PETROLEUM SOCIETY
CANADIAN INSTITUTE OF MINING, METALLURGY & PETROLEUM

PAPER 2004-047

CANADIAN INTERNATIONAL
PETROLEUM CONFERENCE

Direct Prediction of Reservoir Performance with Bayesian Updating Under a Multivariate Gaussian Model

C.V. DEUTSCH
University of Alberta

S.D. ZANON
University of Alberta

This paper is to be presented at the Petroleum Society's 5th Canadian International Petroleum Conference (55th Annual Technical Meeting), Calgary, Alberta, Canada, June 8 – 10, 2004. Discussion of this paper is invited and may be presented at the meeting if filed in writing with the technical program chairman prior to the conclusion of the meeting. This paper and any discussion filed will be considered for publication in Petroleum Society journals. Publication rights are reserved. This is a pre-print and subject to correction.

Abstract

Conventional geostatistics aims at creating models of heterogeneity and uncertainty in static rock properties such as facies, porosity, and permeability. This approach is appropriate for calculating in place resources and providing input to flow simulation. There are times, however, when no flow simulation is going to be performed and we would like to directly predict reservoir flow characteristics. Different techniques are required when the aim is to directly create maps of the (uncertainty in) production potential. This paper summarizes a practical and useful technique for this purpose.

The petroleum industry is reliant on many types of geological and geophysical information to predict reservoir performance. This data covers different areas, provides data on different scales, and is variably correlated to the production characteristics we are trying to predict. Statistical techniques can be used to summarize the relationships between the variables; however, they do not account for spatial correlation. Geostatistical techniques incorporate spatial structure but these techniques are cumbersome in the presence of many secondary variables. We propose that all secondary data be merged statistically by a multivariate Gaussian approach into a single variable that contains all of the secondary variable information; this provides a likelihood distribution. The spatial distribution of each variable by itself is mapped independently of the secondary variable information; this provides a prior distribution. The likelihoods and priors are merged to provide an updated posterior distribution. This technique has been successfully applied in a number of cases. We describe the methodology and show a synthetic example for illustration.

Introduction

Our goal is to directly predict reservoir performance potential summarized by some production variables. The production variables we are predicting are measures of hydrocarbon flow rate and projected cumulative production. Implicitly we assume that the wells are far enough apart so that they are not interacting together in any significant way.

Reservoir characterization uses every data source and interpretive tool possible to improve understanding of the reservoir performance potential at locations where we have no wells. In general, we can group the available data into:

- **Geological variables** that take two forms: (1) maps of interpreted variables where the regional depositional setting is taken into consideration and some expert judgement is accounted for in the map making, and (2) direct well measurements of variables such as porosity, pay thickness and so on. Another grouping of geological variables is into structural and geological variables where the structural variables relate to the container size and shape and the geological variables relate to the internal reservoir quality.
- **Geophysical variables** that have high areal resolution, low vertical resolution, and variable correlation to actual rock properties and production variables. These variables can be direct attributes such as amplitudes or processed variables such as interpreted fracture densities or P/S impedances.
- **Production variables** that we are trying to predict such as initial production rate and projected cumulative production. These variables would typically be interpreted from the production at existing wells, that is, some kind of decline analysis.

The production variables have some spatial correlation that we can exploit; however, we must also exploit the information contained in the geological and geophysical variables that are related to the production variables we are trying to predict. These secondary data sources are also redundant with each other and we need to sort out the true information content in all data sources. All this information must be combined to build maps of what we expect the reservoir performance to be at undrilled locations. A summary of our prediction could take the form of maps of P_{10}, P_{50}, and P_{90} values of our production variables. The uncertainty and risk associated with new well locations could be assessed.

The result of data assembly is a set of variables that can be used to predict performance properties in the reservoir. These variables may include: 3-6 geophysical variables, 2-6 structural, 2-4 geological, and 2-4 production variables of interest that measure reservoir performance. The number of hard calibration data to establish the multivariate characteristics of these 10-30 variables may be few: the wells already drilled in the basin/pool under consideration. Conventional multivariate techniques would require 1000s or more data observations where all variables are present. This is simply not available in petroleum exploration and production.

We must also consider that the coverage area for each variable is different. It is important that all correlated secondary variables be considered. Geoscientists and engineers have been trained to work with data in such complicated settings. Expert judgment and interpretation is extraordinarily valuable. There is a need, however, to supplement such expert assessment with quantitative numerical tools that integrate all information accounting for the various interdependencies and to provide a measure of uncertainty in the predicted variable.

