

Data Integration with Non-Parametric Bayesian Updating

Chad Neufeld and Clayton V. Deutsch

Centre for Computational Geostatistics (CCG)
Department of Civil & Environmental Engineering
University of Alberta

Secondary information is important for geostatistical modeling and Bayesian updating is becoming a popular method for incorporating that secondary information. The methods currently in use are Gaussian based. However, not all data conform to the Gaussian model. A method for non-parametric Bayesian updating is developed. An example shows that the non-parametric method and Gaussian based method are identical if the non-parametric distributions are Gaussian. Another example shows how the method can be applied to a real setting.

Introduction

Bayesian updating is a method for incorporating secondary information in the prediction of reservoir properties. The Bayesian updating schemes currently in use require a multivariate Gaussian distribution between the primary and secondary data. Although, any univariate distribution can be transformed to be Gaussian, there are potential problems related to the transformation. This paper aims to develop a non-parametric approach to Bayesian updating.

Bayesian updating works by taking an estimate at a location, called the prior distribution, and combines it with the secondary information, called the likelihood distribution, to get an updated distribution that is a combination of the likelihood and prior. The updated distribution is defined by a mean and a variance. It accounts for the information contained in the likelihood and the prior estimates. This paper proposes a non-parametric approach for Bayesian updating that can be applied to any distribution shape.

Methodology

Let's start by defining an outline for the Bayesian updating process for a single location \mathbf{u} :

1. Derive the likelihood distribution. This could come from a seismic attribute, soft-data calibration, or any number of sources. The likelihood must be in the same units as the primary variable. This necessitates a calibration of the secondary data to the primary data, or a normal score transformation. As part of the calibration, the relationship between the primary and secondary data is quantified. What this means, is that the primary data value can be predicted using the secondary data and its calibration to the primary data. However, there will be uncertainty in the primary data value. The uncertainty is captured by the full distribution of the primary given the secondary. Figure 1 shows a schematic of how to calculate the likelihood distribution using a secondary variable.

2. Estimate the prior distribution. Any method could be used for estimating the prior distribution; kriging, indicator kriging, simulation, or multiple point statistics. However, the estimation is usually done with kriging.
3. Combine the likelihood, prior, and global distributions. This is the Bayesian updating step. The three input distributions are combined to calculate the updated distribution at the location \mathbf{u} .

Figure 2 shows a schematic example Bayesian updating. The green distribution is the likelihood, the red distribution is the prior, and the blue distribution is the updated.

Calculating the Likelihood Distribution

The likelihood is a function of the relationship between the primary and secondary variables. In most cokriging and cosimulation algorithms, the relationship is assumed to be bivariate Gaussian whether it is or not. It is common to have bivariate relationships that exhibit non-Gaussian behavior. These include, but are not limited to non-linear relationships, constraints, and heteroscedasticity. The calibration discussed here makes no assumptions about the bivariate relationship.

The starting point for any calibration is the crossplot between the primary and secondary variables. Typically, the primary variable is on the y-axis and the secondary is on the x-axis. We are interested in the distribution of the primary variable given a particular value of the secondary variable. The conditional distribution is calculated by binning the primary variable conditional to the secondary variable. After the binning is done, the distribution of the primary variable conditional to the secondary is calculated directly.

The only limitation to this method is sparse data. The direct calculation of the conditional distributions requires a lot of data. With sparse data, scatterplot smoothing or debiasing can be used to help infer the conditional distributions [1].

Estimating the Prior Distribution

Any method can be used for estimating the prior distribution. Kriging, indicator kriging or calibration with multiple point statistics are some methods. There are a few considerations that must be taken into account. The non-parametric approach requires the estimates to have an estimated value, Z^* , an estimated variance, σ^{*2} , and a conditional distribution that correctly relates to the original data. The last consideration is the hardest one to account for properly.

The kriging variance is not an accurate measure of the variance for a conditional distribution in original units. In addition, it does not provide any information about the shape of the estimated distribution. Simple or ordinary kriging in original units will not work. There are alternative kriging approaches that will work. Indicator kriging works extremely well. Using IK, we can estimate the cumulative probability of selected thresholds directly at each location. Multi-Gaussian kriging can also be used. Kriging is done in Gaussian units where the kriging variance is directly related to the distribution of uncertainty. Backtransforming many quantiles of the distribution provides the distribution of uncertainty in original units at each location. Multiple point statistics could also be used to estimate the threshold probabilities.

Bayesian Updating in a Gaussian Context

It is straight forward to apply Bayesian updating in a Gaussian setting [3], [5], [6]. The updated distribution is calculated using Equations (1) and (2):

$$Y_U = \frac{Y_L \sigma_p^2 + Y_p \sigma_L^2}{[1 - \sigma_L^2][\sigma_p^2 - 1] + 1} \quad (1)$$

$$\sigma_U^2 = \frac{\sigma_p^2 \sigma_L^2}{[1 - \sigma_L^2][\sigma_p^2 - 1] + 1} \quad (2)$$

Where Y_p and σ_p^2 are the mean and variance of the prior distribution, Y_L and σ_L^2 are the mean and variance of the likelihood and Y_U and σ_U^2 are the mean and variance of the updated distribution.

There are some important points to note:

- The weight that each distribution receives changes. The weights are a function of the distribution variance. For example, as the variance of the likelihood increases, the amount of information it contains decreases and it receives a lower weight relative to the prior distribution for calculating the updated distribution.
- The global distribution plays a role in the updating process. The Bayesian updating methodology in Equations (1) and (2) assume that the global distribution is standard normal with a mean of zero and a variance of one. This was an important point for developing the non-parametric Bayesian updating workflow.
- The updating process is non-convex. Since the global distribution plays a role in the updating process, it is possible to get an updated distribution that has a mean that is in-between the likelihood and prior, or it is also possible to get an updated mean that is higher or lower than both the likelihood and prior distributions.

