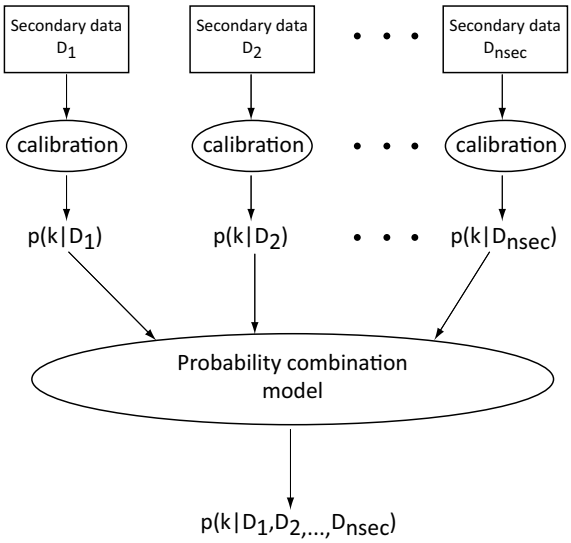

Figure-4: A diagram for showing PCS

The target probability to be estimated is expressed as $p(k|D_1,\ldots,D_{nsec})$ where primary and secondary data are denoted as k=1,…,K and D_i, i=1,…,nsec. To generate elementary probabilities, $p(k|D_i)$, secondary data are calibrated individually. In the PCS, different types of secondary data can be calibrated using the most appropriate method for each data. In its general mathematical form, the probability is approximated by:

$$\frac{p(k \mid D_1,\ldots,D_{n\sec})}{p(k)} = \Phi\left[\frac{p(k \mid D_1)}{p(k)},\ldots,\frac{p(k \mid D_{n\sec})}{p(k)}, C\right] \tag{1}$$

Probability terms independent of primary variable k are absorbed in a normalizing term C. $p(k)$ is a global proportion of k=1,…,K. Function $\Phi[\cdot]$ is a generic notation of probability combination model and it would have various names such as permanence of ratio, tau and lamda model.

Permanence of Ratios (PR-model)

Journel (2002) developed a permanence of ratios model that approximates the probability under assuming that ratios of probability increments from different sources are constant. PR model estimates the probability as:

$$p_{PR}(k \mid D_1,\ldots,D_{n\sec}) = \frac{\left(\frac{1-p(k)}{p(k)}\right)^{n\sec-1}}{\left(\frac{1-p(k)}{p(k)}\right)^{n\sec-1} + \prod_{i=1}^{n\sec}\left(\frac{1-p(k \mid D_i)}{p(k \mid D_i)}\right)} \in [0,1] \tag{2}$$

The estimated probability by PR model meets closure condition and positiveness regardless of number of data D_i. The PR model; however, is limited to only binary case k=1 or 2. For instances, sum of $p_{PR}(k|D_i,i=1,\ldots,nsec)$ over k=1,…,K where K is greater than 2 does not always amount to 1. A numerical example of this violation is demonstrated in Hong and Deutsch (2009).

Conditional Independence

Assumption made in PR model is same as the independence assumption of $(D_1,\ldots,D_{nsec})$ conditioned to primary variable k=1,2. Equivalence of PR model and conditional independence is proved for the binary case. Under conditional independence assumption of $(D_1,\ldots,D_{nsec})$, the generic equation for combining conditional probabilities shown in the equation (1) becomes:

$$p(k \mid D_1,\ldots,D_{n\sec}) = p(k)\frac{p(k \mid D_1)}{p(k)}\times\cdots\times\frac{p(k \mid D_{n\sec})}{p(k)}C = p(k)\prod_{i=1}^{n\sec}\left(\frac{p(k \mid D_i)}{p(k)}\right)C \in [0,1] \tag{3}$$

C is a normalizing term to meet the closure condition and it is independent of primary variable k. It will be removed by normalizing:

$$p_{CI}(k \mid D_1,...,D_{n\sec}) = \frac{p(k)\prod_{i=1}^{n\sec}\left(\frac{p(k \mid D_i)}{p(k)}\right)C}{p(k)\prod_{i=1}^{n\sec}\left(\frac{p(k \mid D_i)}{p(k)}\right)C + (1-p(k))\prod_{i=1}^{n\sec}\left(\frac{1-p(k \mid D_i)}{1-p(k)}\right)C}$$

$$= \frac{1}{1+\left(\frac{1-p(k)}{p(k)}\right)\left(\frac{p(k)}{1-p(k)}\right)^{n\sec}\prod_{i=1}^{n\sec}\left(\frac{1-p(k \mid D_i)}{p(k \mid D_i)}\right)}$$

$$= \frac{\left(\frac{1-p(k)}{p(k)}\right)^{n\sec-1}}{\left(\frac{1-p(k)}{p(k)}\right)^{n\sec-1} + \prod_{i=1}^{n\sec}\left(\frac{1-p(k \mid D_i)}{p(k \mid D_i)}\right)} = p_{PR}(k \mid D_i, i=1,...,n\sec) \in [0,1]$$

This equivalence holds for binary case k=1,2. PR model is not mathematically valid when more than two categories are to be considered. Instead conditional independence should be used because it enforces the closure condition, for example when ternary facies are occurring:

$$p_{CI}(k \mid D_1,...,D_{n\sec}) = \frac{p(k)\prod_{i=1}^{n\sec}\left(\frac{p(k \mid D_i)}{p(k)}\right)}{p(k)\prod_{i=1}^{n\sec}\left(\frac{p(k \mid D_i)}{p(k)}\right) + p(k')\prod_{i=1}^{n\sec}\left(\frac{p(k' \mid D_i)}{p(k')}\right) + p(k'')\prod_{i=1}^{n\sec}\left(\frac{p(k'' \mid D_i)}{p(k'')}\right)} \in [0,1] \quad (4)$$

where $k' \neq k'' \neq k$=1,2,3.

Weighted Combination

Permanence of ratios and conditional independence model all assumed independence among secondary data. By adopting independence assumption, combing probabilities is simplified into the product of individual probabilities, $p(k|D_i)$. In some cases, the simplified model could make a serious bias because multiplication of each probability result in a very high combined probability possibly far away from $p(k)$ and all $p(k|D_i)$, i=1,…,nsec. In particular, the resulting probability becomes very close to 1 or 0 as seen in the introductory example in the figure-3 if many secondary data are considered for integrating and they are highly redundant.

