Accounting for Nonlinear Relations in Bayesian Updating

Sahyun Hong and Clayton V. Deutsch

Bayesian Updating (BU in short) is a widely used technique to combine primary and multiple secondary variables. In this work, advanced BU technique is considered to account for non-linear relation among primary and secondary variables. The work consists of two parts. In the first part, new expression of BU equation is suggested and some advantages are addressed over the conventional expression that is not clearly understood how different conditional pdf from disparate sources are combined. In the second part, an approach is suggested to account for the non-linear relation among variables using the new form of updating equation. Joint pdf between primary and secondary variables is modeled in a nonparametric way and constraints the joint pdf must meet are described in the latter part. Different from the conventional BU, the considered approach gives the locally varying estimation variance that better reflects the relation between local primary and secondary data.

Introduction

Bayesian updating technique was first proposed in 1996 by Doyen and since then the technique has been widely adopted to integrate multiple soft secondary variables. The BU is fundamentally equivalent to the collocated cokriging using the Markov type screening assumption developed by Xu et al. but its formalism is different from the collocated cokriging (Doyen et al., 1996; Chiles and Delfiner, 1999; Xu et al., 1992). Collocated cokriging is an extended form of kriging expressed by a weighted linear combination of nearby primary and collocated secondary data, and Bayesian updating expresses the updated distribution by the combination of probability distributions conditioned to each primary and secondary variable.

This work introduces the detailed derivation of Bayesian updating equation and proposes an alternative form to the conventional BU equation. New expression gives an explicit interpretation such that how probability distributions derived from disparate data sources can be combined leading to the final updated distribution, and where the global mean and variance play in the updating equation.

Because new updating equation decomposes the influence of the primary and secondary variable, separate calibration of secondary variable from primary variable is possible. An approach to account for non-linear relations among primary and secondary variables is proposed using the new form of updating equation. The joint probability density functions are modeled in a nonparametric way using kernel density estimator. Because the (initial) joint pdf does not meet the marginality condition fitting algorithm is described to refine the modeled joint pdf into the corrected joint pdf that satisfies all lower order marginal conditions. The considered marginal fitting algorithm directly accounts for the differences between the empirical and reference marginal pdf. Conditional pdf at a given secondary value can be drawn from the corrected joint pdf and thus mean and variance are numerically estimated based on the extracted conditional pdf. The secondary data derived moments are now easily anchored into the new form of updating equation.

Resolution of Bayesian Updating

The primary and secondary variables are denoted as random variables Z and Y, and they are already transformed into normal score values with mean of 0 and variance of 1. A posterior distribution of interest is the conditional distribution of Z given the surrounding primary and collocated secondary data:

$$f(z(\mathbf{u}) \mid z(\mathbf{u}_1),\ldots,z(\mathbf{u}_n),y(\mathbf{u})),\quad \mathbf{u}\in A \tag{1}$$

where $z(\mathbf{u}_1)$,…, $z(\mathbf{u}_n)$ are nearby primary data at different locations $\mathbf{u}_i$, i=1,…,n, and $y(\mathbf{u})$ is a collocated secondary data retained as conditioning data, respectively. A single secondary variable is considered as an example, but any equations derived in this work can be simply extended into multiple secondary variables using vector and matrix notation. The equation (1) is re-expressed as:

$$\begin{aligned} f\left(z(\mathbf{u})\mid z(\mathbf{u}_1),\dots,z(\mathbf{u}_n),y(\mathbf{u})\right) &= \frac{f\left(z(\mathbf{u}),z(\mathbf{u}_1),\dots,z(\mathbf{u}_n),y(\mathbf{u})\right)}{f\left(z(\mathbf{u}_1),\dots,z(\mathbf{u}_n),y(\mathbf{u})\right)} \\ &= \frac{f\left(z(\mathbf{u}_1),\dots,z(\mathbf{u}_n),y(\mathbf{u})\mid z(\mathbf{u})\right)f\left(z(\mathbf{u})\right)}{f\left(z(\mathbf{u}_1),\dots,z(\mathbf{u}_n),y(\mathbf{u})\right)} \end{aligned} \tag{2}$$

The conditional distribution $f(z(\mathbf{u}_1),\dots,\ z(\mathbf{u}_n),y(\mathbf{u})|z(\mathbf{u}))$ in the numerator is approximated as $f(z(\mathbf{u}_1),\dots,z(\mathbf{u}_n),y(\mathbf{u})|z(\mathbf{u})) \cong f(z(\mathbf{u}_1),\dots,z(\mathbf{u}_n)|z(\mathbf{u}))\times f(y(\mathbf{u})|z(\mathbf{u}))$ under independence assumption between collocated $y(\mathbf{u})$ and local surrounding primary data $[z(\mathbf{u}_1),\dots,z(\mathbf{u}_n)]$ conditioned to estimate of primary variable Z at **u**, $z(\mathbf{u})$. This independence assumption alleviates the requirement of inferring the joint distribution $f(z(\mathbf{u}_1),\dots,z(\mathbf{u}_n),y(\mathbf{u})|z(\mathbf{u}))$ which is difficult to model because it requires joint modeling of mixed variables from different locations. The equation (2) is approximated as:

$$f\left(z(\mathbf{u})\mid z(\mathbf{u}_1),\dots,z(\mathbf{u}_n),y(\mathbf{u})\right)=\frac{f\left(z(\mathbf{u}_1),\dots,z(\mathbf{u}_n)\mid z(\mathbf{u})\right)f\left(y(\mathbf{u})\mid z(\mathbf{u})\right)f\left(z(\mathbf{u})\right)}{f\left(z(\mathbf{u}_1),\dots,z(\mathbf{u}_n),y(\mathbf{u})\right)} \tag{3}$$

Conditional independence assumption decouples the posterior distribution into two terms: (1) distribution associated with the primary data at different locations, $f(z(\mathbf{u}_1),\dots,z(\mathbf{u}_n)|z(\mathbf{u}))$, and (2) the distribution associated with the primary and secondary variable relation, $f(y(\mathbf{u})|z(\mathbf{u}))$. Probabilistic terms in the right hand side of equation (3) are *likelihood function* that treats the unknown estimate $z(\mathbf{u})$ as fixed. They are re-expressed as probability functions of the unknown estimate given the fixed data:

$$f\left(z(\mathbf{u})\mid z(\mathbf{u}_1),\dots,z(\mathbf{u}_n),y(\mathbf{u})\right)=\frac{f\left(z(\mathbf{u})\mid z(\mathbf{u}_1),\dots,z(\mathbf{u}_n)\right)}{f\left(z(\mathbf{u})\right)}\frac{f\left(z(\mathbf{u})\mid y(\mathbf{u})\right)}{f\left(z(\mathbf{u})\right)}f\left(z(\mathbf{u})\right)\cdot C \tag{4}$$

where normalizing C term is $f(z(\mathbf{u}_1),\dots,z(\mathbf{u}_n))\times f(y(\mathbf{u}))/(f(z(\mathbf{u}_1),\dots,z(\mathbf{u}_n),y(\mathbf{u}))$. Although Bayesian updating technique has been commonly used in many cases, there are few references explaining how the final updating equations are reached (Neufeld and Deutsch, 2004; Deutsch and Zanon, 2004; Ren et al., 2006). Equation (4) provides a posterior distribution through the multiplication of three probability distributions. $f(z(\mathbf{u})|z(\mathbf{u}_1),\dots,z(\mathbf{u}_n))$ is a univariate conditional distribution of Z conditioned to local nearby data $z(\mathbf{u}_1),\dots,z(\mathbf{u}_n)$. Under multiGaussianity (MG) assumption, kriging parametrically constructs the Gaussian distribution with mean of kriging estimate and variance of kriging varianc (Journel and Huijbregts, 1981; Verly, 1983):