Non-Parametric Bayesian Updating

The Gaussian based Bayesian updating is applied on distributions. The approach relies on a multi-variate Gaussian assumption. We cannot make that assumption for a non-parametric approach. Two approaches were developed for non-parametric Bayesian updating: (1) a methodology based on an assumption of independence and (2) an approach based on the permanence of ratios assumption. The methodology used for both approaches is similar. The same probability values are input, but they are combined in a different way.

The non-parametric approach cannot be applied in a single step like the Gaussian method. It is instead applied on probability intervals for specified thresholds or categories. This means that at a particular location, we need to perform multiple steps to complete the Bayesian updating.

Consider the distribution of the variable Z shown in Figure 3. The PDF is on the left and the CDF is on the right. A probability interval represents the probability that Z falls within a certain range for continuous variable or that Z takes a specific discrete value for categorical variables. The interval probability for a category $z(i)$ is the probability that Z is between a lower bound, called $z(i-1)$, to an upper bound, called $z(i)$:

The interval probability can be calculated from the PDF:

$$P(Z(i)) = \int_{z(i-1)}^{z(i)} f(z) dz \quad (3)$$

Or from the CDF:

$$P(Z(i)) = F(z(i)) - F(z(i-1)) \quad (4)$$

We will use probability intervals from now on. Three input probabilities are required to calculate the updated probability. Figure 4 shows the four probabilities we are interested in. Although these are shown as a single line, they represent interval probabilities. $P(A|B,C)$ is the updated probability of A , that Z is equal to z_c , given the likelihood, prior, and global distribution. $P(A)$ is the probability of A on a global scale. $P(A|B)$ is the probability of A given the likelihood estimate. $P(A|C)$ is the probability of A given the prior estimate. $P(A|B,C)$ is calculated from the other three input probabilities.

Independence

The first method assumes that the information from the likelihood and prior estimates is independent. This means that the two data sources B and C provide information about A , while being unrelated. For example, if both B and C say that A is supposed to be high, then A will be very high. The updated probability is calculated as follows [4]:

$$P_l^*(A|B,C) = \frac{P(A|B) \cdot P(A|C)}{P(A)} \quad (5)$$

Unfortunately, this is not usually the case. If both A and B are related to C , then they must be related to each other. The permanence of ratios, or conditional independence, approach is an alternative to the independence approach.

Permanence of Ratios

Permanence of ratios assumes that the incremental information about A from B and C are conditionally independent, see [4] for more information. The updated probability is calculated with the following equation:

$$P_{PR}^*(A|B,C) = \frac{\frac{1-P(A)}{P(A)}}{\frac{1-P(A)}{P(A)} + \left(\frac{1-P(A|B)}{P(A|B)} \right) \left(\frac{1-P(A|C)}{P(A|C)} \right)} \quad (6)$$

This method does not exaggerate the extreme high and low values like full independence method.

Combining the Distributions

The following steps are required to combine the likelihood and prior distributions:

1. **Assemble the required information:** This includes the global distribution for the cut-offs or categories of interest, the likelihood distributions at each location \mathbf{u} , and the estimated values at each location.
2. **Combine the distributions:** Calculate the updated probability for each cutoff or category. This is done using Equation (5) or (6). The independent updating approach is theoretically correct, but the permanence of ratios is more stable when working with outliers and extreme values.

In addition to updating the probabilities for the *ncat* thresholds, the probability to above the highest category must also be calculated. This is required for the restandardization in step #3.

3. **Restandardize the probability values:** The updated probabilities do not sum to 1 by construction. A simple restandardization is required to ensure closure. The following equation can be used:

$$P(Z(i)) = \frac{P(Z(i))}{\sum_{i=1}^{ncat} P(Z(i)) + P(Z > Z(ncat))} \quad (7)$$

The updated probabilities define final distribution of uncertainty at **u**.

Example 1 – Comparison with the Gaussian Case

A test and comparison of the non-parametric Bayesian updating was done. This was done to ensure that the methodology produced the desired results. The non-parametric approach was tested against the standard Gaussian based Bayesian updating. Two cases were tested here: (1) a fine discretization of the CDF, and (2) a coarse discretization of the CDF for the non-parametric approach. The prior and likelihood distributions were chosen. They were not calculated or estimated. The following input parameters were used for the Bayesian updating:

$$\begin{aligned} Y_L &= -0.50 & Y_p &= 1.50 \\ \sigma_L^2 &= 0.60 & \sigma_p^2 &= 0.30 \end{aligned}$$

The updated distribution was calculated directly using the Gaussian based Bayesian updating Equations (1) and (2):

$$\begin{aligned} Y_U &= \frac{(-0.5)(0.3) + (1.5)(0.6)}{(1-0.6)(0.3-1) + 1} \\ &= 1.04 \\ \sigma_U^2 &= \frac{(0.3)(0.6)}{(1-0.6)(0.3-1) + 1} \\ &= 0.25 \end{aligned}$$

The updated distribution was then calculated using the non-parametric approach. A very fine discretization was used for the first check. Using a fine discretization ensures that the updated distribution should be accurately reproduced.

The prior, likelihood, and global distributions were discretized in increments of 0.05 from -4.0 up to 4.0. At each increment, the CDF value was calculated. The CDF values were used to determine the increment probabilities for the non-parametric Bayesian updating approach. The updated distribution was calculated for each interval and then restandardized to get the final updated distribution.

The 2 updated distributions were compared visually and with the Kolmogorov-Smirnov test. The updated distributions are identical. Figure 5 shows the updated distributions along with the input

distributions. The updated distributions for the parametric and non-parametric approaches plot on top of each other. The K-S test confirms this. The distributions are statistically similar.

The next step was to compare the methods using a coarse discretization for the non-parametric approach. For this case, the discretization interval was changed from 0.05 to 0.5. The results are shown in Figure 6. The underlying lines show the analytical results from Figure 5. The non-parametric results are shown as the bullets on top of the lines. The coarse non-parametric approach matches the analytical result visually. The K-S confirms this.

This example has shown that the results from the non-parametric approach match the parametric Gaussian based approach. The results were the same for a fine or coarse discretization of the distributions.