Weighted combination approaches are advanced to adjust the influence of elementary probabilities. The probability is approximated by combining individual probabilities with data specific weights:

$$\frac{p(k \mid D_1,...,D_{n\sec})}{p(k)} = \Phi\left[\left(\frac{p(k \mid D_1)}{p(k)}\right)^{w_1},...,\left(\frac{p(k \mid D_{n\sec})}{p(k)}\right)^{w_{n\sec}}, C\right] \quad (5)$$

One of the weighted model is the Tau model and the model has been actively used for combining data specific probabilities since its first development by Journel (2002) (Krishnana, 2004; Caers et al., 2004; Castro et al., 2006; Chugunova and Hu, 2008). Tau model approximates $p(k|D_i, i$=1,…,nsec) by imposing power τ_i:

$$p_{Tau}(k \mid D_1,...,D_{n\sec}) = \frac{\frac{p(k)}{1-p(k)}\left(\frac{1-p(k)}{p(k)}\right)^{\sum_{i=1}^{n\sec}\tau_i}}{\frac{p(k)}{1-p(k)}\left(\frac{1-p(k)}{p(k)}\right)^{\sum_{i=1}^{n\sec}\tau_i} + \prod_{i=1}^{n\sec}\left(\frac{1-p(k \mid D_i)}{p(k \mid D_i)}\right)^{\tau_i}} \quad (6)$$

Tau model has the same form of PR model except introducing power weights τ_i that controls the contribution of elementary probabilities $p(k|D_i)$. Similar to the PR model, Tau model only works for binary case, k=1 and 2. There is no guarantee that sum of $p_{Tau}(k|Di, i$=1,…,nsec) over k=1,…,K amounts to 1.

Lamda model was developed by Hong and Deutsch (2007). It can be interpreted as an expanded model of conditional independence with introducing data specific weights, λ_i. If $\lambda_i=\tau_i$ for i=1,…,nsec and binary category is only considered then p_{PR} and p_{Lamda} are exactly same. Ladma model is not limited to binary case. For example, the combined probability can be obtained as following if ternary facies is considered:

$$p_{Lamda}(k \mid D_1,...,D_{n\sec}) = \frac{p(k)\prod_{i=1}^{n\sec}\left(\frac{p(k \mid D_i)}{p(k)}\right)^{\lambda_i}}{p(k)\prod_{i=1}^{n\sec}\left(\frac{p(k \mid D_i)}{p(k)}\right)^{\lambda_i} + p(k)\prod_{i=1}^{n\sec}\left(\frac{p(k' \mid D_i)}{p(k')}\right)^{\lambda_i} + p(k'')\prod_{i=1}^{n\sec}\left(\frac{p(k'' \mid D_i)}{p(k'')}\right)^{\lambda_i}} \in [0,1] \quad (7)$$

where k′≠k′′≠k=1,2,3. Thus, the Lamda model is a more generalized weighted combination model.

In weighted combination models, choosing appropriate weights is a critical issue (Krishnan, 2004). Optimal weights would be found if data redundancy among secondary data are fully characterized and can be expressed just by power. However, this is almost impossible. As an alternative to direct quantifying redundancy, calibration method may be considered. For instances, weights are obtained in order to minimize the errors between the true value at well locations and the estimated probability p_{Lamda} of true value at that location (Hong and Deutsch, 2007).

Non-convex Property

Non-convex property represents the integrated probability is not within the used input probability $p(k|D_i)$ and global proportion $p(k)$. Non-convexity is natural in data integration. Integrating diverse data amplifies the impact of data sources if they represent the same direction of probability. This can be seen in the form of mathematical expressions shown in equations (2) – (7): multiplication of probability ratios make much higher or lower resulting probability. Degree of non-convexity is also affected by how redundant data D_i are among themselves. Although all of elementary probabilities $p(k|D_i)$ are in the same direction (they are all higher than the global $p(k)$), non-convexity may not be significant when data are highly redundant, and thus the combined probability should not be very high or very low for this case. Incorrect weights can make the resulting probability be very high or low. Again, finding weights is very important in combining probability method.

Applications of PCS

There are no generalized approaches to calculate data specific weights in the PCS. Polyakova and Journel (2007) derived an exact analytical expression dominated only by a single parameter ν, however, approximation is still required for ν value that is case-dependent. Nevertheless, the principle of combining probability can be applied to other applications such as full trend modeling in 3D by combining lower order trends.

Large scale trend is important feature that should be accounted for in the final geostatistical model, however, unfortunately there is no geostatistical technique to account for the trend in an implicit way. A typical way for 3D trend modeling could be divided into three steps (Deutsch, 2002): (1) model the areal trend, (2) model the vertical trend against vertical coordinate, and (3) merge 2D areal and 1D vertical trend into 3D trend. Soft information such as seismic map often helps for identifying areal variations. For full 3D proportion modeling, combining lower order proportions has been considered since it is easier to fit the vertical and areal proportion than direct modeling a 3D proportion. Hong and Deutsch (2009) demonstrated detailed examples about 3D trend modeling using a weighted proportion combination method. In this paper, some key features are summarized. Integrating the 2D and 1D proportion that may be modeled by different data sources can be viewed as a probability combination problem. Full 3D proportion $p_k(x,y,z)$ can be approximated as following:

$$p_k(x,y,z) = \frac{p_k(x,y)}{p_k}\frac{p_k(z)}{p_k}p_k C \in [0,1] \quad (8)$$

In this form, the 3D proportion is estimated from the multiplication of areal and vertical proportion standardized by global proportion. This approximation also adopts an independence assumption between 2D and 1D proportions conditioned to the estimated 3D proportion. Main drawback of the simple method is that the combined trend value can be unfairly larger or lower than both input aerial and

vertical trend value leading to a large variability in 3D trend model of which smoothness is an intrinsic characteristic.