The methodology we develop below builds on very classical statistical and geostatistical tools for probabilistic prediction. A Bayesian approach is adopted whereby the secondary data are combined together to form a liklihood and the primary variables are mapped independently to form a prior distribution. These can be merged using Bayesian inference to arrive at posterior or updated probability distributions of the variables we are trying to predict.

Comments on Multivariate Statistics

The field of statistics provides a number of techniques that address the relationships between large sets of multiple variables. A set of *n* correlated variables can be transformed to be uncorrelated through techniques such as principal component analysis (PCA). PCA and other techniques such as factor analysis can be used to reduce the number of variables that must be considered. The variables can be non-linearly transformed to maximize linear correlation through techniques such as alternating conditional expectation (ACE). The data can be grouped together with techniques such as cluster analysis. There are a number of multivariate regression techniques for prediction of response variables considering multiple input variables. There are experimental design methods that aim at providing the best setup of test runs to understand multivariate statistical properties.

Most geoscientists and engineers would have difficulty choosing the best technique; it is often unclear which technique or sequence of techniques is appropriate for a particular problem. Moreover, as in any field, there are many fads and a new technique is sometimes thought of as a cure-all for any problem.

Another complicating factor is that multivariate techniques do not account for spatial relationships between the variables. Multivariate techniques were largely developed in the sciences where each observation of multiple variables can be thought of as independent of other observations. A central feature of reservoir characterization is spatial correlation in the underlying reservoir properties and the consequent production potential. We must integrate geostatistical measures of correlation in the multivariate statistical tools we choose/develop to account for the multiple secondary data.

A further complication is that not all variables are available at all locations. This unequal sampling requires us to locally change the parameters of our multivariate analysis.

Finally, there is a need to provide a measure of uncertainty in any prediction we make. There is often a high degree of uncertainty and risk associated with predicting performance characteristics at unsampled locations. A measure of this uncertainty is required to protect the geoscientist and engineer, that is, to avoid conveying a false sense of certainty in the predictions. A measure of uncertainty also provides management with some performance metrics that can be used to track the exploration/development program. The performance of new wells should be within the P_{10} and P_{90} values 80% of the time.

Comments on Geostatistics

There are a number of geostatistical techniques designed to work with multiple variables. These techniques account for the spatial relationships between the variables and provide a measure of uncertainty at every estimated location. The main technique is cokriging that can be applied in a multivariate Gaussian or an indicator framework. There are simplifying assumptions such as collocated cokriging and the Markov-Bayes approach. A concern with all these techniques is the inference of the direct and cross variogram measures of correlation, which requires a large number of data. They often require a total of $n(n+1)/2$ variogram models, which is extremely difficult in practice.

Collocated cokriging, in the Gaussian or Bayesian form, simplifies the process to consider only the collocated secondary variables. This also removes the need to model the large number of variograms mentioned above. There is some implementation problems associated with this simplification, but the method has proved very practical.

These geostatistical methods for considering multiple variables really only consider 1 to 3 secondary variables; there is no simple way to consider 10 to 30 secondary variables simultaneously. We must tailor the multivariate statistical and geostatistical tools to the problem of production performance prediction.

Proposed Approach

We will keep the notation to a minimum. We use *n* to represent the number of secondary variables; *m* is the number of production variables that we are estimating. We will develop a solution in the well established multivariate Gaussian framework. This requires each variable to be transformed to a univariate Gaussian distribution and, then, the parameters of the multivariate Gaussian distribution must be inferred. The univariate transformation is accomplished with the very classical normal scores transformation as implemented in the NSCORE program in GSLIB.

All secondary variables are merged into a single *likelihood* estimate at each location. Of course, the number of secondary variables available at each location could vary. The mean of the likelihood distribution is calculated as:

$$y_L^* = \sum_{i=1}^{n} \lambda_i \cdot y_i \qquad (1)$$

where *n* can be a subset or all secondary variables. The weights, λ_i $i=1,..,n$, are provided by the normal equations:

$$\sum_{j=1}^{n} \lambda_i \cdot \rho_{i,j} = \rho_{i,o}\,,\quad i = 1,....,n \qquad (2)$$

where $\rho_{i,j}$ is the correlation between the secondary variables and $\rho_{i,0}$ the correlation between the secondary variables and a primary variable. The *n* x *n* set of linear equations on the left hand side must be inverted and multiplied by the right hand side to solve for the weights. These weights are used to calculate the mean (equation 1 above) and the estimation variance:

$$\sigma_L^2 = 1 - \sum_{j=1}^{n} \lambda_i \cdot \rho_{i,o} \qquad (3)$$

At each location and for every primary variable, Equations (1) and (2) provide the mean and variance of a Gaussian likelihood distribution. These distributions are a collapsed version of all available secondary variables. The final likelihood distributions account for the relationships between the secondary variables and will be used to help inform the primary estimate.