Example 2 – Amoco Data Set

For this example, the porosity will be mapped using indicator kriging and then updated using the secondary information from the seismic attribute. The advantage of using indicator kriging and the non-parametric Bayesian updating is removing the reliance on the Gaussian model: different variograms may be used for the high and low thresholds, and the relationship between the seismic attribute and porosity does not have to be multi-variate Gaussian.

The example was done using the publicly available Amoco data set. The Amoco data set contains a set of 63 wells that have been averaged vertically for this example. In addition to the well data, a set of 2D seismic is available at all location within the modeling area. The well locations and seismic data are shown in Figure 7. The 2D seismic and 2D well data were used to illustrate the application of non-parametric Bayesian Updating.

Figures 8 through 10 show the exploratory data analysis. The histograms of porosity and seismic at the well locations are shown in Figure 8. Figure 9 shows the histogram of the seismic attribute everywhere. The crossplot between porosity and seismic is shown in Figure 10. Note that the bivariate relationship is not Gaussian. It has a definite non-linear shape. The smoothed crossplot is shown in Figure 10.

Ten indicator thresholds were used for modeling the porosity. The thresholds were chosen to correspond to the centered quantiles. Table 1 lists the thresholds and their cumulative probability values.

Table 1: Indicator thresholds used for modeling porosity.

Threshold	Cumulative Probability
5.25	0.05
6.00	0.15
6.71	0.25
7.33	0.35
8.23	0.45
9.14	0.55
9.66	0.65
9.93	0.75
10.33	0.85
11.12	0.95

Indicator variograms were calculated and fit for each threshold. The fitted variograms are shown in Figure 11. Omni-directional variograms were used for all thresholds. Each variogram had a

nugget effect of zero and a single spherical structure. The parameters for the fitted models are in Table 2.

Table 2: Fitted variogram models for the Amoco example.

Threshold	Sill	Range
5.25	0.0475	625
6.00	0.1275	1125
6.71	0.1875	2000
7.33	0.2275	2250
8.23	0.2475	2000
9.14	0.2475	1250
9.66	0.2275	1000
9.93	0.1875	375
10.33	0.1275	325
11.12	0.0475	250

The next step is to calculate the likelihood function. This involves a calibration of the seismic and porosity data, the cross plots, and then transferring the calibration to the estimation grid. Consider the smoothed crossplot shown in Figure 10. When the seismic is low, <44000, the porosity is also low. In fact, the porosity does not go higher than 7%. When the seismic is high, >50000, the porosity is also high. The histograms in Figure 12 show the conditional porosity distributions for 2 different seismic values.

After the calibration was done, the conditional porosity distributions were transferred to the estimation grid. At each location, the seismic value was read and the corresponding porosity distribution was looked up. The cumulative probabilities for each threshold were calculated at each location. The gridded likelihood function is shown in Figure 13. Note that the signature of the seismic map is reproduced in the likelihood.

Indicator kriging was used to estimate the conditional CDF values. The results of the indicator kriging are shown in Figure 14. The plots show the probability of the primary variable to be less than the specified threshold. The well locations are coded as blue when they are below the threshold and red when they are above it.

The likelihood and prior distributions were then merged using the non-parametric approach. Figure 15 shows the results of the updating. The updated results provide the CDF at each location. PostIK was used to calculate the estimated porosity value at each location. 200 quantiles were calculated and averaged to get the estimated value. Figure 16 shows the estimated porosity map with no seismic on the left, and the estimated porosity with seismic on the right. The impact of the seismic attribute is easy to see.

The non-parametric updating also improved the bivariate relationship between seismic and the estimated porosity. The crossplots of seismic and estimated porosity are shown in Figure 17 with the input smoothed distribution underneath. The output crossplot with no seismic is on the left and the crossplot where seismic was used is on the right. The non-parametric Bayesian updating was able to reproduce the non-Gaussian input relationship.

There is not a marked improvement from the non-parametric Bayesian updating. This is due to the fact that there are many wells in the study area that constrain the estimation. When there are

fewer wells, or the range of correlation is shorter, the Bayesian updating will have a larger impact.

Programs

Several programs were written to do the non-parametric Bayesian updating. The first program, `build_lh_np`, takes a calibration table and calculates the gridded likelihood for a set of specified thresholds. The program, `update_np`, takes the likelihood and prior distributions and calculates the updated distribution.

The calibration table takes a very specific format. Excel was used to build the calibration table for Example 2. However, any program may be used as long as the output conforms to the standard. The first line in the file contains the number of classes for the secondary variable, `nsec`, and the number of classes for the primary variable for each class of the secondary variable, `npri`. The next `nsec` lines contain the values of the secondary variable that define the class boundaries. The following `nsec * npri` lines contain the calibration information. The first column contains a class boundary for the primary variable within the secondary variable and the second column contains the probability (PDF) of the primary variable. A short example is below:

```

2      2
  45000.0000
  50000.0000
    4.0000    0.80000
    8.0000    0.20000
    6.0000    0.30000
   10.0000    0.70000

```

The `build_lh_np` program reads in the calibration table and the gridded secondary variable to calculate the gridded likelihood. The program calculates the CDF value for each threshold specified. The results are written to a single output with 1 column for each threshold.

```

Parameters for BUILD_LH_NP
*****

```

```

START OF PARAMETERS:
  5                                -number of cutoffs
  0.5  1.0  2.5  5.0  10.0        - thresholds / categories for modeling
sec_data.dat                        -file with secondary variable
  1                                - column for secondary variable
116  213  62                       - nx, ny, nz
build_lh_np.out                     -file for output
calibration.dat                     -file for calibration data

```

The updating program, `update_np`, reads in the likelihood and the prior estimates and calculates the updated distribution. A required input is the global distribution at the specified thresholds.

```

Parameters for UPDATE_NP
*****

```

```

START OF PARAMETERS:
  5                                -number thresholds/categories
  0.5  1.0  2.5  5.0  10.0        - thresholds / categories
  0.12  0.29  0.50  0.74  0.88    - global cdf / pdf
likelihood.out                       -file with likelihood
  1  2  5  4  5                    - columns for the likelihood probabilities

```

```

ik3d.out           -file with prior distribution
  1  2  3  4  5    - columns for the prior probabilities
 50 50  1          -grid size: nx, ny, nz
update_np.out      -file for output

```

The output from the updating is a file with a column for the updated probability of each threshold. A program like `PostIK` can be used to calculate a single estimated value or a specified quantile.