As a way of reducing the possibility of falling outside the input proportion, a weighted combination approach is advanced:

$$p_k(x,y,z) = \left(\frac{p_k(x,y)}{p_k}\right)^{w_1} \left(\frac{p_k(z)}{p_k}\right)^{w_2} p_k C \in [0,1] \tag{9}$$

where w_1 and w_2 are weights imposed on each proportion ratio. The weighted model reverts to the conditional independence model with letting w_1=w_2=1. The proportion ratio has a minimum bound of 0, but theoretically it has no maximum bound. The normalizing term C is disappeared by enforcing the closure condition ($\sum_{k=1}^{K} p_k(x,y,z) = 1$). Figure-5 below shows how the proportion ratio (denoted as X in the figure) changes according to the change of weight. By imposing weights lying within [0,1], the ratio is decreased if it is larger than 1, and the ratio is increased if it is smaller than 1. In other words, proportion ratios are forcibly reduced when they tend to be extremes.

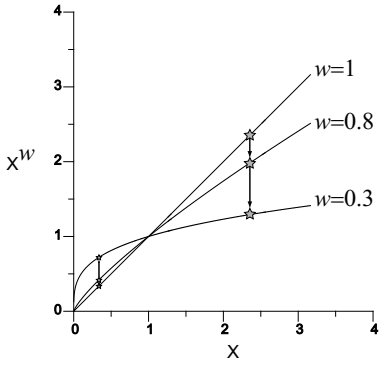

Figure-5: An illustration of how the proportion ratio change according to weight in [0,1]. As smaller weights are used, the weighted proportion ratio is bounded in shorter range.

In a weighted combination approach, we can prevent the multiplication of proportion ratios, finally the combined proportion, from being too high or too small, and consequently, it may lead to better reproduction of input trends.

Another potential application of probability combination schemes is to aggregate the information that is obtained from the data integration process. Probability combination at this level should be differentiated from PCS at data integration level. *Data integration* is referred to as assimilating the available data such as well data, seismic data, TI and others. Integrated results are to be various and different based on the choice of integration methodology. Different integrated results sometimes need to be considered together as integration procedure often requires the subjective user-input parameters and algorithm settings. Evaluating of multiple results will ensure the production of the most satisfactory possible. Pooling different integration results, conditioned to the same data sources, involves procedure with the goal of combining results to find consensus estimates. This is termed *information integration*. Figure-6 illustrates data integration and information integration.

In data integration step, non-convexity is a significant characteristic. Various integrated predictions are resulted from different choice of integration models or parameters with the same data. Information integration aggregates those decisions to provide consensus predictions. The consensus must be found by weighted linear averaging: convexity is a desirable property in information integration,

$$p_{\text{consensus}}(k \mid D_1,\ldots,D_m) = \frac{1}{\sum_{i=1}^{N}\alpha_i} \sum_{i=1}^{N} \alpha_i p_i^{\text{integrated}}(k \mid D_1,\ldots,D_m)$$

Weights $\alpha_i \in [0,1]$, i=1,…,N might be selected in terms of integration model reliability, accuracy or other criterion. Linear averaging lets consensus probability exist between the integrated results.

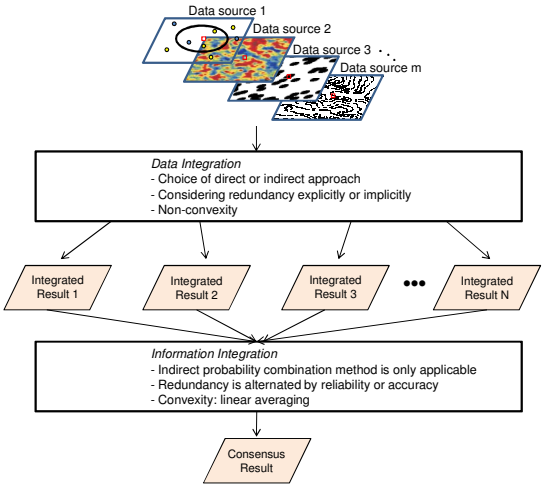

Figure-6: Data integration and information integration.

Direct Modeling of Joint Distribution

In probability combination approaches, the probability of interest jointly conditioned to secondary data is approximated by combining elementary probabilities that are conditioned to single secondary data. Another method for estimating the probability of interest is to directly model the joint distribution consisting of all of primary and secondary variables. Once the joint distribution is modeled, the conditional probability is immediately derived from the joint pdf using Bayes law. The direct pdf modeling method is motivated by some points: (1) there are many nonparametric techniques for modeling the joint distribution among variables, (2) one characteristic of secondary data is exhaustiveness and so the joint pdf of secondary data is ignored in the PCS approaches can be modeled very reliably, (3) by modeling the joint distribution directly, data redundancy among variables is inherently accounted for not requiring an external redundancy calibration. The sketch shown in the figure-7 illustrates the diagram of the method.

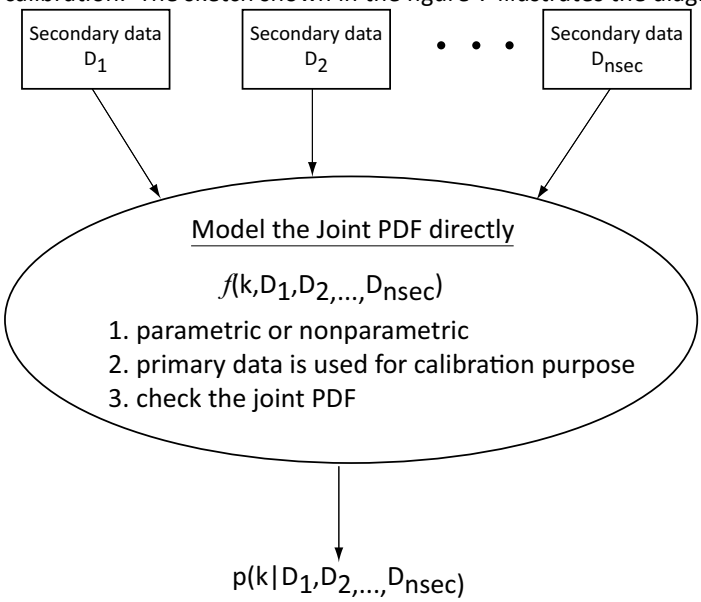

Figure-7: A schematic illustration for integrating secondary data through building the joint pdf directly.