$$f(z(\mathbf{u})\mid z(\mathbf{u}_1),\dots,z(\mathbf{u}_n))=\frac{1}{\sqrt{2\pi\sigma_P^2(\mathbf{u})}}\exp\left\{-\frac{(z(\mathbf{u})-z_P(\mathbf{u}))^2}{2\sigma_P^2(\mathbf{u})}\right\} \tag{5}$$

where $z_P(\mathbf{u})$ and $\sigma^2_P(\mathbf{u})$ are estimate and estimation variance obtained by simple kriging at **u**. Subscript *P* means that those are primary data derived moments. The $f(z(\mathbf{u})|y(\mathbf{u}))$ in equation (4) is:

$$f(z(\mathbf{u})\mid y(\mathbf{u}))=\frac{1}{\sqrt{2\pi\sigma_S^2}}\exp\left\{-\frac{(z(\mathbf{u})-z_S(\mathbf{u}))^2}{2\sigma_S^2}\right\} \tag{6}$$

$z_S(\mathbf{u})$ and $\sigma^2_S(\mathbf{u})$ are estimate and variance obtained by overall relation between primary and secondary variables. Subscript *S* are added since they are secondary data derived moments. Due to linear relation assumption between Z and Y, the conditional mean and variance $z_S(\mathbf{u})$ and $\sigma^2_S(\mathbf{u})$ are simply calculated as $z_S(\mathbf{u}) = \rho \times y(\mathbf{u})$ and $\sigma^2_S(\mathbf{u})=1-\rho^2$, where ρ is a linear correlation coefficient between Z and Y. $z_S(\mathbf{u})$ depends on the local secondary value $y(\mathbf{u})$ but variance $\sigma^2_S(\mathbf{u})$ is constant over $\mathbf{u}\in A$. The last term function of $z(\mathbf{u})$ in equation (4) is a global distribution of the primary variable $f(z(\mathbf{u}))$:

$$f(z(\mathbf{u}))=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(z(\mathbf{u})-m)^2}{2\sigma^2}\right\} \tag{7}$$

Even though the primary variable *Z* has zero mean and unit variance (σ^2=1, *m*=0), *m* and σ^2 symbols are left in the above equation. One can follow how the global mean and variance play in the final updating equation by letting symbols remain.

Elementary probability distributions consisting of a posterior distribution are all Gaussian (equations (5), (6) and (7)). Multiplication of Gaussian distributions is another Gaussian; thus, the posterior distribution becomes Gaussian. Equations shown in (5), (6) and (7) are inserted into the (4) as following:

$$
\begin{aligned}
f\left(z(\mathbf{u})\,|\,z(\mathbf{u}_1),\ldots,z(\mathbf{u}_n),y(\mathbf{u})\right) &= \frac{f\left(z(\mathbf{u})\,|\,z(\mathbf{u}_1),\ldots,z(\mathbf{u}_n)\right)}{f\left(z(\mathbf{u})\right)}\frac{f\left(z(\mathbf{u})\,|\,y(\mathbf{u})\right)}{f\left(z(\mathbf{u})\right)}f\left(z(\mathbf{u})\right)\cdot C \\
&= \frac{\dfrac{1}{\sqrt{2\pi\sigma_P^2(\mathbf{u})}}\exp\left\{-\dfrac{(z(\mathbf{u})-z_P(\mathbf{u}))^2}{2\sigma_P^2(\mathbf{u})}\right\}}{\dfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\dfrac{(z(\mathbf{u})-m)^2}{2\sigma^2}\right\}}\frac{1}{\sqrt{2\pi\sigma_S^2(\mathbf{u})}}\exp\left\{-\frac{(z(\mathbf{u})-z_S(\mathbf{u}))^2}{2\sigma_S^2(\mathbf{u})}\right\}\cdot C
\end{aligned}
\tag{8}
$$

Terms inside exponential function are grouped and terms independent of *z*(**u**) are absorbed in the proportionality then:

$$
\begin{aligned}
&f(z(\mathbf{u})\,|\,z(\mathbf{u}_1),\ldots,z(\mathbf{u}_n),y(\mathbf{u})) \propto \\
&\qquad \exp\left\{-\frac{\left(z(\mathbf{u})-z_P(\mathbf{u})\right)^2}{2\sigma_P^2(\mathbf{u})}+\frac{\left(z(\mathbf{u})-m\right)^2}{2\sigma^2}-\frac{\left(z(\mathbf{u})-z_S(\mathbf{u})\right)^2}{2\sigma_S^2(\mathbf{u})}\right\}
\end{aligned}
\tag{9}
$$

Equation (9) is arranged with respect to *z*(**u**):

$$
\begin{aligned}
&f(z(\mathbf{u})\,|\,z(\mathbf{u}_1),\ldots,z(\mathbf{u}_n),y(\mathbf{u})) \propto \\
&\qquad \exp\left\{-\underbrace{\left[\frac{1}{2\sigma_P^2(\mathbf{u})}-\frac{1}{2\sigma^2}+\frac{1}{2\sigma_S^2(\mathbf{u})}\right]}_{A}z^2(\mathbf{u})+\underbrace{\left[\frac{z_P(\mathbf{u})}{\sigma_P^2(\mathbf{u})}+\frac{z_S(\mathbf{u})}{\sigma_S^2(\mathbf{u})}-\frac{m}{\sigma^2}\right]}_{B}z(\mathbf{u})\right\}
\end{aligned}
\tag{10}
$$

Terms independent of *z*(**u**) were absorbed in the proportionality again. Equation (10) follows a quadratic form of exp{-Az^2 + B*z*} where A and B are parameterized coefficients. This can be easily converted into the basic form of Gaussian function:

$$
\begin{aligned}
&f(z(\mathbf{u})\,|\,z(\mathbf{u}_1),\ldots,z(\mathbf{u}_n),y(\mathbf{u})) \propto \\
&\qquad \exp\left\{-Az^2(\mathbf{u})+Bz(\mathbf{u})\right\} \propto \exp\left\{-\frac{\left(z(\mathbf{u})-B/2A\right)^2}{2\left(1/2A\right)}\right\}
\end{aligned}
\tag{11}
$$

A posterior distribution *f*(*z*(**u**)|*z*(**u**$_1$),…,*z*(**u**$_n$),*y*(**u**)) becomes a Gaussian distribution with the mean of B/2A and variance of 1/2A while A and B are defined in the equation (10):

$$
\sigma_{BU}^2(\mathbf{u})=\frac{1}{2A} \quad \text{and} \quad z_{BU}(\mathbf{u})=\frac{B}{2A}
\tag{12}
$$

The Bayesian updated variance and estimate at location **u** are finally:

$$
\begin{aligned}
\frac{1}{\sigma_{BU}^2(\mathbf{u})} &= \frac{1}{\sigma_P^2(\mathbf{u})}+\frac{1}{\sigma_S^2}-\frac{1}{\sigma^2} \\
\frac{z_{BU}(\mathbf{u})}{\sigma_{BU}^2(\mathbf{u})} &= \frac{z_P(\mathbf{u})}{\sigma_P^2(\mathbf{u})}+\frac{z_S(\mathbf{u})}{\sigma_S^2}-\frac{m}{\sigma^2}
\end{aligned}
\tag{13}
$$