Future Work

The methodologies developed so far have been focused on a simple case. There was only one secondary attribute used to predict a single primary variable. Additional work could involve:

1. **Multiple secondary variables:** Either expanding the algorithm to use multiple secondary variables directly, or developing a methodology for combining secondary variables before the likelihood calculation.
2. **Categorical variables:** This method could also be applied to estimation of categorical variables. Updating the indicator probabilities for categorical variables would require modifications to the method presented in this paper.
3. **Simulation:** An advantage to Bayesian updating is that we have the conditional distributions at each location. A simulation method, like p-field, could be used to generate conditional simulations from the conditional distributions very quickly.

Conclusions

Bayesian updating has become a popular method for incorporating secondary information. The previous Bayesian updating methods were based on the Gaussian model. This requires a transformation and a strong distributional assumption. With complicated data/relationships, the transformations becomes complex and error prone. The non-parametric approach presented in this paper makes no assumptions about the shape of the distributions. It is capable of using any distribution shape or type for the estimation.

The updating method requires three inputs: (1) the global distribution, (2) a prior distribution, and (3) a likelihood distribution. The global distribution is from the data. The prior distribution is estimated from the data. Any estimation method can be used, see the requirements mentioned earlier. Kriging, indicator kriging, simulation, and multiple point statistics are all acceptable methods. The likelihood distribution is a function of the secondary variable and the calibration between the primary and secondary variables. The conditional distributions for the primary variable can be calculated directly from the data, or determined through debiasing if the data are sparse.

The output from the updating is a distribution of uncertainty at each location that accounts for the primary data and the secondary information. A postprocessing method, like `PostIK`, can be used to calculate a final estimate from the distributions.

Example 1 showed that Gaussian based Bayesian updating and the non-parametric Bayesian updating are equivalent. When applied to the same problem, they give the same results. Example 2 showed how the method can be applied to a real data set for estimating the porosity using the well data and the seismic information.

References

- [1] C. V. Deutsch and T. Dose. Programs for Debiasing and Cloud Transformation, Paper 404, In *Centre for Computational Geostatistics, Report 7*, 2005.
- [2] C. V. Deutsch. *Geostatistical Reservoir Modeling*. Oxford University Press, New York, 2002.
- [3] P. M. Doyen, L. D. den Boer, and W. R. Pillet. Seismic porosity mapping in the Ekofisk field using a new form of collocated cokriging. *Society of Petroleum Engineers*, 1996. SPE 36498.
- [4] A. G. Journel. Combining knowledge from diverse sources: An alternative to traditional data independence hypotheses. *Mathematical Geology*, 34(5):573–596, July 2002.
- [5] C. Neufeld and C. V. Deutsch. Incorporating Secondary Data in the Prediction of Reservoir Properties Using Bayesian Updating. In *Centre for Computational Geostatistics, Report 6*, 2004.
- [6] S. Zanon and C. V. Deutsch. Direct Prediction of Reservoir Performance with Bayesian Updating Under a Multivariate Gaussian Model. In *Petroleum Society's 5th Canadian International Petroleum Conference (55th Annual Technical Meeting)*, Calgary, Alberta, Canada, June 8, 2004.

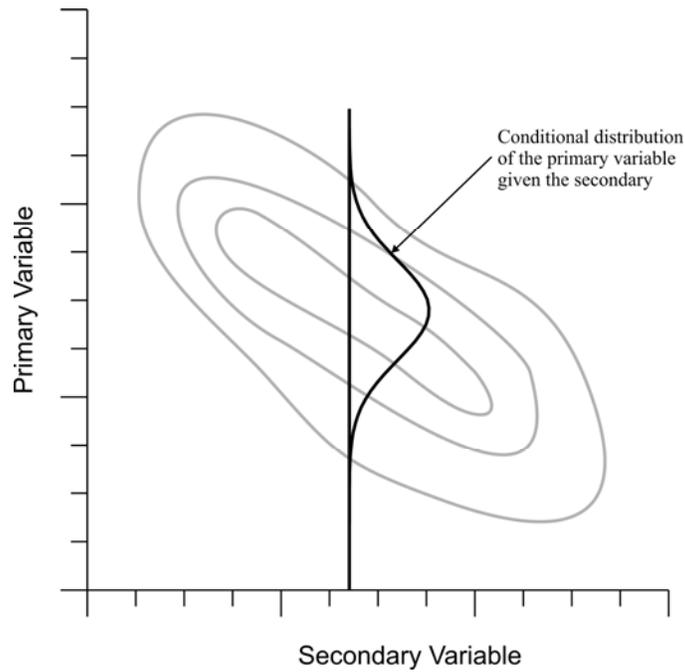


Figure 1: Calibration of the primary variable to the secondary variable. There will always be uncertainty in the primary variable when it is estimated from the secondary variable.

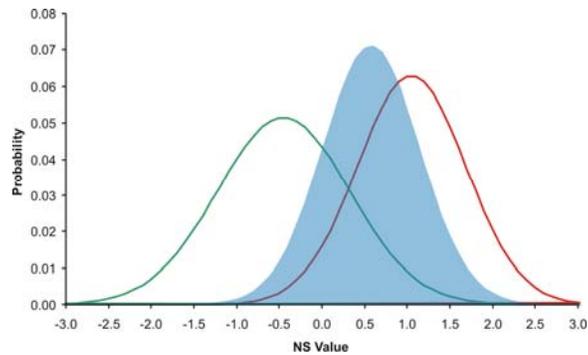


Figure 2: Schematic illustration of Bayesian updating. The likelihood distribution is represented by the green line, the prior distribution is represented by the red line, and the updated distribution is the shaded blue distribution.