Build joint PDF by Nonparametric Modeling Method

Nonparametric density modeling techniques are distribution free methods that do not rely on the assumption that the data are drawn from a given probability distribution. Among various nonparametric density modeling techniques, kernel density estimator has been widely used and most investigated

mathematically (Scott, 1992). The flexibility of kernel density estimator is placed at the expenses of computational costs. The method evaluates densities at every bin where we want to estimate density. For example, if we have 3 variables (1 primary and 2 secondary variables), 30 sample data and we want density estimates at every 50 bins of each variable, then total of $50^{(3)} \times 30 = 3{,}750{,}000$ calculations are required for constructing a 3-dimentional probability density functions. The complexity is expressed in general:

$$B^{(\#\text{ of variables})} \times (\#\text{ of samples}) \tag{10}$$

where B is number of bins. Practical implementation would limit the number of variables by merging secondary variables (data aggregation shown in the figure-2). More than 5 secondary variables would be better merged into less than 5 by aggregating closely related variables.

Constraining joint PDF with Marginal Conditions

The modeled joint probability distribution $f(k,D_1,\ldots,D_{nsec})$ must satisfy all axioms of probability distributions: non negative density functions, closure condition and reproduction of lower order marginal distributions. Kernel density estimator meets first two axioms if the used kernel function $W(\cdot)$ follows $W(x)\geq 0$ and $\int W(x)dx=1$. The third condition is a marginality condition that the p-variate joint distribution should reproduce p'-variate distribution where $p' < p$. The followings are possible marginal conditions that the modeled joint distribution $f(k,D_1,\ldots,D_{nsec})$ must meet:

$$\int \cdots \int f(k,D_1,\ldots,D_{n\sec})dD_1 \cdots dD_{n\sec} = p(k) \tag{11}$$

$$\int f(k,D_1,\ldots,D_{n\sec})dk = f(D_1,\ldots,D_{n\sec}) \tag{12}$$

There is no guarantee, however, that the modeled joint distributions meet these marginal conditions. $f(k,D_1,\ldots,D_{nsec})$ is modeled based on the limited samples (n) that is much less than the number of secondary values that constitute the marginal distribution $f(D_1,\ldots,D_{nsec})$. The collocated secondary data at the sample locations normally do not cover the full range of the secondary data. The marginal distribution from the joint distribution may not match the secondary marginal distribution. Figure-8 illustrates this case. The bivariate distribution f_{KDE} (light black line) is modeled using four data points (black dots). Integration of the bivariate distribution over the primary variable (shown as dashed line on horizontal axis) has less variability and nearly zero densities outside the collocated secondary values even if there are some non zero frequencies over that range. Thick solid line represents a secondary data pdf denoted as $f_{reference}$ and it is built from large number of samples. $f_{reference}$ have variations in densities through the entire range of secondary values.

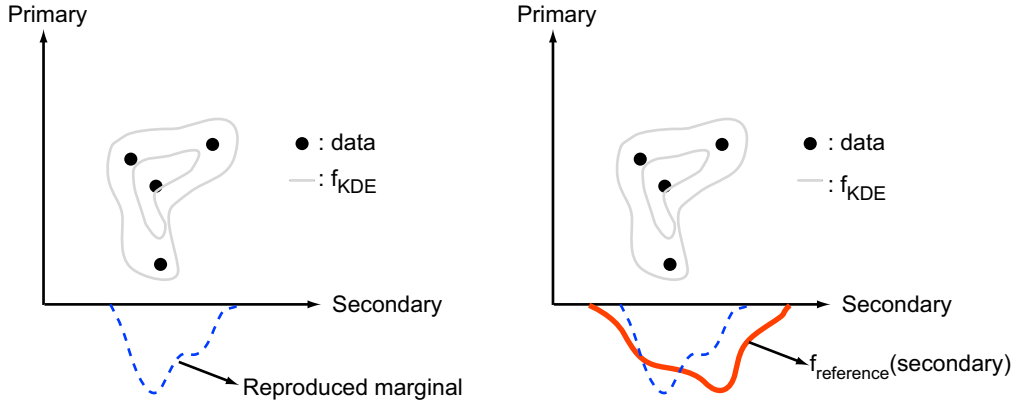

Figure-8: A schematic illustration for comparing the reproduced marginal with the known marginal. Since the joint distribution is modeled with the limited well samples its reproduced marginal is not consistent with the (very well-known) marginal pdf which is a distribution of secondary data.

Besides, if the global proportion is different form a naïve proportion, for example, after declustering then the marginality condition (11) is not achieved.

Given the marginal relations described in (11) and (12), an algorithm is proposed to impose them on the joint probability distribution. The marginals are derived from initial joint distribution and compared with the reference marginals. The differences are directly accounted for in order to adjust the initial distributions. This correcting process is performed in the following steps:

Step1. Model the joint distribution of secondary data, $f(D_1,\ldots,D_{nsec})$ and global proportion of k, $p(k)$. Declustering is considered for obtaining unbiased $p(k)$ if required.

Step2. Model the joint distribution $f(k,D_1,\ldots,D_{nsec})$ using kernel method and define it as $f^{(0)}$ to differentiate from the resulting joint distribution.

Step3. Scale the $f^{(0)}$ to ensure the marginal distribution shown in equation (12). The scaling equation below is proposed for ensuring the imposed marginal condition:

$$f^{(0)}(k,D_1,\ldots,D_{n\sec})\times\frac{f(D_1,\ldots,D_{n\sec})}{\int f^{(0)}(k,D_1,\ldots,D_{n\sec})dk}\rightarrow f^{(1)}(k,D_1,\ldots,D_{n\sec})$$

Step4. Scale the $f^{(1)}$ to ensure the marginal distribution shown in equation (11). Similar to the step 3, the scaling equation below is for updating the $f^{(1)}$:

$$f^{(1)}(k,D_1,\ldots,D_{n\sec})\times\frac{p(k)}{\int\cdots\int f^{(1)}(k,D_1,\ldots,D_{n\sec})dD_1\cdots dD_{n\sec}}\rightarrow f^{(2)}(k,D_1,\ldots,D_{n\sec})$$

Step5. Finish the procedures if stopping rule is met, otherwise go to step 6.

Step6. Reset $f^{(2)}$ into $f^{(0)}$ and repeat through steps 3 and 5.