The primary variables are predicted independently using kriging. For every location the kriging step provides an estimate, y_p^*, and variance, σ_P^2, that describes the prior distribution of the variable. The prior distribution will be Gaussian in shape. The kriging process accounts for spatial structure through the variogram model for each primary variable.

The likelihood and prior distributions are then combined to get the final updated distribution. Since the two input distributions are Gaussian in shape, the resulting updated distribution will be Gaussian. The updated distribution is defined by the updated mean:

$$y_U^* = \frac{y_L^* \cdot \sigma_P + y_P^* \cdot \sigma_L}{(1-\sigma_L)(\sigma_P - 1) + 1} \qquad (4)$$

and the updated variance:

$$\sigma_U^* = \frac{\sigma_P \cdot \sigma_L}{(1-\sigma_L)(\sigma_P - 1) + 1} \qquad (5)$$

Note that standard deviations are used and not variances in equation (5). The resulting updated distribution defined by (4) and (5) must be back-transformed to return the production variable to their original distributions. Any summary statistics of the local distributions can be calculated including the expected value, P_{10}, P_{50}, and P_{90} values. These summaries can be used to assist with land decisions, well placement, and reservoir development.

The proposed technique is referred to as **B**ayesian Updating under a **M**ultivariate **G**aussian model - or a **BMG** model for lack of a better acronym. The elements of this technique are not new; however, this is a novel way of putting everything together for reliable and simple estimation.

Limitations and Assumptions

Most practitioners will appreciate the simplicity of the proposed approach. More complicated procedures inevitably require additional parameters and greater risk of misapplication. We are accounting for all major aspects of the problem including redundancy between the secondary data variables, correlation to the production variables, spatial correlation of the production variables, and uncertainty in the prediction. No technique is without inherent assumptions and limitations. Here are the major ones for the proposed approach.

There is a strong assumption of *representative data*, that is, we assume that the statistical distributions of each parameter are fairly sampled with no systematic biases. Declustering techniques could be used to correct for minor sampling bias; however, the technique cannot correct for any systematic bias in the data. A systematic bias may be intentional (wells are supposed to be drilled in good areas) or unintentional (just bad luck). The data should be looked at carefully.

There is an implicit assumption of *spatial homogeneity* or *stationarity*, that is, that the statistical properties are the same across the entire study area. Gradational trends or abrupt changes in the depositional style will invalidate this assumption. Subdividing the study area and trend modeling may help, but there is always a point where we must group the data together for (geo)statistical inference.

A further assumption is that the data are multivariate Gaussian, that is, all relationships are linear, with constant variance, and with no abrupt constraints. Moreover, under the multivariate Gaussian model, all multivariate relationships are summarized by correlation coefficients. Inspecting each bivariate distribution for the reasonableness of this assumption is good practice. Of course, we really should check the trivariate and higher distributions for multivariate Gaussianity, but that is difficult in practice.

The estimates of production assume *no interaction between wells*, that is, the production at one location does not take away from the production at another location. Adding up all of the cumulative production estimates on the generated maps will lead to more volume than is present in the reservoir. Basic material balance calculations must be undertaken to supplement these calculations of local production. Moreover, some local flow simulations may be warranted to establish if the wells are indeed independent.

Uncertainty prediction is problematic because our estimates at any one location can always be wrong: there is a 10% probability to be below the predicted P_{10} value. The reasonableness of probability estimates must be checked over a number of outcomes. It would be a significant problem if 10 wells in a row were drilled below their predicted P_{10} values.

Example

This procedure has been applied on real reservoirs; however, the results are considered highly confidential. This synthetic example was created to demonstrate the updating process with two primary production variables using six variables as secondary data. The two production variables are initial production rate (IR) and total production (TP). The secondary variables consist of two geological, two structural, and two geophysical variables: sandstone indicator, reservoir quality, top elevation, formation thickness, impedance, and distance to a fault, respectively. These six variables were normalized and can be seen in Figure 1. All of the variables are inside of the 10,000m by 15,000m study area but the coverage changes between the variables. These synthetic variables were created using a combination of hand contouring, kriging, and cosimulation.

To create production data, 20 wells were randomly placed inside of the study area. Total production at each location was assigned randomly and the five highest locations had another two wells drilled in the surrounding area. The initial production rate was then assigned based on the total production and some random variables. This provided a total of 30 wells in the study area. The secondary variables were then sampled to obtain values at the well locations.