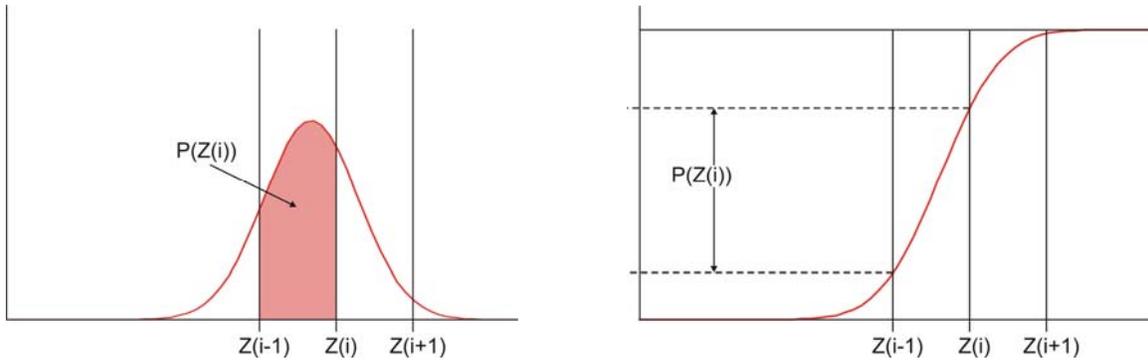


Figure 3: Interval probability calculation. Calculating the interval probability from a PDF, on the left, or a CDF, on the right, will produce the same result.

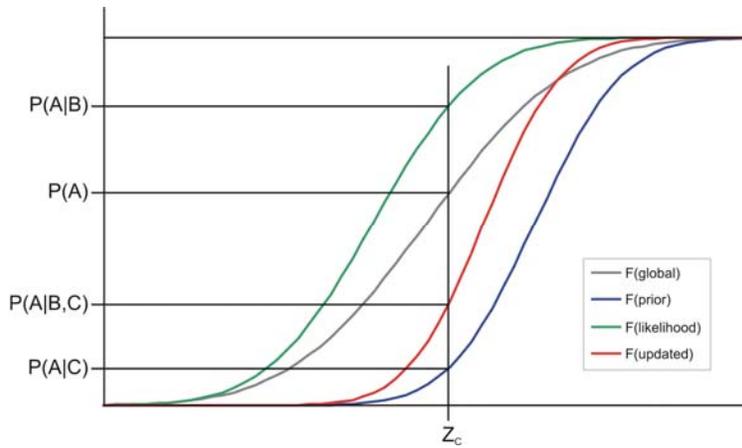


Figure 4: Schematic showing all four probability values that are associated with the Bayesian updating process. $P(A)$ is from the global distribution. $P(A|B)$ is from the likelihood calculation. $P(A|C)$ is from the prior distribution. $P(A|B,C)$ is calculated from the other three probabilities.

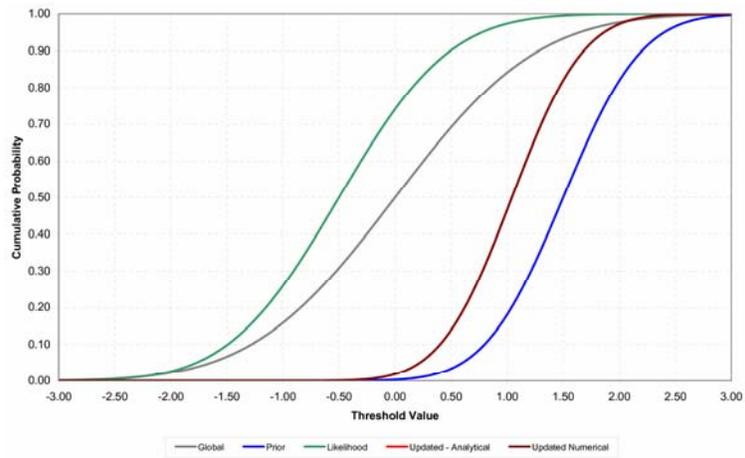


Figure 5: Comparison of Bayesian updating methods. A fine discretization was used for the non-parametric approach. The analytical and numerical results are identical. The analytical distribution is underneath the numerical distribution.

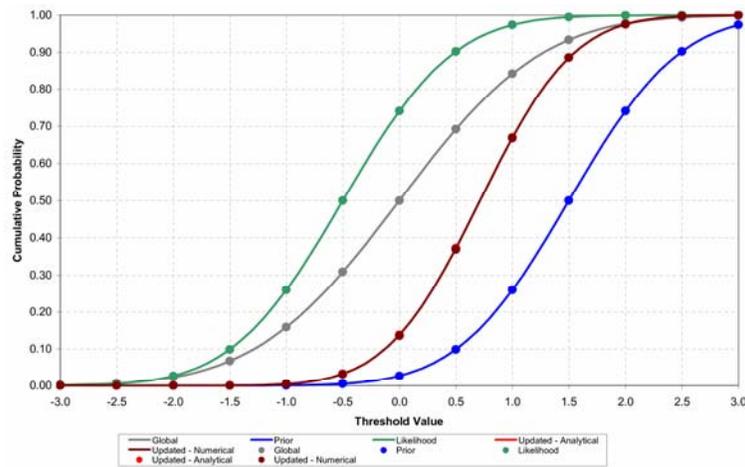


Figure 6: Comparison of Bayesian updating methods. A coarse discretization was used for the non-parametric approach. Again, the analytical and numerical results are identical. The non-parametric results are shown as the points, while the analytical approach is shown as the lines.