Step 1 and 2 are initial steps to establish the marginal distributions $p(k)$ and $f(D_1,\ldots,D_{nsec})$. Steps 3 through 5 are employed to correct the initial distribution with the considered marginal distributions. The correction is performed by directly accounting for the differences. Step 5 terminates successive adjustments done through step 3 and 4 when the joint distribution becomes stable. A satisfactory stopping rule is to decide on a threshold δ, for example, δ = 0.1 or 0.01, and stop when a complete correcting cycle (steps 3 and 4) does not cause changes in distribution by more than pre-defined threshold δ. Another stopping rule could be the error between the reproduced and reference marginal distributions. The algorithm stops when the error becomes less than a specified tolerance:

$$e_1=\mid f^{repro}(D_1,\ldots,D_m)-f^{ref}(D_1,\ldots,D_m)\mid,\ e_2=\mid p^{repro}(k)-p^{ref}(k)\mid$$

and

$$e=(e_1+e_2)/2$$

The proposed sequential updating algorithm can be visualized for better understanding. Figure-9 demonstrates the joint probability distribution with one primary (k) and two secondary data (D_1,D_2). Initial joint distributions are modeled based on the limited collocated samples leading to smooth distributions. Big arrows shown in the middle of the figure represent the updating procedures under two marginal conditions; vertical direction is for applying the secondary variable marginal $f(D_1,D_2)$ (step 3) and horizontal direction is for applying the primary variable marginal $p(k)$ (step 4). Updated joint distributions are shown in the right of the figure.

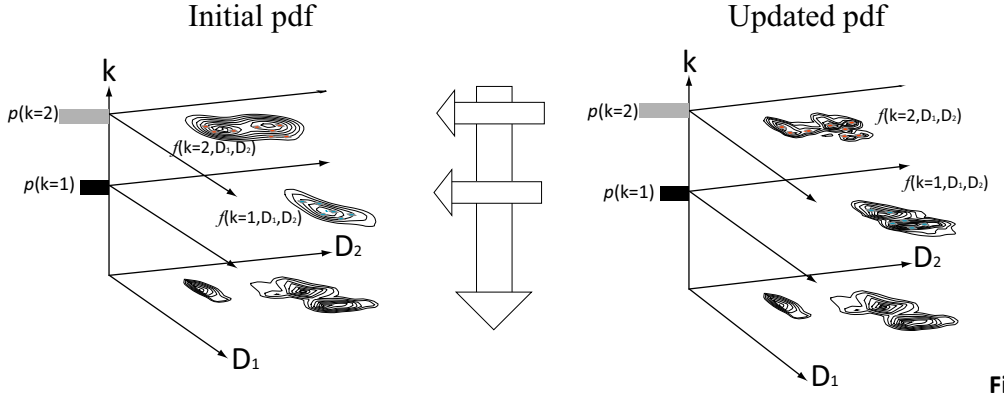

Figure-9: The joint probability distributions modeled by kernel estimator (left column) and updated distributions constrained by the imposed marginal conditions (right column). The arrow shown in the middle

represents the direction of marginal fitting: two horizontal directions are to compare the probability of facies $p(k)$, and vertical direction is to compare the secondary data distribution $f(D_1,D_2)$.

Marginal fitting procedures are sequentially repeated in the direction of big arrows until the stopping conditions are met. Figure-10 shows a result for an applied iterative fitting algorithm with real test data. Averaged marginal errors between empirical distributions and reference distributions are plotted against iteration numbers. First iteration drops the errors quickly and the averaged errors become less than 0.1% before 30 iterations. 100 iterations make errors very small where the resulting joint distributions become stable. Practices showed that 100 iterations are large enough to generate an updated distribution and it took a few minutes with 5 variables (1 primary + 4 secondary) on a 3.2 GHz personal computers.

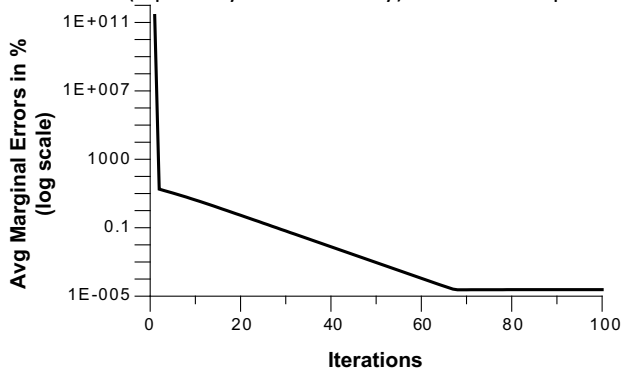

Figure-10: Averaged marginal errors versus iterations for a test data.

Incorporating Large Scale Secondary Data

It is important to integrate information from seismic data that delivers better areal coverage due to sparse sampling of well data. Seismic data; however, inexactly measures facies proportions because of geological complications and inherent limitations in seismic data acquisition (Deutsch, 1996). Poor vertical resolution of the seismic data warrants using a single 2D map representing vertically averaged facies proportions and the volume vertically averaged is significantly larger than the typical geological modeling cell.

The explained direct pdf modeling method can be applied to integrate large scaled seismic data. The workflow involves following steps: (1) model the secondary data distribution which is a distribution of block values and get the representative global facies proportions, (2) model joint distribution using kernel density estimator, (3) constraining the joint distribution with the established marginal conditions, and (4) derive the conditional probabilities from the constrained joint distribution. Figures-11 and 12 show a synthetic reference image consists of three facies types, randomly selected wells from reference image, and simulated secondary variables. The example is a vertical section of the reservoir with size of 100m in lateral and 50m in vertical direction. Modeling cell size is 1m-by-1m in horizontal and vertical direction. Vertical resolution of secondary data is assumed to be 5 times larger than the modeling size and so secondary values are generated at every 1m in horizontal and 5m in vertical. Figure-12 shows the comparison of the facies proportions calculated from block average of the reference image and the secondary derived proportions. Visual inspection gives that they show good agreement.