The primary and secondary data were complied into a master data sheet for the well locations. All of the variables were normal score transformed and the correlations were calculated (Figure 2). The correlation matrix indicates the variables with correlation above 0.2 and below -0.2 with dots. These correlations will be used to create the likelihood distributions.

A program was created that utilized the correlation matrix and available data to create likelihood distributions at every location for the two primary variables. These distributions are described by the likelihood estimate and variance maps shown in Figure 3. Note that the variance changes depending on the availability of the secondary variables.

Prior distributions must be created before the likelihood data can be applied. The two production variables were kriged separately with different variograms and only the well data. The kriged estimates and variances are seen in Figure 4. Note that the variance is zero at the data locations and increases as you move away.

The likelihood and prior distributions were then combined to create updated distributions at every location. The updated estimates and variances are seen in Figure 5. The updated maps show some interesting features. If both the likelihood and prior maps show high values in the same area for the estimate, the updated map will be even higher. The same situation will occur in the low value areas. Alternatively, if one map is high and the other map is low, the updated estimate will be in the middle. The updated variances are decreased at every location, except at the wells. The central area with the highest number of secondary data is reduced the most and the reduction is decreased as fewer data become available. These features come from the likelihood variances. The contribution from the prior variances is seen near the well locations. The effect is less noticeable compared to the prior maps due to the reduced variance everywhere.

The updated maps can be used as is, but it is difficult to interpret the estimate and variance maps at the same time. To make this process easier, maps were created for the two primary variables to show the P_{10}, P_{50}, and P_{90} at every location (Figures 6 and 7), respectively. To apply these maps you should start by looking at the P_{50} to look for areas you are interested in. If you are trying to identify poor production areas then the P_{90} map is used. Low values on this map are most likely low since there is a 90% chance that the value will be lower than the one shown. If an area is low on the P_{90}, it is highly likely to find low values in that area. If you are trying to identify high value areas, the P_{10} map is used. High value locations on this map are most likely high since there is a 90% chance the value will be higher than the one shown.

Conclusions

Bayesian updating under a multivariate Gaussian model provides a simple and robust solution to this inference problem. There are, of course, limitations and assumptions such as representative data, statistical homogeneity and multivariate Gaussianity.

Traditional geostatistics requires much professional time and is aimed at providing inputs to flow simulation. In many cases, this approach is to intensive; there is a need for a simpler method to directly predict reservoir performance. This technique has seen much practical application of late.

Acknowledgement

The authors would like to thank the industrial sponsors of the Centre for Computational Geostatistics for financial support of this research: http://www.uofaweb.ualberta.ca/ccg/.

NOMENCLATURE

n	=	number of data used in a calculation
y	=	normal score transform of a variable
σ	=	standard deviation
λ	=	weight calculated by normal equations
ρ	=	correlation coefficient
subscripts		
P	=	prior distribution from same data type
L	=	likelihood distribution from same secondary
U	=	updated distribution
i,j	=	data indicies
0	=	index for location being estimated
superscripts		
*	=	estimate from available data

REFERENCES

1. DEUTSCH, CV, Geostatistical Reservoir Modeling, Oxford University Press, New York, 376 pages, 2002.
2. DEUTSCH, CV and JOURNEL AG, GSLIB: Geostatistical Software Library and User's Guide, Oxford University Press, New York, second edition, 376 pages, 1998.
3. DOYEN, PM., DEN BOER, LD, PILLEY, WR, Seismic porosity mapping in the Ekofisk field using a new form of collocated cokriging. *Society of Petroleum Engineers*, SPE 36498, 1996.
4. XU, W, TRAN, TT, SRIVASTAVA, RM, and JOURNEL, AG, Integrating seismic data in reservoir modeling: The collocated cokriging alternative. *Society of Petroleum Engineers*, SPE 24742, 1992.