A set of equation (13) is a new form of Bayesian updating equation compared with the conventional form below:

$$
\begin{aligned}
z_{BU}(\mathbf{u}) &= \frac{z_P(\mathbf{u})\sigma_S^2+z_S(\mathbf{u})\sigma_P^2(\mathbf{u})}{\sigma_P^2(\mathbf{u})-\sigma_P^2(\mathbf{u})\sigma_S^2(\mathbf{u})+\sigma_S^2(\mathbf{u})} \\
\sigma_{BU}^2(\mathbf{u}) &= \frac{\sigma_P^2(\mathbf{u})\sigma_S^2(\mathbf{u})}{\sigma_P^2(\mathbf{u})-\sigma_P^2(\mathbf{u})\sigma_S^2(\mathbf{u})+\sigma_S^2(\mathbf{u})}
\end{aligned}
\tag{14}
$$

Equations (13) and (14) are exactly same under letting *m* = 0 and σ^2 = 1 in the new updating equation (13). The main advantage of new expression is that the updated parameters z_{BU} and σ^2_{BU} are clearly decomposed into combination of local and global parameters derived from diverse sources. This form is better than the original equation in terms of simplicity, lucidity and potential applications such as accounting for non-stationary global mean and variance or non-linear relations among primary and secondary variables. The application of the new form to the nonstationary modeling was discussed in Sahyun and Deutsch (2008).

Accounting for Non-linear Relations between Primary and Secondary Variables

Several assumptions have been made to derive the Bayesian updating equation and key assumptions are: (1)multiGaussianity of primary variables at different locations, which allows constructing the local conditional distribution $f(z(\mathbf{u})|z(\mathbf{u}_1),\ldots,z(\mathbf{u}_n))$ by simple kriging at **u**, (2)linear relation between primary and secondary variables, which allows building the conditional distribution $f(z(\mathbf{u})|y(\mathbf{u}))$ by linear correlation coefficient ρ. Only under Gaussian assumption on distributions (multivariate Gaussian in spatial context and multivariate Gaussian in variable context), analytical derivation of updating equation (equations (5) through (13)) is possible. Relaxing the linear relation assumption between primary and secondary variables is focused here. MultiGaussianity assumption is still adopted to build the local conditional distribution $f(z(\mathbf{u})|z(\mathbf{u}_1),\ldots,z(\mathbf{u}_n))$.

Conventional Bayesian updating assumes the multivariate Gaussian relation after univariate normal score transformation for each variable. An illustration shown in the figure-1 represents a highly non-linear feature between two normal scored variables (normal scored values of porosity and permeability from 3D Amoco data set). Normality of univariate marginal distribution is a necessary condition, but not a sufficient condition for joint Gaussianity.

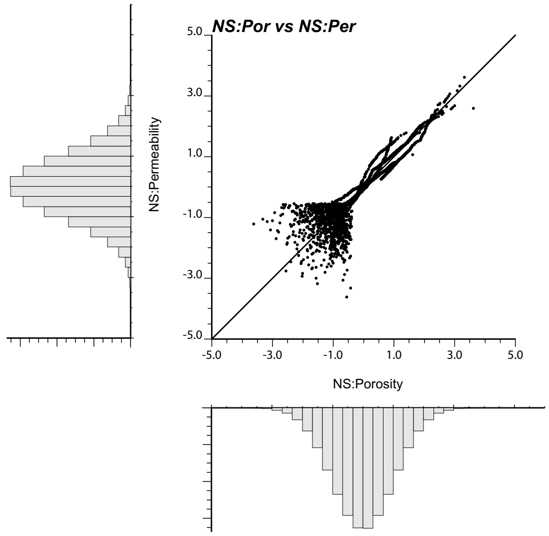


Figure-1: An illustration of non-linear relation after individual normal score transform of each variable.

The Proposed Methodology

To capture the non-linear relation among primary and secondary variables, the joint relation is first modeled in a nonparametric way. Kernel density estimator is considered to model joint pdf $f(z,y)$ without data distribution assumption; the method applies the specified kernel function to the pairs of collocated primary and secondary data and then approximate the underlying true distribution through adding up the applied kernel function values. Figure-2 schematically illustrates the data-driven pdf.

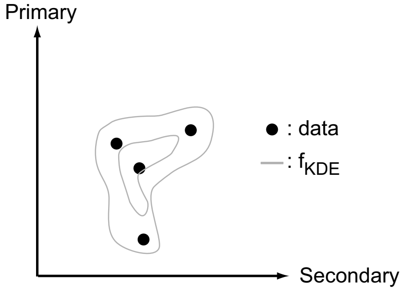


Figure-2: A schematic illustration of the bivariate pdf modeled in a nonparametric way

The kernel estimate for a single variable is defined by (Scott, 1992):

$$f_{KDE}(z)=\frac{1}{nh}\sum_{i=1}^{n}W\left(\frac{z-Z_i}{h}\right)\ \text{for}\ 1-\text{D} \tag{15}$$

where *h* is a bandwidth of the applied kernel function, also referred to a smoothing parameter because *h* controls the smoothness of the resulting density function. *W*(·) is a kernel function satisfying *W*(·)≥0 and ∫*W*(*x*)d*x* = 1 and these conditions make the resulting density estimates be positive and grand total of densities be 1. *n* is the number of used samples for pdf modeling. Bivariate pdf is constructed by product kernel density estimator (Scott, 1992):

$$f(z,y)=\frac{1}{nh_1\times h_2}\sum_{i=1}^{n}\left(W\left(\frac{(z-Z_i)}{h_1}\right)\times W\left(\frac{(y-Y_i)}{h_2}\right)\right) \tag{16}$$

where *W*(·) is a univariate kernel applied to each variable Z and Y. Equation (16) is referred to a product kernel estimator for joint pdf modeling (Scott, 1993). h_1 and h_2 are kernel bandwidths for variable Z and Y. Scott (1993) recommended a data-based optimum bandwidth:

$$h_i=\hat{\sigma}_i n^{-1/(d+4)} \tag{17}$$

where $\hat{\sigma}$ is a standard deviation of samples and *d* is the number of variables. Next step is aimed at checking axioms of the modeled joint pdf: 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*(·) follows *W*(*x*)≥0 and ∫*W*(*x*)d*x*=1. The third condition is a marginality condition that the *p*-variate joint pdf should reproduce *p*´-variate pdf where *p*´< *p*. The followings are marginal conditions that the modeled bivariate pdf must meet:

$$\int f(z,y)dy=f(z) \tag{18}$$

$$\int f(z,y)dz=f(y) \tag{19}$$

The marginal relation described in (18) states that integration of the joint probability distribution over possible outcomes of secondary data should amount to the global probability distribution of the primary variable Z, *f*(z). The second marginal relation (19) states that integration of the joint probability distribution over possible outcomes of primary variable should reproduce the distribution of secondary variable Y, *f*(y).

The global distribution of the primary variable *f*(z) is experimentally obtained from well samples. Declustering techniques such as cell or polygonal declustering is used to obtain the representative *f*(*z*) if there is a bias in data distribution caused by spatial clustering of wells. The secondary data distribution *f*(*y*) is modeled with the densely sampled values over the area. The modeling of *f*(y) is very reliable.