Geologic analog data may be the first available data for reservoir study. Geologists typically interpret geology and support their interpretations using the analog data. Geologic analog or interpreted information is often treated deterministically. Geostatistically constructed reservoir models are then validated in light of geologic sense. Otherwise, geologic data is sometimes used for better inferring the geostatistical modeling parameters such as horizontal range, ratio of horizontal to vertical range in 3-D variogram model and areal trend (Deutsch and Kupfersberger, 1998).

Incorporating Geologic Map with Soft Secondary Data

Joint pdf modeling technique under marginal constraints can be extended for incorporating geologic data into soft secondary data. Geologic data involves an interpreted geologic map or database of

geologic patterns which are type of exhaustively sampled images such as soft secondary data. The joint relation between geologic data and soft secondary data was modeled with nonparametric method and the known marginal conditions were applied to modify the joint distribution. By direct modeling the joint distribution, the geologic information is fused into the final probabilistic model without external data redundancy calibration that the probability combination requires.

Suppose the true reference image is available as shown in the figure-13. 61 wells are randomly extracted from the reference image and treated as well data. Channel and non-channel are coded as 1 and 0, respectively.

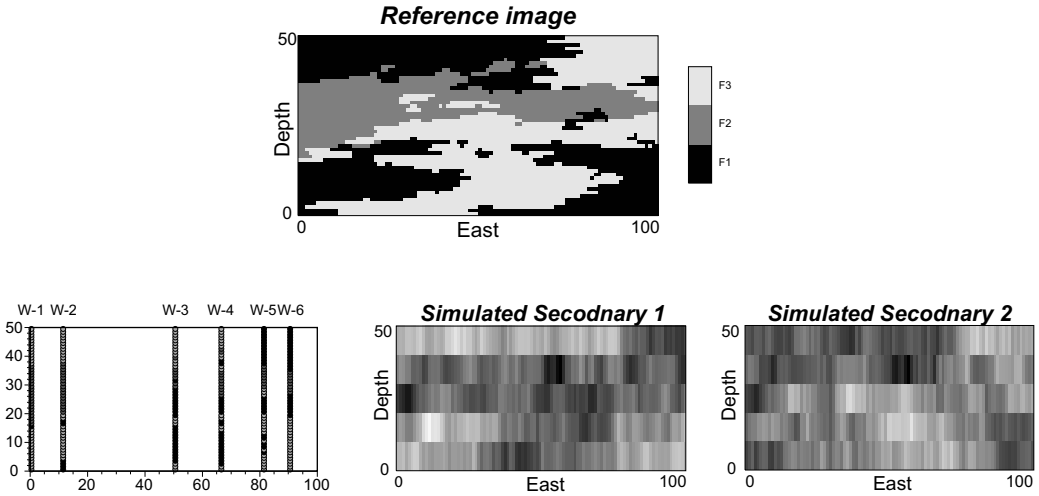

Figure-11: A synthetic reference image, well data extracted from the reference image and large scaled soft secondary data.

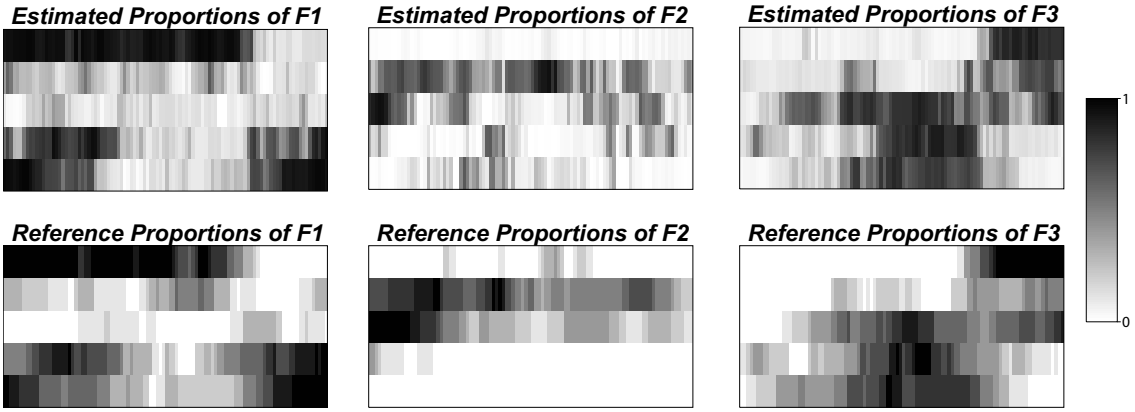

Figure-12: Facies proportions from the true image and estimated proportions

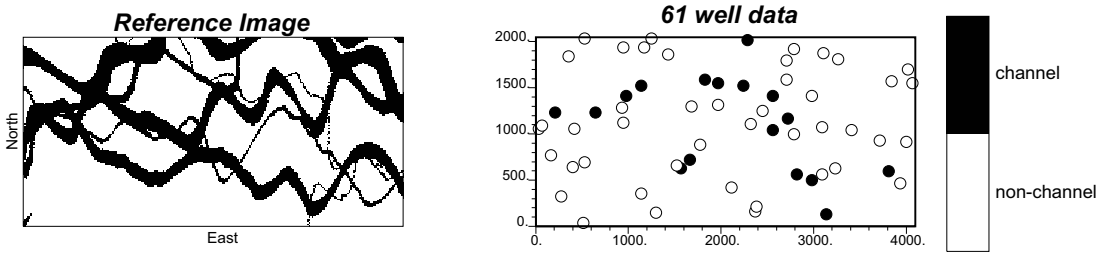

Figure-13: Reference image showing channel reservoir and 61 well data extracted from the reference image.

Two seismic variables (Y_1 and Y_2) are generated and are non-linearly correlated with the linear correlation of 0.572 (see figure-14). The synthetic soft secondary variables are made to differentiate somehow the channel and non-channel facies. Improvements in channel identification are expected by integrating those secondary data.