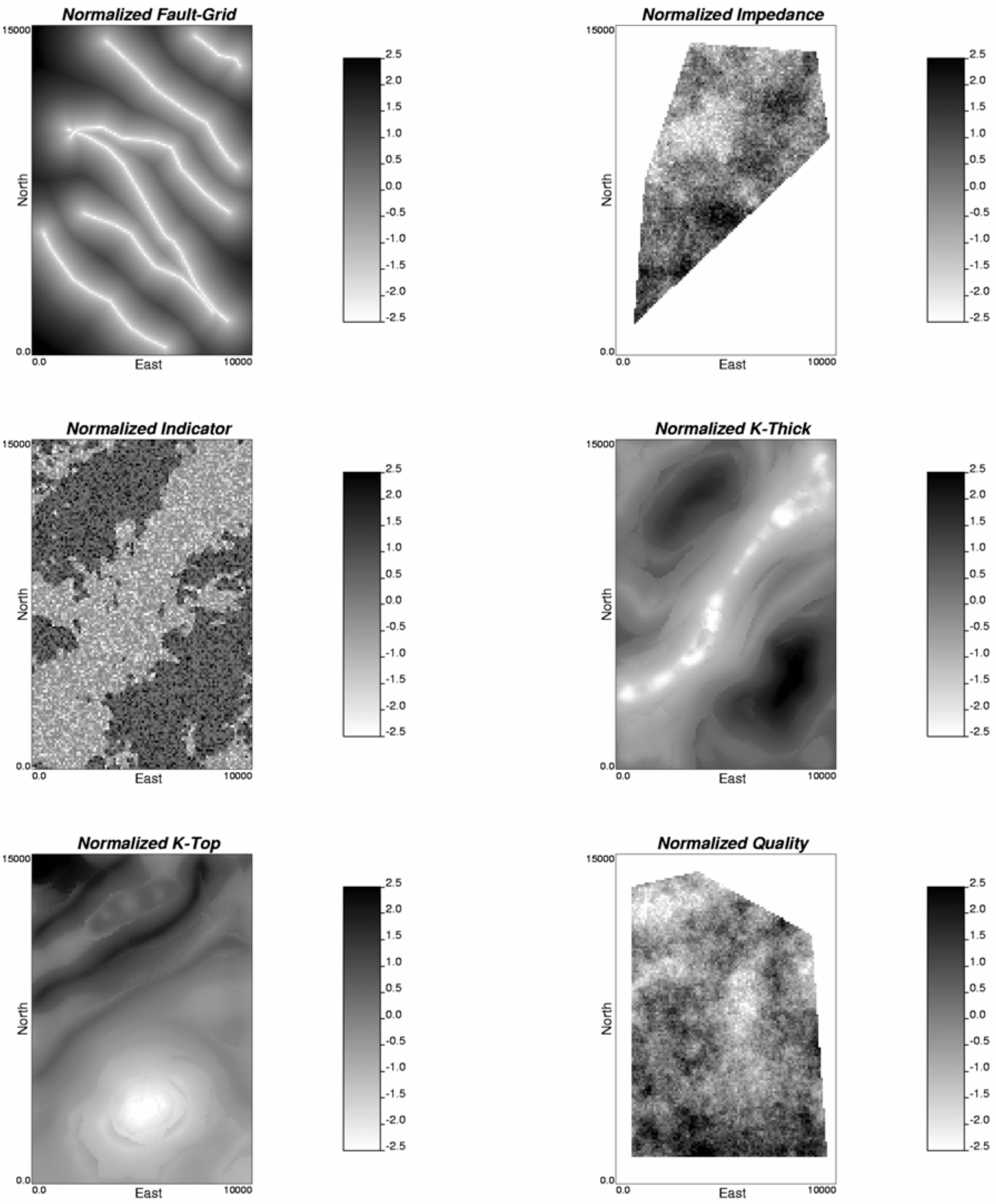


Figure 1: Maps of the six normalized variables used for the example of direct prediction of production characteristics.

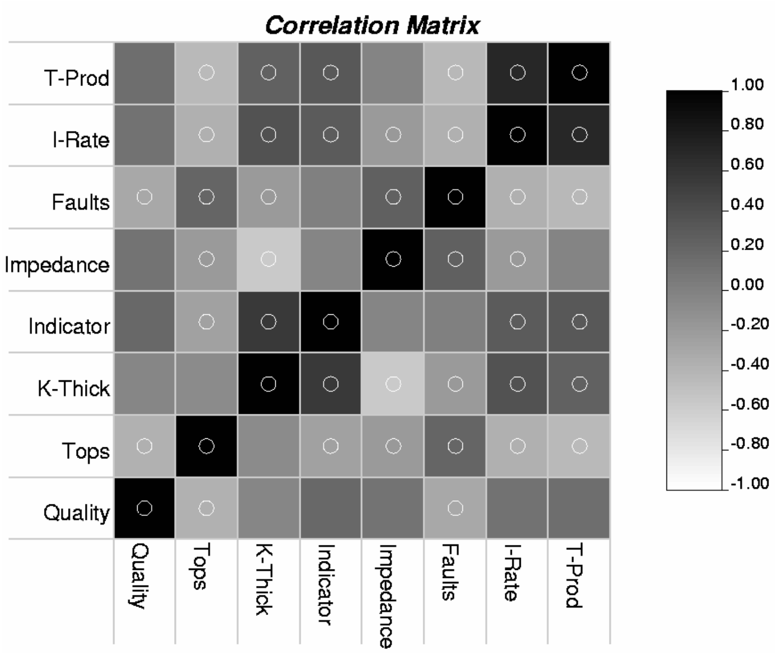


Figure 2: A gray scale map of the correlation coefficients between six secondary variables and two production variables being predicted in the example. The circles represent correlation coefficients that have an absolute value above 0.2.