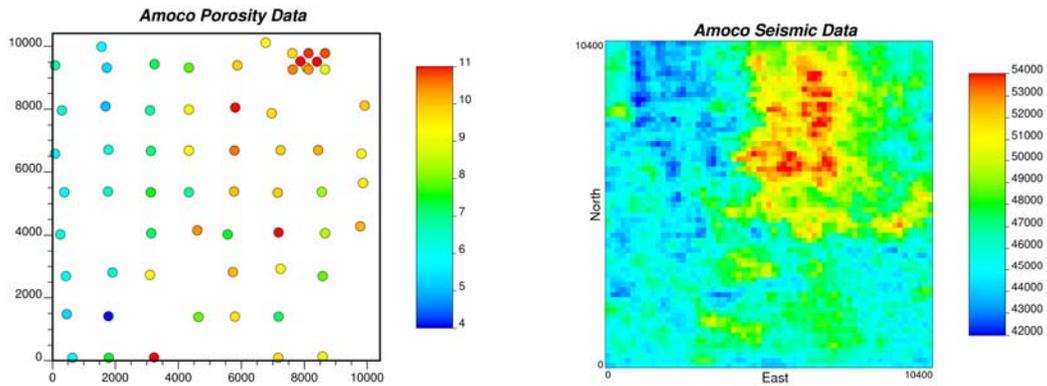


Figure 7: Amoco data set. The well locations with porosity are shown on the left, and the 2D seismic data is on the right.

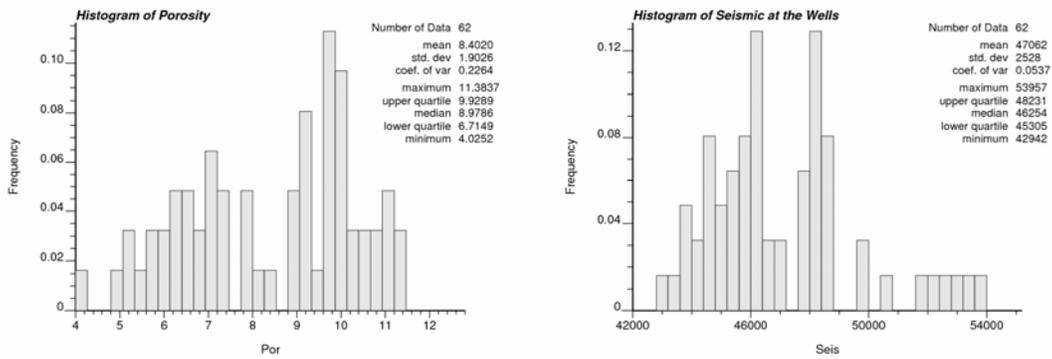


Figure 8: Histograms of porosity and seismic at the well locations.

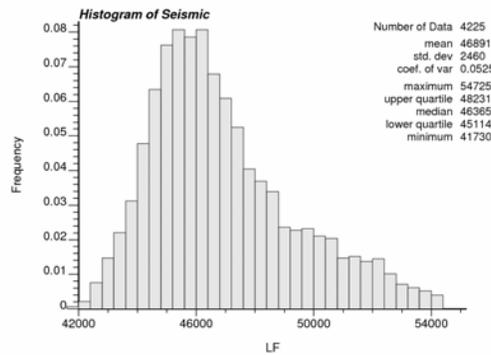


Figure 9: Histogram of the seismic everywhere.

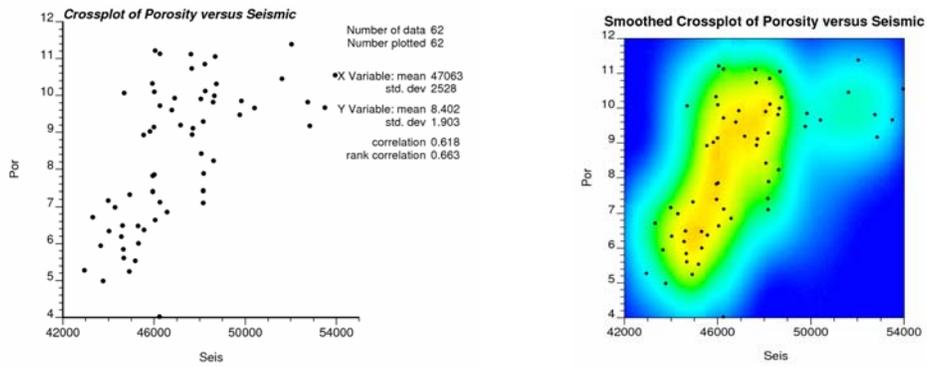


Figure 10: Crossplots of porosity versus seismic. The smoothed crossplot is shown on the right.

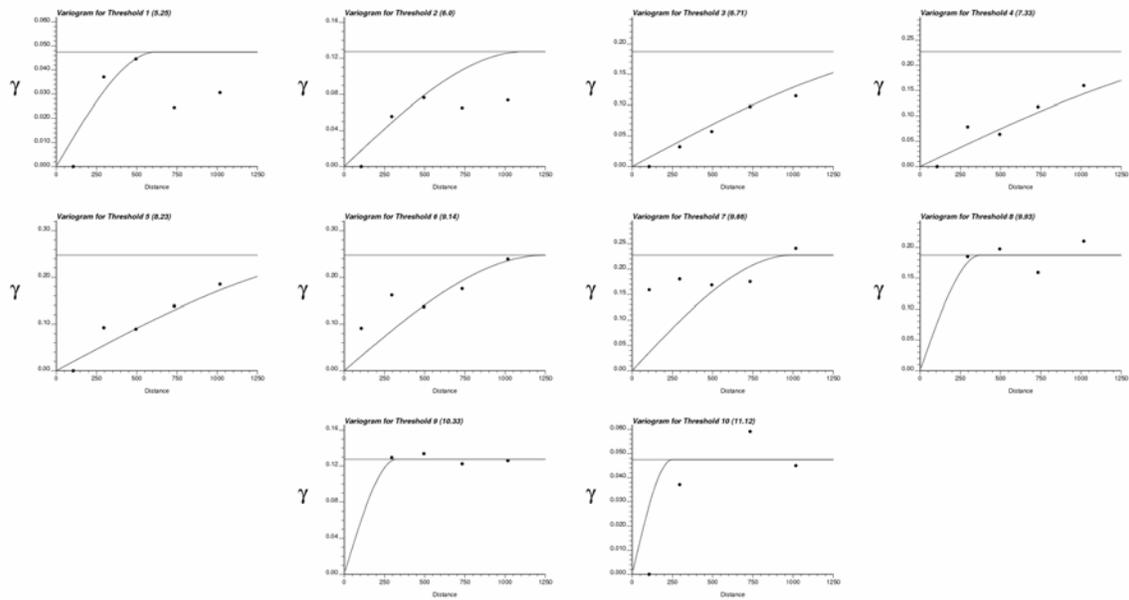


Figure 11: Fitted indicator variograms for the Amoco data set.