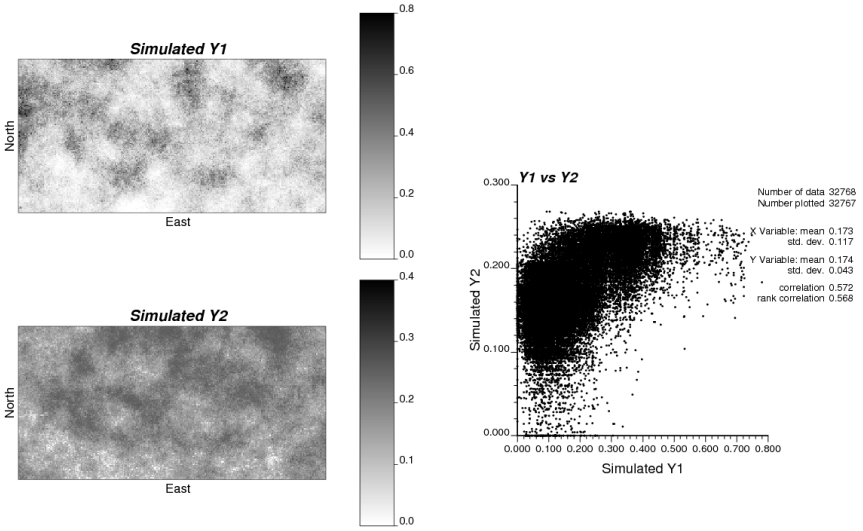

Figure-14: Two soft secondary variables are generated using the sequential Gaussian simulation and are correlated in a non-linear manner as shown in the cross plot of two variables.

One possible geologic map is prepared in the figure-15 showing a certain pattern of complex geologic features such as curvilinear channel. The image is not conditioned to well information; it is simply mapped to represent the prior geological/structural concept in terms of expected orientation and curvature of channels. Geologic map data typically need not carry any locally accurate information on the reservoir; the main purpose of using geologic map is to reflect a prior knowledge about complex geology (Caers and Zhang, 2004). The prior geologic map data has the same grid definition with the final modeling in X and Y direction.

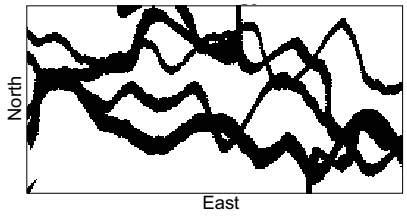

Figure-15: A geologic map used for integrating with soft secondary data.

Given the soft secondary and geologic map data, the joint distribution is modeled using kernel density estimator. Marginal conditions are established and are applied to constrain the initial distribution. The constrained joint distributions are shown in the figure-16. Primary, secondary and geologic data are denoted as random variables S,**Y**=[Y_1,Y_2],G, respectively in the figure. Joint distribution is modeled in a trivariate space, but it is clipped and shown based on the outcomes of primary and secondary geology variables.

Channel probability is derived from the corrected joint distribution. Figure-17 illustrates the map of channel probability compared with the reference image. To see the effect of geology information, bottom of figure-17 shows the probability of channel estimated from integrating soft secondary and geology data (left), and from integrating soft secondary data only (right). Probabilities are shown if pixels have 65% or higher chances of channel. Geologic heterogeneity is better captured and complex geologic patterns are accounted for by considering geology information.

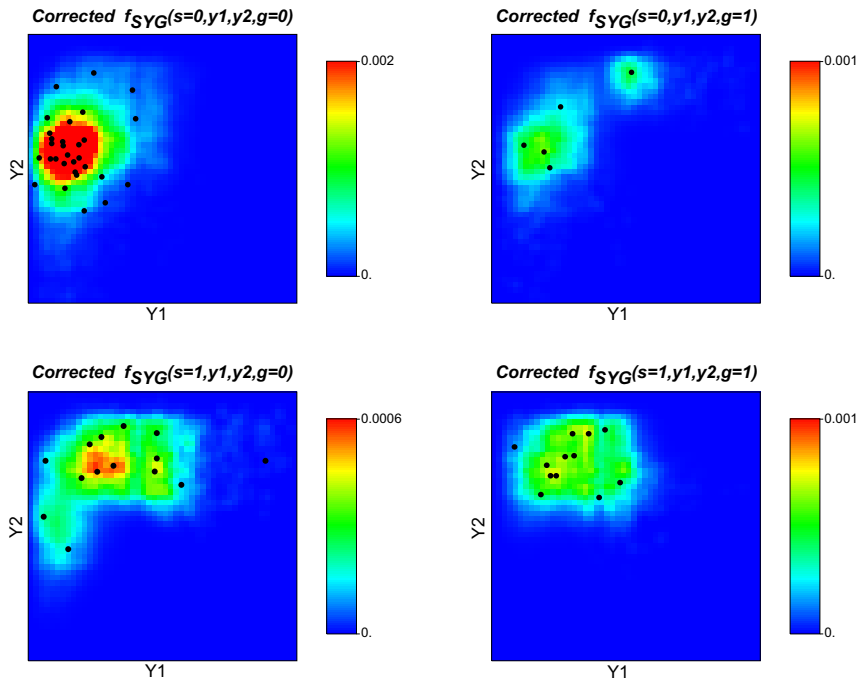

Figure-16: The corrected joint distributions under the imposed marginal constraints.