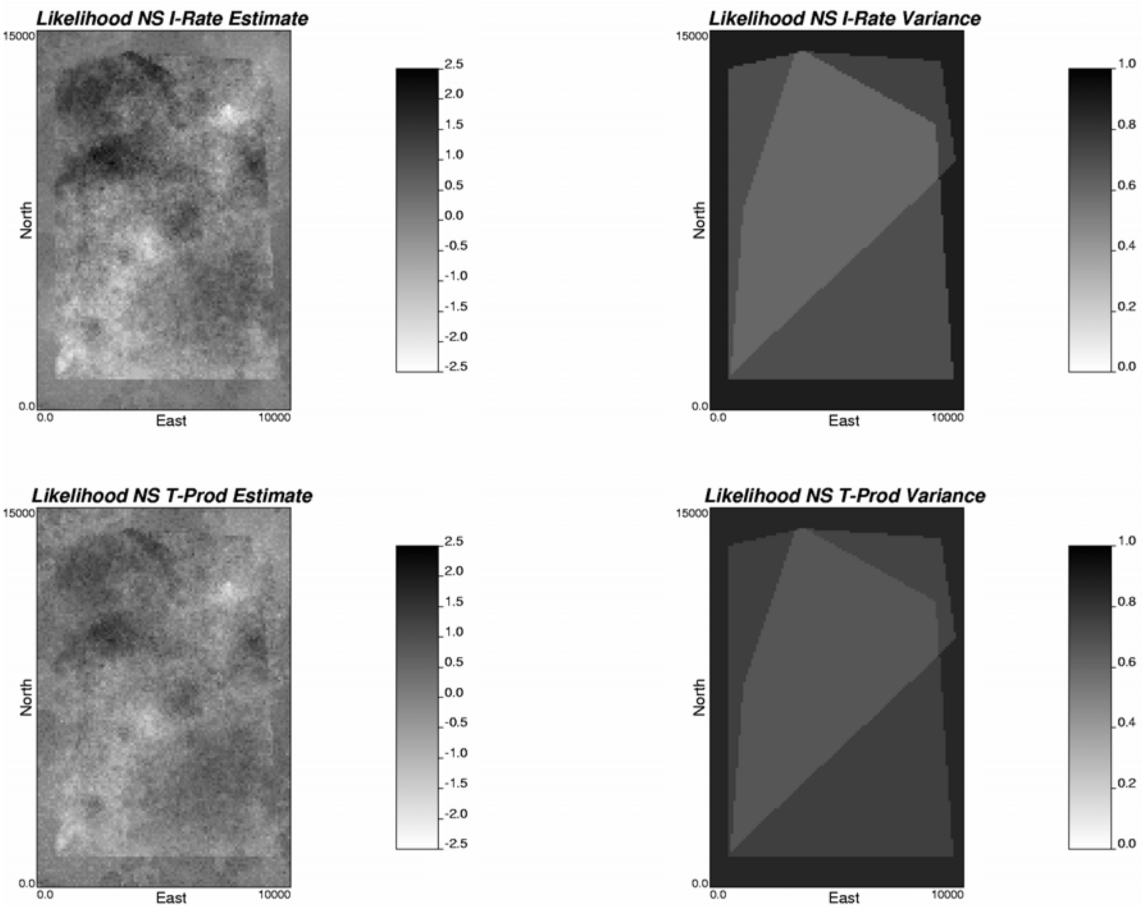


Figure 3: The likelihood maps for the initial production rate (top) and total production (bottom) in normal space. The left hand side maps show the estimate of the production variables based on all available secondary data. The right hand side maps is the corresponding variance at every location.