There is no guarantee, however, that the modeled joint pdf *f*(*z*,*y*) meet these marginal conditions. *f*(*z*,*y*) is modeled based on the limited samples (*n*) that is much less than the number of secondary values that constitute the marginal distribution *f*(*y*). The collocated secondary data at the sample locations $\mathbf{u}_i$, *i*=1,…,*n* normally do not cover the full range of the secondary data; therefore, the marginal distribution from the joint distribution may not match the secondary marginal distribution. Figure-3 illustrates this case. The bivariate distribution f_{KDE} 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 $f_{\text{reference}}$ and it is built from large number of samples. $f_{\text{reference}}$ have variations in densities through the entire range of secondary values.

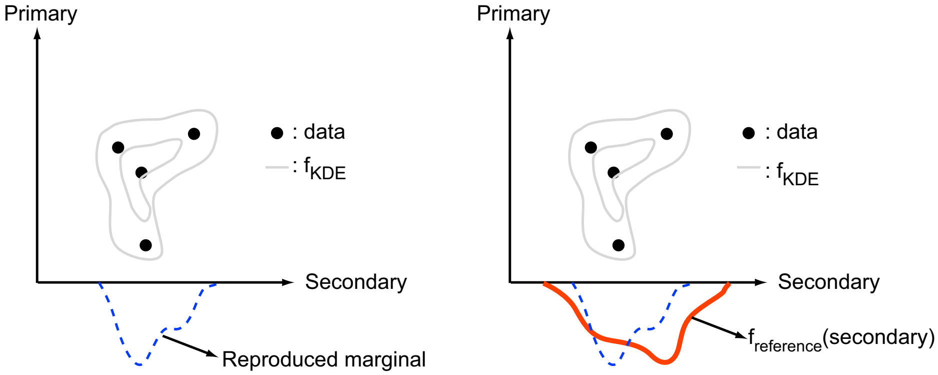


Figure-3: Schematic illustration for comparing the reproduced with the known marginal distribution. Since the joint distribution is modeled with the limited well samples its reproduced marginal is not consistent with the (very reliable) marginal pdf which is a distribution of secondary data.

Given the marginal relations described in equations (18) and (19), an algorithm is proposed to impose them on the joint pdf. 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 joint distributions. This correcting process is performed as described in the following steps:

Step1. Model the distribution of secondary variable, $f(y)$ and global distribution of primary variable, $f(z)$. Declustering is considered for obtaining an unbiased $f(z)$ if required.

Step2. Model the joint distribution $f(z,y)$ 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 (19). The scaling equation below is proposed for ensuring the imposed marginal condition:

$$f^{(0)}(z,y)\times\frac{f(y)}{\int f^{(0)}(z,y)dz}\rightarrow f^{(1)}(z,y) \tag{20}$$

The ratio $f(y)/\int f(z,y)dz$ is a modifying factor. If the marginal relation (19) is satisfied then this factor becomes 1 leading to no changes in $f^{(0)}$. Otherwise, $f^{(0)}$ is adjusted by the modifying factor that accounts for the differences between the reference $f(y)$ and reproduced marginal distribution $\int f^{(0)}(z,y)dz$. The corrected distribution under the marginal condition is set as $f^{(1)}$ for the next step.

Step4. Scale the $f^{(1)}$ to ensure the marginal condition shown in equation (18). Similar to the step 3, the scaling equation below is for updating the $f^{(1)}$:

$$f^{(1)}(z,y)\times\frac{f(z)}{\int f^{(1)}(z,y)dy}\rightarrow f^{(2)}(z,y) \tag{21}$$

The ratio $f(z)/\iint f^{(1)}(z,y)dy$ is another modifying factor. If the marginal relation (18) is met then the factor becomes 1 leading to no change in $f^{(1)}$. Otherwise, $f^{(1)}$ is adjusted by the modifying factor accounting for the differences between the reference marginal distribution $f(z)$ and the reproduced marginal distribution $\iint f^{(1)}(z,y)dy$. The corrected distribution under marginal condition (18) is set as $f^{(2)}$.

Step5. Terminate 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 $f(z)$ and $f(y)$. 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 averaged differences of density functions by more than the pre-defined threshold δ:

$$avergaed\ diff = | f^{(t)}(z,y) - f^{(t-1)}(z,y) | \frac{1}{C} < \delta$$

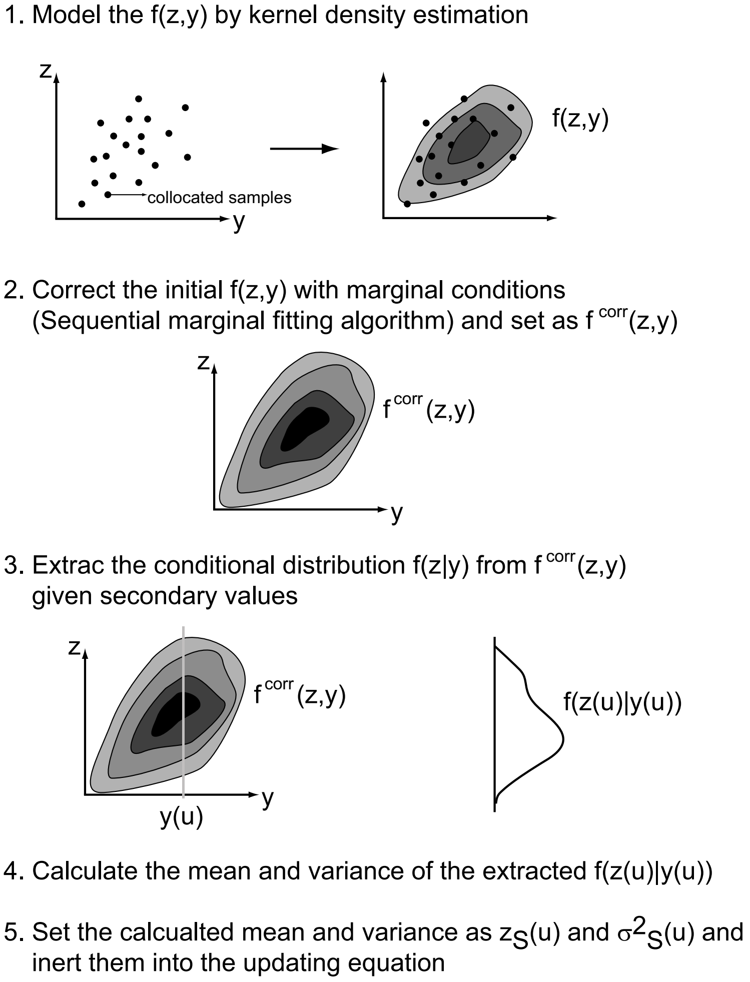
where *t* and (*t-1*) are the current and previous correction steps, respectively. Another stopping condition might be test an error between the reproduced and reference marginal distributions. The algorithm would stop iterations when the error becomes less than a specified tolerance:

$$averaged\ e_1 = | f^{repro}(y) - f^{ref}(y) | \frac{1}{C},\ averaged\ e_2 = | f^{repro}(z) - f^{ref}(z) | \frac{1}{C}$$

and

$$Total\ Averaged\ error = (e_1 + e_2)/2$$

The proposed marginal fitting algorithm corrects the initial joint pdf under marginal conditions. Once the joint pdf is achieved, conditional pdf can be immediately derived by Bayes law. The conditional distribution given secondary values, *f*(*z*(**u**)|*y*(**u**)) at **u**, is extracted from the corrected joint pdf *f*(*z*,*y*). Conditional mean and variance are calculated using the extracted conditional distribution, which are finally secondary data derived local estimate and variance, z_S(**u**) and σ^2_S(**u**). The obtained mean and variances are put into the Bayesian updating equation (13). Following charts (figure-4) gives whole idea of the proposed approach.