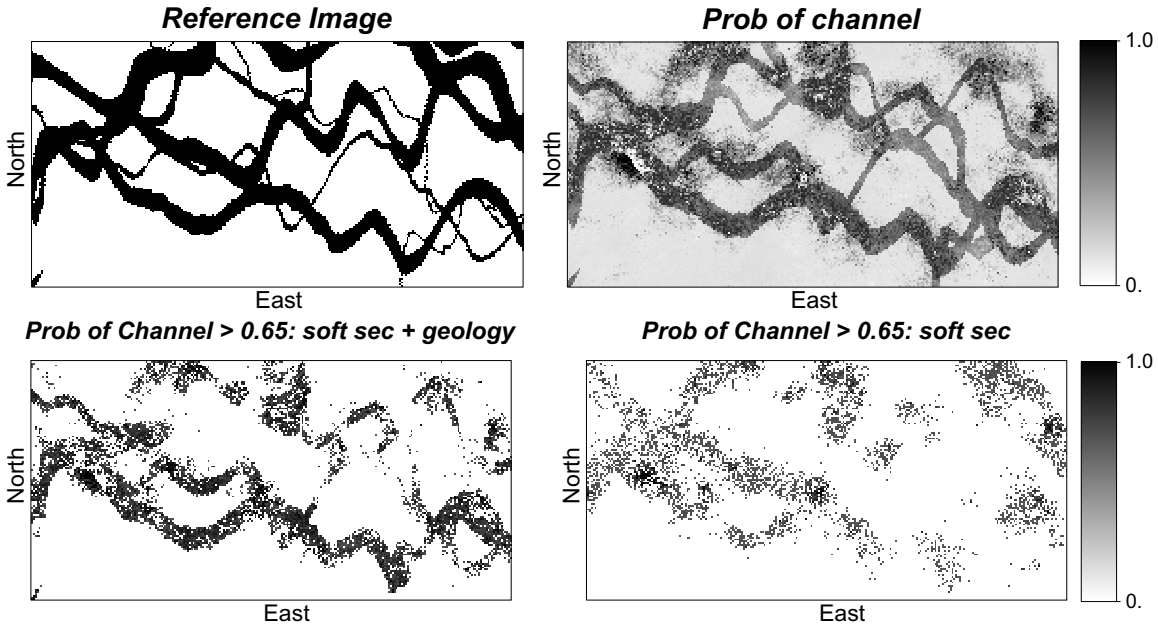

Figure-17: The estimated probability of channel is shown at the top. Result is compared with reference image. Bottom figures compared two different results from integrating soft secondary and geology, and from integrating soft secondary data only.

Applications to the Continuous Variable Modeling

The described workflow, modeling the joint distribution and constraining it under marginal conditions, can be applied to the continuous variable modeling. Joint pdf among primary and all of secondary data are modeled using kernel estimator. The proposed marginal fitting algorithm imposes marginal conditions on initial joint pdf resulting in an updated joint pdf. Once the joint pdf is obtained, estimates of primary variable can be immediately derived from the joint pdf given secondary values at $\mathbf{u}$, $\mathbf{u} \in \boldsymbol{A}$.

Because the approach accounts for non-linear relation among primary and secondary data, the derived estimates and estimation variances are not just limited to linear correlation coefficient. An example of applying this approach to Bayesian updating technique is demonstrated in Hong and Deutsch (2009).

Discussion

Incorporating secondary data is a longstanding problem in petroleum geostatistics. The main difficulties arise from the various levels of scales, coverage and precision. Geostatistical reservoir modeling involves two steps in the presence of multiple secondary data. First, secondary data are integrated to generate an estimate in a primary variable unit, e.g. facies probability or estimates of porosity. Given several secondary data, the conditional probability or probability distribution of primary variable are modeled. In the second step, results from the first step are integrated with well data through spatial modeling.

There are many approaches for the first secondary data integration step. In particular, probability combination schemes and direct joint pdf modeling are reviewed. The PCS is an indirect estimation method of the conditional probability of interest through combining individual probabilities that are obtained from each secondary data. Despite of potential applicability of this approach such as 3D trend modeling or information integration, it is hard to apply to the practical applications because there are no generalized ways to estimate data redundancy parameters which is a crucial part in the PCS. The proposed direct joint pdf modeling approach would be more preferable in terms of clarity and robustness. The method directly models the joint relations among variables in a nonparametric way and data redundancy is implicitly accounted for during joint distribution modeling. An external redundancy calibration process that is often ad-hoc and case-dependent is not required. Marginality conditions that the modeled probability distribution must meet no matter what methods are used are directly evaluated in a new method.

References

O. Babak and C. V. Deutsch, 2007, Merging multiple secondary data for collocated cokriging, *Centre for Computational Geostatistics*.

J. A. Benediktsson and P. H. Swain, 1992, Consensus Theoretic Classification Methods, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 22, No. 4.

J. Caers and T. Zhang, 2004, Multiple-point geostatistics: A quantitative vehicle for integrating geologic analogs into multiple reservoir models, in *Integration of outcrop and modern analogs in reservoir modeling: AAPG Memoir 80*, p. 383-394.

J. Caers, T. Hoffman, S. Strebelle, and X-H, Wen, 2006, Probabilistic integration of geologic scenarios, seismic, and production data-a West Africa turbidiate reservoir case study, *The Leading Edge*, 25, 240.

S. Castro, J. Caers, C. Otterlei, T. Hoye, T. Andersen, and P. Gomel, 2006, A probabilistic integration of well log, geological information, 3D/4D seismic and production data: application to the Oseberg field, SPE 103152.

T. L. Chugunova and L. Y. Hu, 2008, Multiple-point simulations constrained by continuous auxiliary data, *Mathematical Geology*, Vol. 40.

C. V. Deutsch, 2002, *Geostatistical reservoir modeling*, Oxford University Press, New York.

C. V. Deutsch, S. Srinivasan and Y. Mo, 1996, Geostatistical reservoir modeling accounting for precision and scale of seismic data. SPE 36497.

S. Hong and C. V. Deutsch, 2007, Methods for Integrating Conditional Probabilities for Geostatistical Modeling, *Centre for Computational Geostatistics*.

S. Hong and C. V. Deutsch, 2009, 3D trend modeling by combining lower order trends, *Centre for Computational Geo*statistics.

S. Hong and C. V. Deutsch, 2009, Semi-nonparametric Bayesian Updating, *Centre for Computational Geostatistics.*

J. M. Harris and R. T. Langan, 1997, Crosswell seismic profiling: principle to applications, in Geophysical Corner, *AAPG Explorer*.

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

S. Krishnana, 2004, Combining diverse and partially redundant information in the earth sciences, PhD dissertation, Stanford University.

H. Kupfersberger, C. V. Deutsch, A. G. Journel, 1998, Deriving constraints on small-scale variograms due to variograms of large-scale data, *Mathematical Geology*, Vol. 30, No. 7.

T. Lee, J. A. Richards, and P. H. Swain, 1987, Probabilistic end evidential approaches for multisource data analysis, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. GE-25, No. 3.

K. J. McConway, 1981, Marginalization and Linear Opinion Pools, *Journal of American Statistical Association*, Vol. 76, No. 374.

E. I. Polyakova and A. G. Journel, 2007, The nu expression for probabilistic data integration, *Mathematical Geology*, Vol. 39.

B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, 1986.

D. W. Scott, 1992, *Multivariate Density Estimation*, John Wiley and Sons, Inc., New York.

R. L. Winkler, 1981, Combining probability distributions from dependent information sources, *Management Science*, Vol. 27, No.4.