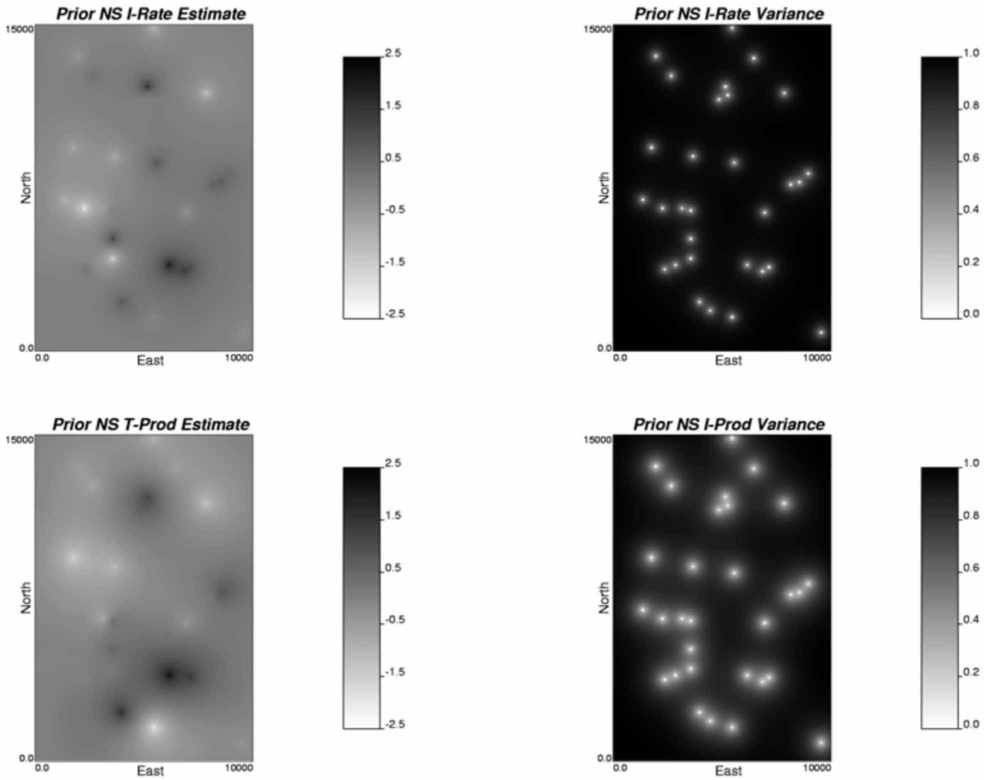


Figure 4: The prior distribution maps for the initial production rate (top) and total production (bottom) in normal space. The left hand side are the estimates and the right hand side are the corresponding variances.