1. Model the f(z,y) by kernel density estimation

2. Correct the initial f(z,y) with marginal conditions (Sequential marginal fitting algorithm) and set as f^{corr}(z,y)

3. Extrac the conditional distribution f(z|y) from f^{corr}(z,y) given secondary values

4. Calculate the mean and variance of the extracted f(z(u)|y(u))

5. Set the calcualted mean and variance as z_S(u) and $\sigma^2{}_S$(u) and inert them into the updating equation

Figure-4: Workflow for the proposed approach.

The suggested algorithm accounts for the non-linear features between primary and secondary variables and its modeled joint pdf exactly reproduces lower order marginal distributions. The conditional distribution given any value of secondary data can be drawn from the obtained joint distribution. There is no guarantee that the extracted conditional distribution is a univariate Gaussian; it can be any shape of distribution (see the extracted 1-D distribution at step 3 in the above flow chart). Univariate Gaussian fitting is made for the extracted conditional distribution because the conditional pdf $f(z(\mathbf{u})|y(\mathbf{u}))$ can be fused with another local conditional pdf $f(z(\mathbf{u})|z(\mathbf{u}_1),\ldots,z(\mathbf{u}_n))$ into a single updated Gaussian distribution only when those distributions are all Gaussian (see the equations through (5) and (11)).

Examples

Amoco data set is used for the application of a new Bayesian updating method. Permeability is considered as primary variable to be modeled and porosity is considered as a secondary variable. Sampled porosities at 62 wells are simulated to initialize the exhaustive secondary data. 65-by-65 grids are defined in X and Y directions. Figure-5 illustrates the simulated porosity field and permeability samples at well locations. Figure-6 shows a scatter plot of the normal scored permeability and porosity at collocation. Bivariate relation does not seem to be bivariate Gaussian despite of univaraite normality. Evaluating the bivariate normality can be performed through plotting the squared generalized distances from data pairs against the chi-square distances (Johnson and Wichern, 2002). An illustration shown in the right of the figure-6 is a plot for the calculated distances from normal scored data pairs and analytical chi-square distances (see the reference of Johnson and Wichern, 2002, through p182 – 187 for the details about chi-square distances). The plot should resemble the straight line through the origin if bivariate normality is guaranteed. Systematic curved pattern shown on the chi-square distance plot suggests lack of normality for the normal scored data pairs.

The joint relation among variables was modeled by kernel density estimator and the described marginal fitting algorithm was applied. Figure-7 shows the modeled $f(z,y)$. Kernel bandwidths are chosen using the analytical suggestion shown in the equation (17). Range of primary and secondary variables are discretized into 50 bins and joint densities are estimated at every 50×50 bins. Given the actual secondary value at location $\mathbf{u}$, $y(\mathbf{u})$, the closest binned y value is looked-up in order to extract the conditional distribution of Z at the conditioned $y(\mathbf{u})$. Too few bins may induce a discretization error. The chosen level of 50 bins did not induce discretization errors by plotting the actual secondary values versus the looked-up binned values. The bivariate pdf shown in the figure-7 was refined pdf under the marginal constraints using the described marginal fitting method. Averaged marginal errors quickly drop during first few iteration steps. 100 marginal fitting iterations make the bivariate pdf be stable and it took a few seconds on a 3 GHz personal computer.

Based on the bivariate distribution, we derived the conditional mean and variance of the primary permeability given the secondary porosity. In a nonparametric modeling, the conditional mean and variance are not determined by the coefficient of correlation but are calculated based on the extracted conditional distribution $f(z(\mathbf{u})|y(\mathbf{u}))$, $\mathbf{u}\in A$. Conditional mean is numerically calculated through the binned z_k, k=1,…,50:

$$E\{Z\,|\,Y=y(\mathbf{u})\}=\sum_{k=1}^{50} z_k f(z_k\,|\,y(\mathbf{u}))=\sum_{k=1}^{50} z_k \frac{f(z_k, y(\mathbf{u}))}{f(y(\mathbf{u}))} \tag{22}$$

Extracted 1-D pdf curve $f(z_k,y(\mathbf{u}))$ is a function of binned primary variable z_k with a fixed $y(\mathbf{u})$. The conditional pdf $f(z_k|y(\mathbf{u}))$ is acquired from dividing $f(z_k,y(\mathbf{u}))$ by $f(y(\mathbf{u}))$. Conditional variances are also calculated by:

$$Var\{Z\,|\,Y=y(\mathbf{u})\}=\sum_{k=1}^{50}\left[z_k-\overline{z}\right]^2 f(z_k\,|\,y(\mathbf{u}))=\sum_{k=1}^{50}\left[z_k-\overline{z}\right]^2 \frac{f(z_k, y(\mathbf{u}))}{f(y(\mathbf{u}))} \tag{23}$$

where $\overline{z}$ is a conditional mean of Z given the $y(\mathbf{u})$. Figure-8 demonstrates the conditional means and variances calculated based on the local conditional pdf. Black dots superimposed with the bivariate pdf map represent the calculated conditional means. Conditional variances are plotted in the right. Conditional means are not linear and conditional variances are not constant during the whole variety of conditioned secondary values. Moreover, it is somehow related to the variation of secondary values. For example, the change of conditional variance is sharp when the given secondary values (normal scored porosity) are high or very low. Given the intermediate range of secondary data, about from -2.1 to 1.8, the conditional variance tends to decrease. Figure-9 shows maps of estimates and estimation variance derived from the secondary data.

Conventional BU assumes linear relation between primary and secondary variable so that it provides a constant estimation variance over the domain. Figure-10 demonstrates the comparison of the conditional estimates and estimation variance from using Gaussian assumption and nonparametric approach. Linear correlation (0.642) is calculated from the data pairs of normal scored porosity and permeability as plotted by open circles in the figure. Black solid line is an estimate given the secondary data with the linear regression, and grey dashed line is an estimate from nonparametric approach. Regression line over-estimates when the conditioned secondary is high or low. Estimates from two different approaches are close when intermediate range of secondary values is given. Right plot in the figure represents the estimation variances (horizontal axis) versus the given secondary values (vertical axis). Locally varying variances (dashed line) fluctuates from -59.9% to 70.1% compared with the constant variance 0.588 = $1-\rho^2$ (solid line). The real benefit of the nonparametric approach would be observed when the secondary derived estimation variance is subsequently anchored into the local uncertainty distribution obtained by simple kriging with well data using BU equation.

Integrating more than two secondary variables is nothing but a problem of dimension. Seismic amplitude data is added as another secondary variable. The modeling of joint distribution is required in 3-D probability space. Three variables are transformed into normal unit and denoted by Z, Y_1 and Y_2 for the primary permeability and secondary porosity and seismic amplitude. $f(z,y_1,y_2)$ are modeled and constrained by marginal distributions, $f(z)$ and $f(y_1,y_2)$. Figure-11 shows a 3-D visualization of trivariate pdf cut by arbitrary sections. Lower order marginal distributions are reproduced from the modeled $f(z,y_1,y_2)$ and they all reasonably honor the experimental data scatter plots and reference secondary marginal distribution. Resulting estimates and estimation variance are shown in the figure-12. Figure-13 demonstrates the accuracy plots of secondary data derived results from different approaches. Goodness statistics is shown on each accuracy plot (Deutsch, 1997).