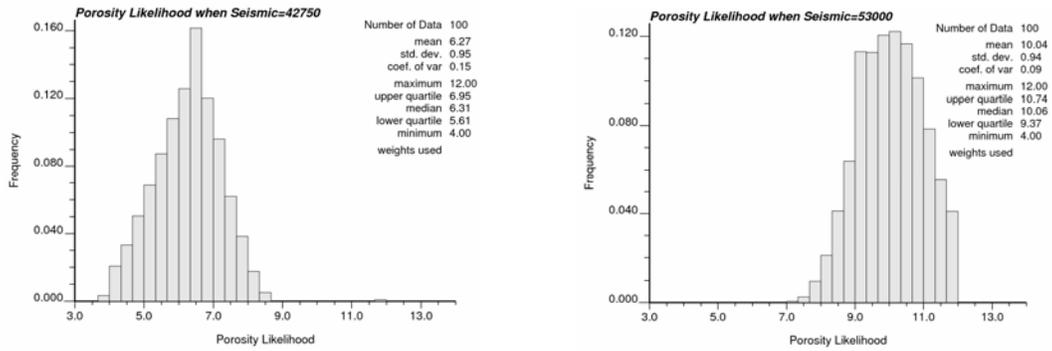


Figure 12: Porosity likelihood for 2 different seismic values. The likelihoods are calculated from the crossplot between seismic and porosity.

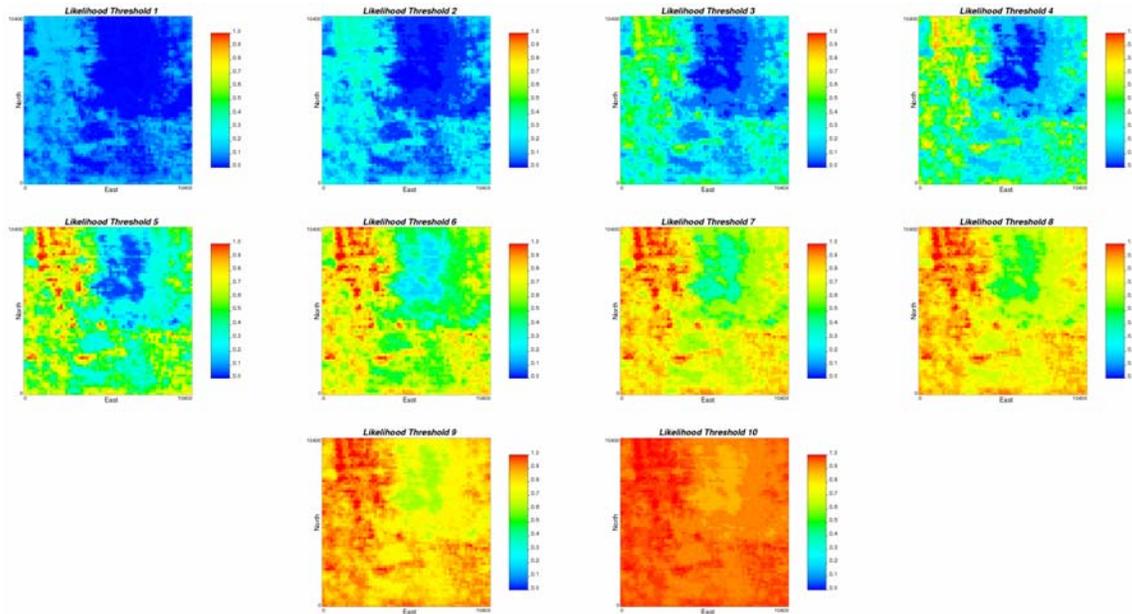


Figure 13: The non-parametric likelihood. The plots show the probability of the primary variable to be less than the specified threshold.

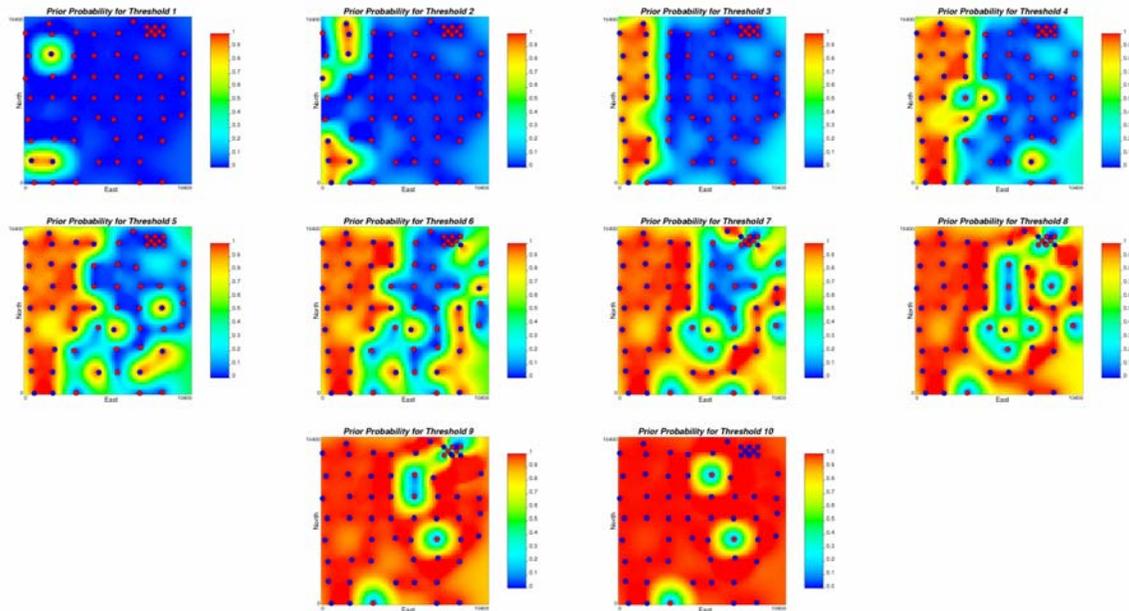


Figure 14: The non-parametric prior estimates. The plots show the probability of the primary variable to be less than the specified threshold. The well locations are coded as blue when they are below the threshold and red when they are above it.