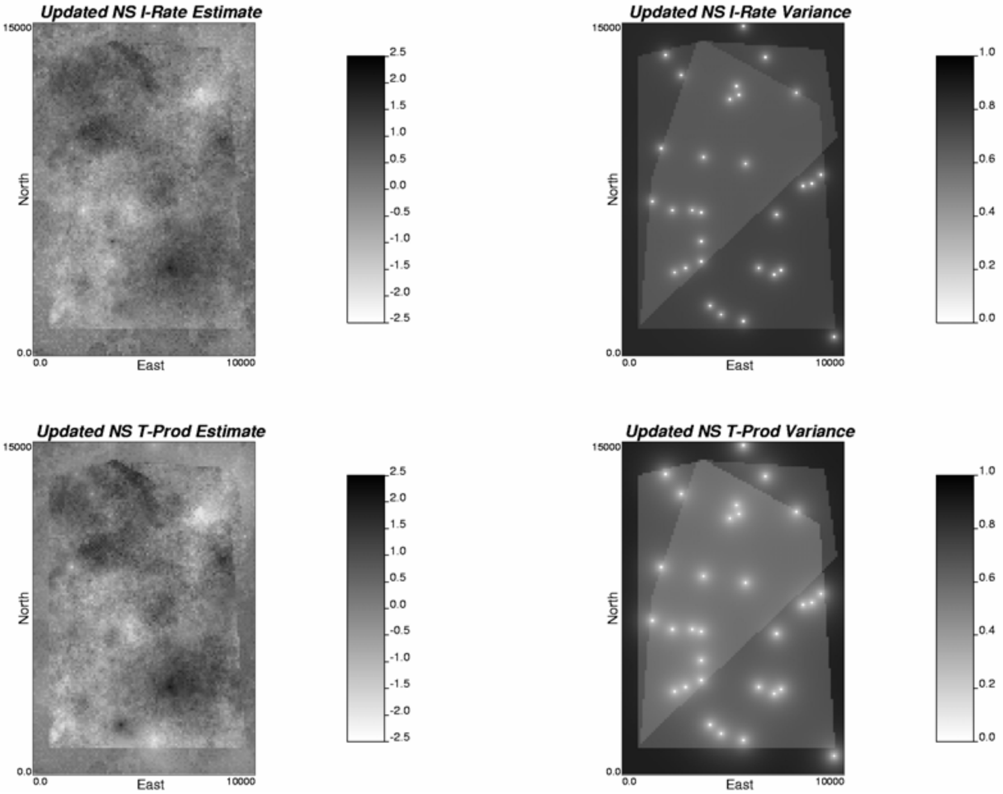
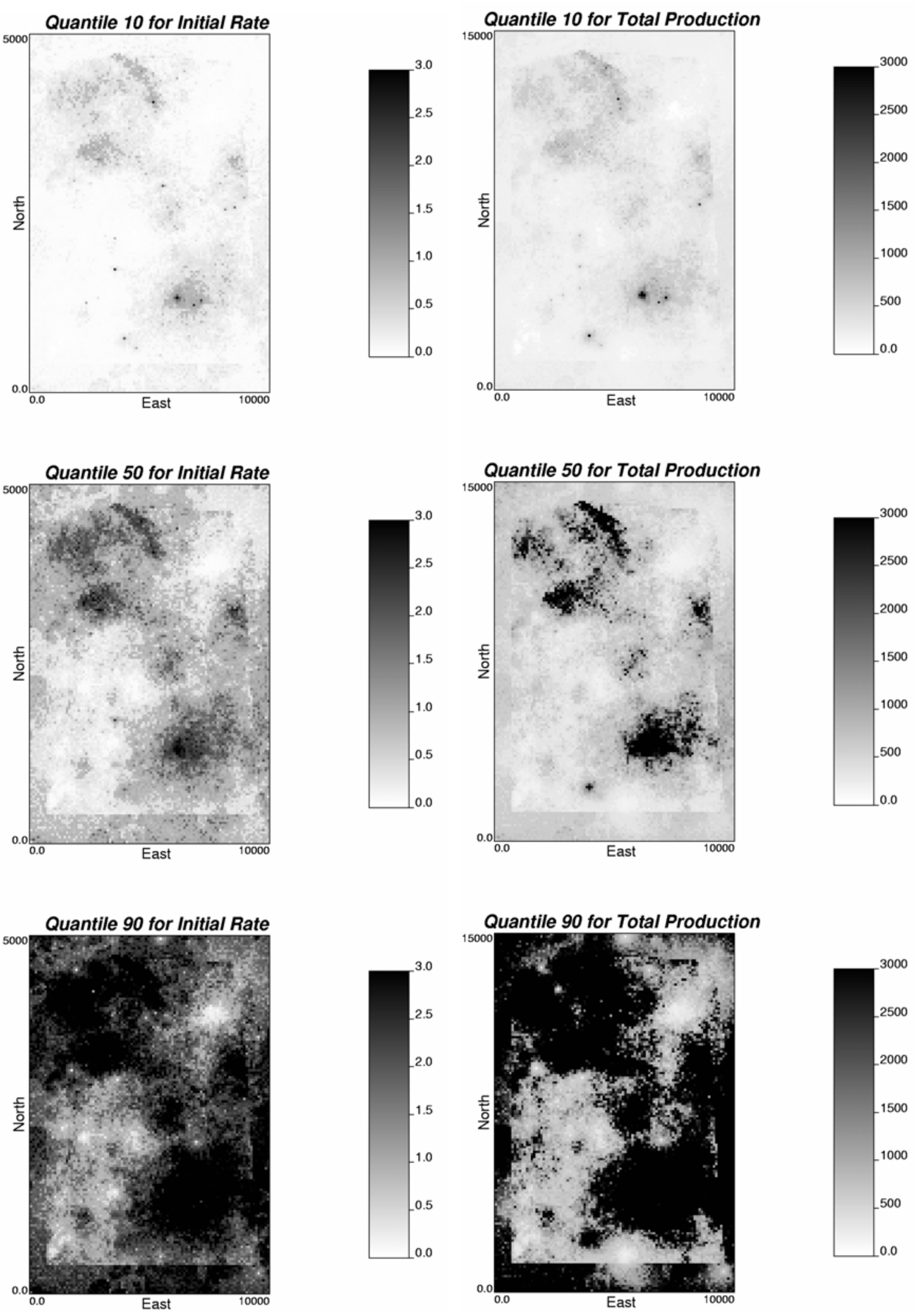

Updated NS I-Rate Estimate
Updated NS I-Rate Variance
Updated NS T-Prod Estimate
Updated NS T-Prod Variance
North
East
15000
0.0
10000
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
1.0
0.8
0.6
0.4
0.2
0.0

Figure 5: The updated maps of the initial production rates (top) and total production (bottom) in normal space. The left hand side are the estimates and the right hand side are the corresponding variances. Notice how features from the prior and likelihood maps can be seen.


Quantile 10 for Initial Rate

Quantile 10 for Total Production

Quantile 50 for Initial Rate

Quantile 50 for Total Production

Quantile 90 for Initial Rate

Quantile 90 for Total Production

Figure 5: The p10 (top), p50 (middle), and p90 (bottom) quantile plots for the initial production rate (left) and total production (right).