Excessive Local Variance

The considered nonparametric approach provides locally varying estimate and estimation variance based on the extracted conditional pdf at a given secondary value. There is no bound of the estimates; however, the estimation variance has a limit to maximum 1, that is $0 \le \sigma^2_S(\mathbf{u}) \le 1$, $\mathbf{u} \in A$. It happens that the calculated local estimation variance is greater than 1 when the extracted conditional pdf at a given secondary value is spread out. As smaller kernel bandwidths are considered when joint pdf is modeled using kernel density estimator, larger dispersion in conditional pdf is often observed. Figure-12 schematically illustrates this case. Bivariate pdf is modeled satisfying all marginal relations. Given the bivariate pdf, conditional distribution at a certain value of $y(\mathbf{u})$ is drawn. Although sum of conditional pdf $f(z|y(\mathbf{u}))$ over $y(\mathbf{u})$ amount to the global pdf $f(z)$, the shape of constituent conditional pdf is various in shape. The variance numerically calculated is high and perhaps larger than 1 if the extracted conditional pdf resembles bimodal shape.

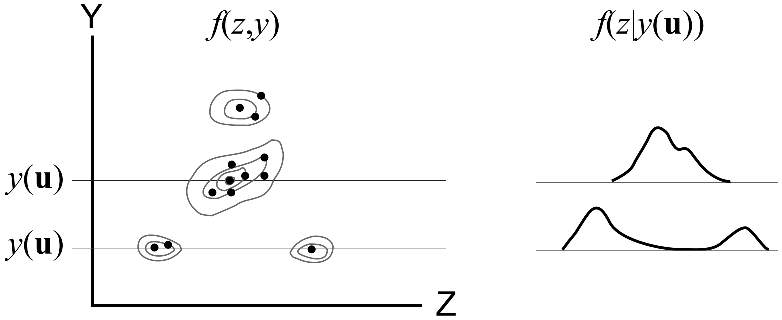


Figure-14: Schematic illustration showing the case of when the excessive local conditional variance happens.

One option for correcting the excessive conditional variance is to reset the variance into 1 if it is greater than 1. This option can be simply implemented with the loss of smooth changes in estimation variance. Another option is to change the conditional variance smoothly. New conditional variances are obtained from adjusting the initial variance by the reduction factor. For instances,

$$\sigma^2_{\mathrm{new}}(\mathbf{u}) = \sigma^2_{\mathrm{ini}}(\mathbf{u})\left(\frac{1}{\max.\mathrm{var}}\right)^{\left(\frac{\sigma^2_{\mathrm{ini}}(\mathbf{u})}{\max.\mathrm{var}}\right)} \tag{24}$$

where *max. var* is the maximum value among the initial conditional variances $\sigma^2_{S}(\mathbf{u}), \mathbf{u} \in A$. The reduction factor in equation (24) modifies variance smoothly as shown in the figure-15.

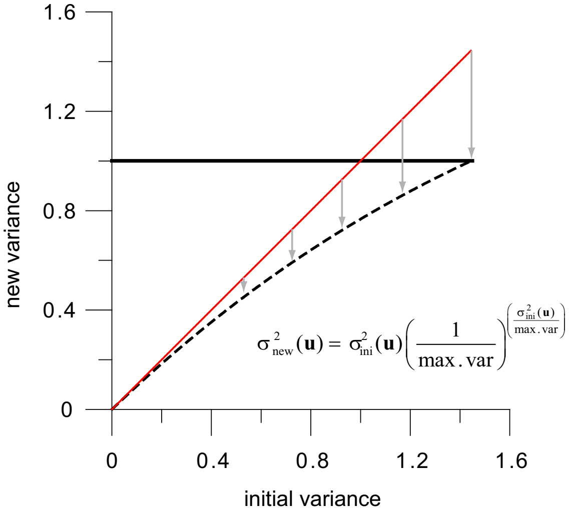


Figure-15: Correcting the initial conditional variances greater than 1 by the reduction factor.

In the figure-15, the maximum conditional variance is 1.4 and all initial variances are adjusted by a factor resulting in a dashed line. One may want to minimize the change in the initial variances if they are valid falling in [0,1]. α in the equation of figure-16 regulates the degree of changes. α=0 is equivalent to linear scaling by (1/*max. var*). Linear scaling (α=0) may be considered if the maximum value of initial conditional variance is not far greater than 1, and smooth change with α>0 might be considered otherwise. The results shown in the figure-9 and 12 are all corrected variances using the equation (24).

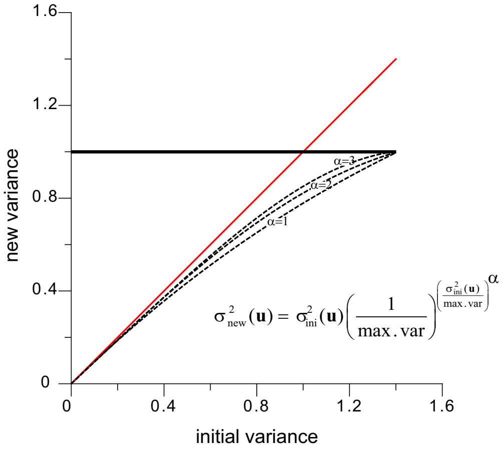


Figure-16: α controls the degree of change in the initial local conditional variances.

Conclusions

Bayesian updating has been widely adopted to integrate secondary data. The updating equation; however, is difficult to understand how information source is combined. New interpretation of Bayesian updating equation is introduced. Although conventional and new expressions are mathematically equivalent, new forms have some advantages rather than old ones. New form of updating equation decomposes the posteriori pdf into the combination of elementary pdf related to the primary and secondary data and it can be clearly understood how different information source is combined.

This work applied new expressions to the modeling of non-linear features among primary and secondary variables. Joint pdf is modeled in a nonparametric way and sequential marginal fitting algorithm refines the modeled joint pdf into the corrected joint pdf that meets all marginal constraints. The described marginal fitting algorithm directly accounts for the differences between empirical and reference marginal distributions. Given the joint pdf, the conditional pdf at a secondary value is extracted, and estimate and estimation variances are calculated using the extracted conditional pdf. The resulting estimation variance; thus, is not constant but locally varying which better reflects the relation of primary and secondary variables. In a future work, the estimated moments from secondary data should be combined with the moments from the primary data through the updating equation, and comparative study of simulation results from conventional and nonparametric approach will be performed.

References

J.-P. Chiles and P. Delfiner, 1999, Geostatistics: modeling spatial uncertainty, John Wiley & Sons, New York.

C. V. Deutsch and S. D. Zanon, 2004, Direct prediction of reservoir performance with Bayesian updating under a multivariate Gaussian model, Paper presented at the Petroleum Society's 5th Canadian International Petroleum Conference, Calgary, Alberta, 8p.

C. V. Deutsch, 1997, Direct assessment of local accuracy and precision. In: Baffi, E. Y., Schofield, N.A. (Eds.), Geostatistics Wollonggong 1996, Kluwer Academiic Publishing, Dordrecht, pp. 115-125.

P. M. Doyen, L. D. den Boer and W. R. Pillet, 1996, Seismic porosity mapping in the Ekofisk field using a new form of collocated cokriging. SPE 36498.