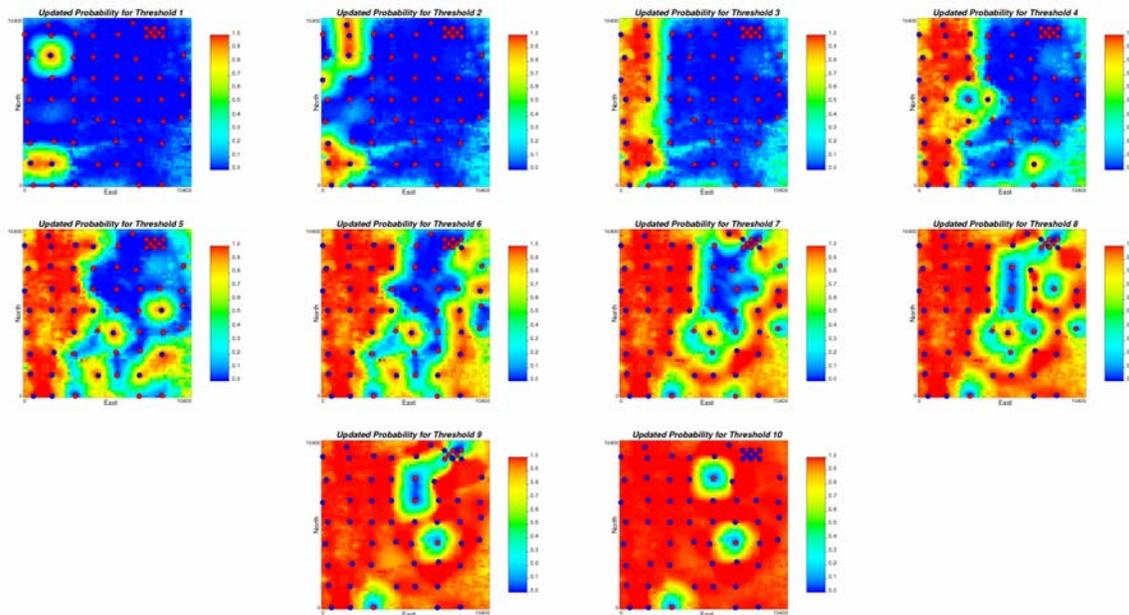


Figure 15: The updated non-parametric estimates. The plots show the probability of the primary variable to be less than the specified threshold.

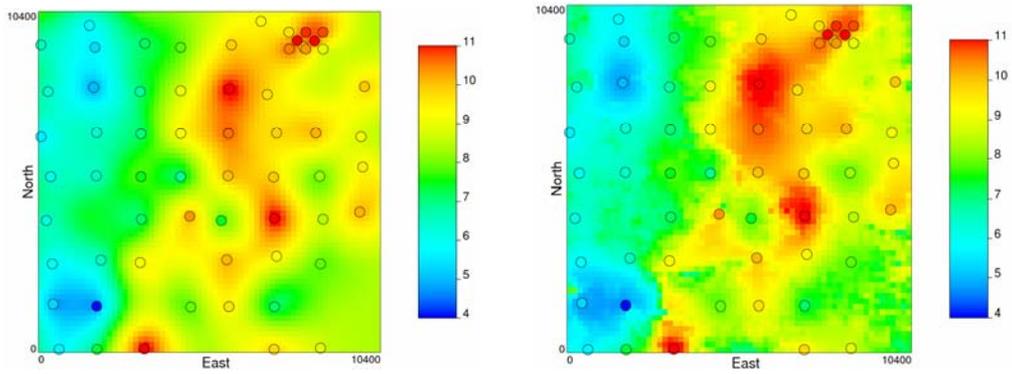


Figure 16: A map of the estimated porosity without any seismic information is shown on the left and the same map where seismic was used is shown on the right.

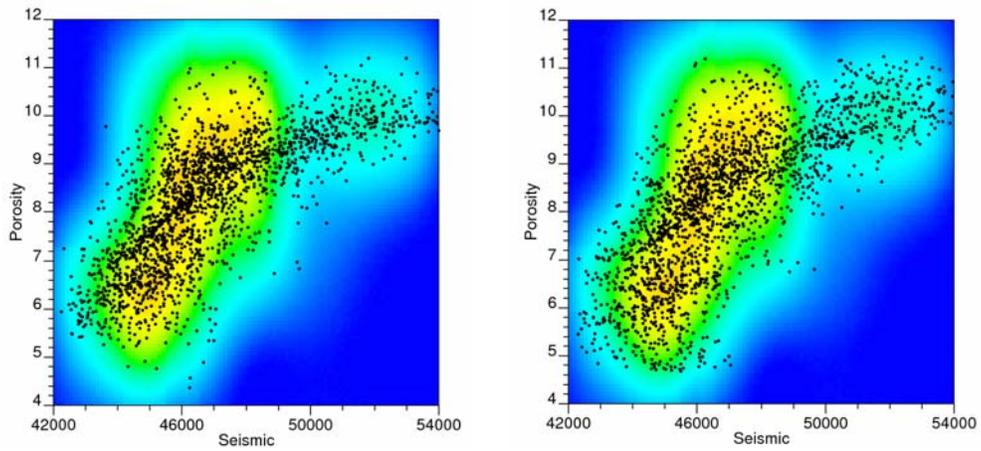


Figure 17: A map of the estimated porosity is shown on the left and a crossplot of the estimated porosity versus seismic is on the right.