S. Hong and C. V. Deutsch, 2008, An alternative interpretation of Bayesian updating, Centre for Computational Geostatistics.

R. A. Johnson and D. W. Wichern, 2002, Applied to multivariate statistical analysis, Prentice Hall, New Jersey, 767p.

A. G. Journel and Ch. J. Huijbregts, 1981, Mining Geostatistics, Academic Press, London.

C. Neufeld and C. V. Deutsch, 2004, Incorporating secondary data in the prediction of reservoir properties using Bayesian updating. Centre for Computational Geostatistics.

W. Ren, J. A. Mclenan, O. Leuangthong and C. V. Deutsch, 2006, Reservoir Characterization of McMurray Formation by 2D Geostatistical Modeling, Natural Resources Research, Vol. 15, No. 2.

D. W. Scott, 1992, Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley and Sons, Inc., New York.

G. Verly, 1983, The Multigaussian approach and its applications to the estimation of local reserves, Mathematical Geology, Vol. 15, No. 2.

W. Xu, T. T. Tran, R. M. Srivastava, and A. G. Journel, 1992, Integrating seismic data in reservoir modeling: the collocated cokriging alternative. SPE 24742, Washington, DC, October 4-7.

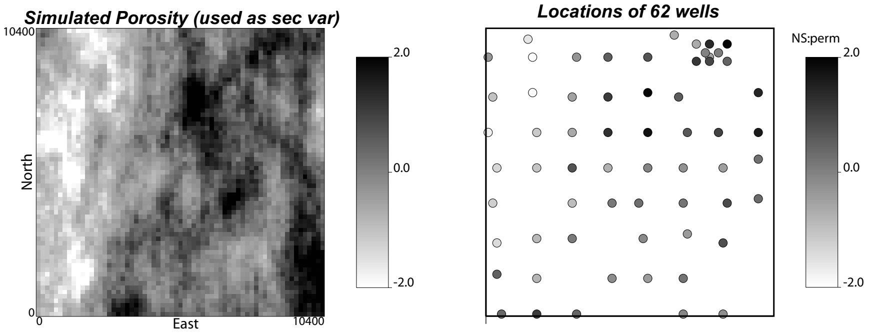


Figure-5: Simulated porosity used as secondary variable (left) for predicting permeability and sampled permeability at 62 well locations (right).

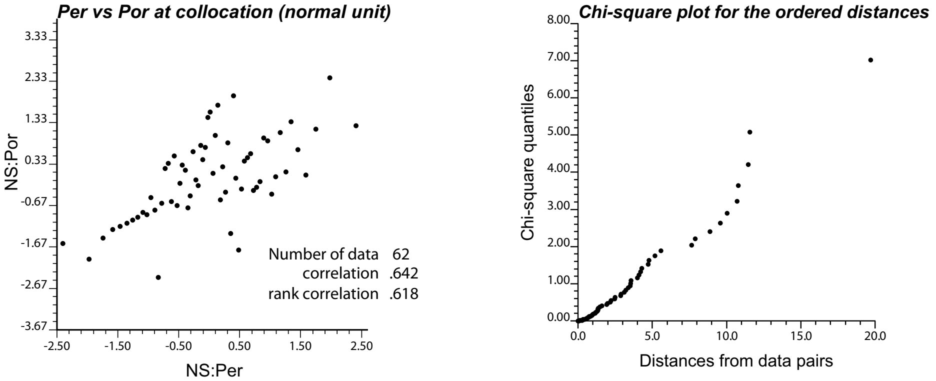


Figure-6: Cross plot of normal scored permeability (primary) and porosity (secondary) variables. To check the bivariate normality, the generalized square distances are calculated from data and plotted against the analytical chi-square distances. Systematic differences represents non-biGaussian relation.

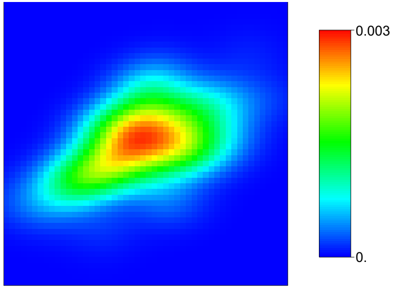


Figure-7: Modeled bivariate pdf. Horizontal and vertical axes represent the primary and secondary variables. The described marginal fitting algorithm was applied to obtain this joint pdf.

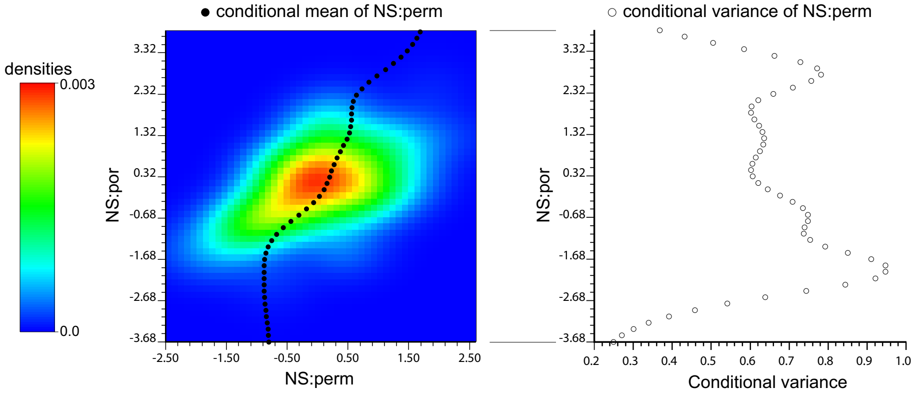


Figure-8: Conditional means and variances obtained from the joint pdf modeled in a nonparametric way. Black dots (left) are the calculated conditional means of NS:per with respect to the varying secondary values NS:por. Open circles (right) are the conditional variances with respect to the secondary values.

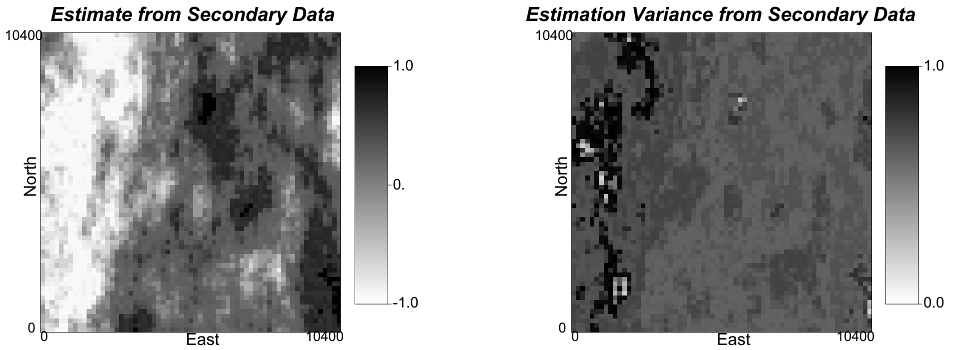

Estimate from Secondary Data
10400
North
0
0
East
10400
1.0
0.
-1.0
Estimation Variance from Secondary Data
10400
North
0
0
East
10400
1.0
0.0

Figure-9: Secondary data derived estimates and variances.

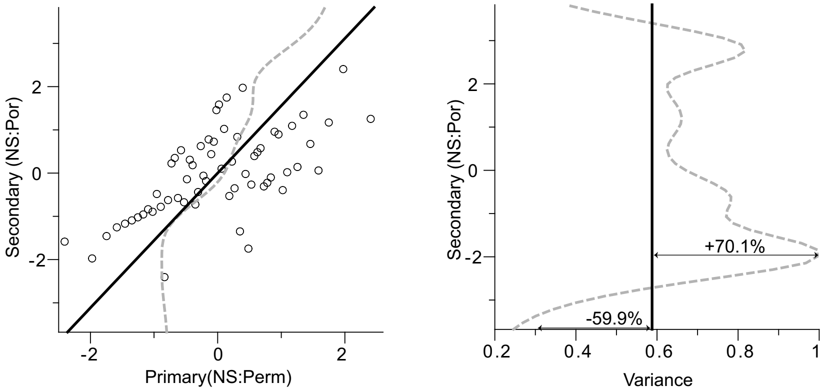


Figure-10: Comparison of estimates and estimation variance from two different approaches. Solid and dashed lines represent estimate and estimation variance using linear assumption among variables and nonparametric approach, respectively.

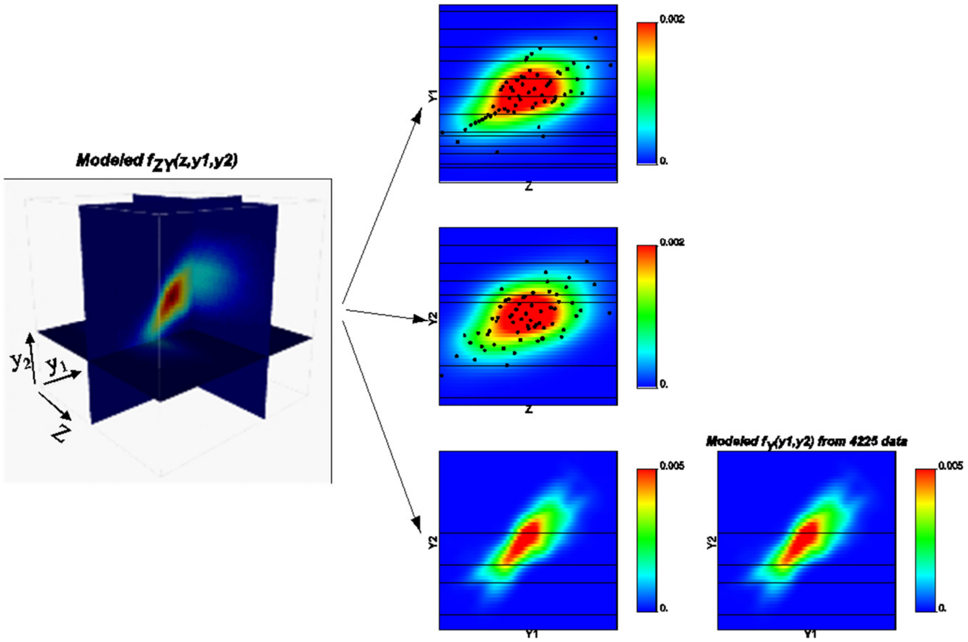


Figure-11: The joint pdf modeling with three variables: primary of permeability and secondary of porosity and seismic. Lower order of distributions are well reproduced.

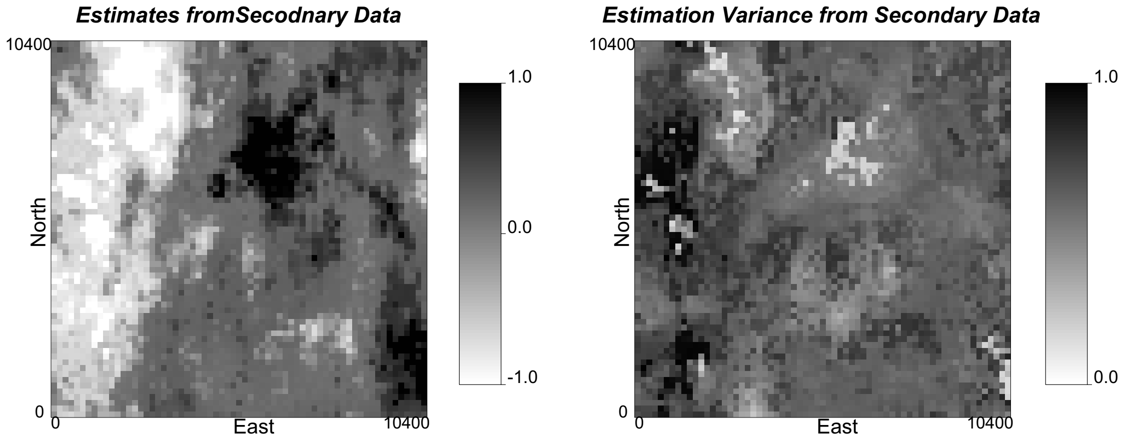

Estimates fromSecodnary Data
10400
0
North
0
East
10400
1.0
0.0
-1.0
Estimation Variance from Secondary Data
10400
0
North
0
East
10400
1.0
0.0

Figure-12: Results from integrating two secondary variables using nonparametric modeling approach.

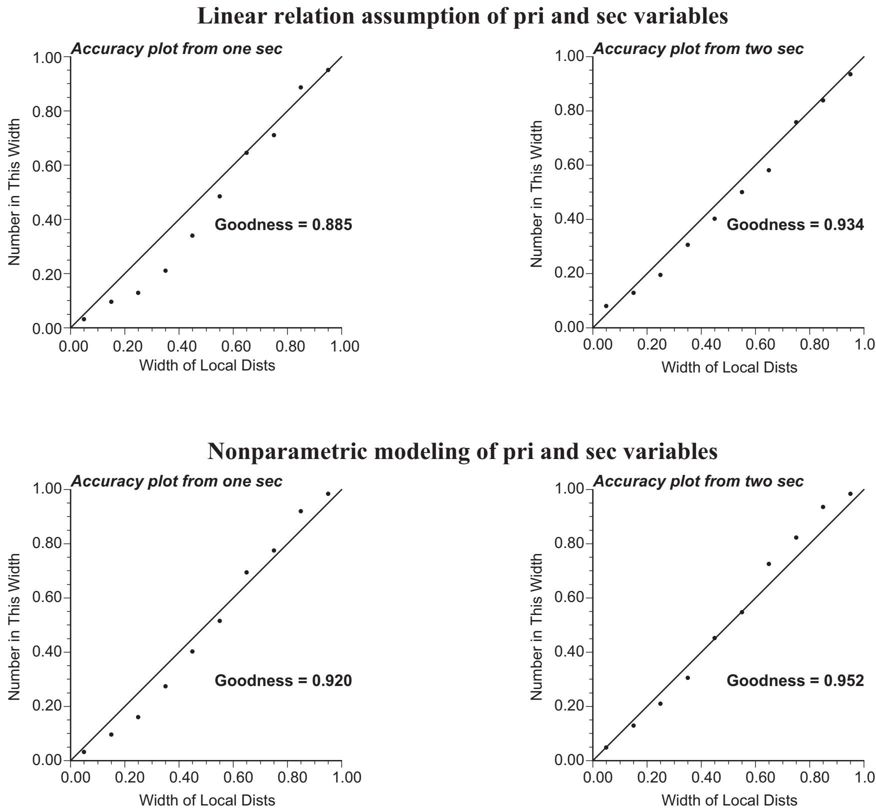


Figure-13: Accuracy plots for the results from different approaches: linear assumption (top) and nonparametric (bottom). Goodness statistics are shown on each plot (Deutsch, 1997).