The use of predictive analytics for hydrocarbon exploration in the Denver-Julesburg Basin

Emmanuel T. Schnetzler[1] and David L. Alumbaugh[1]

Abstract

We present an approach to predict spatial distribution of a variable from a set of geophysical and interpreted grids using alternating conditional expectations (ACE). This technique is based on nonparametric transformations of the predictor and response variables in order to maximize the linear correlation of the transformed predictors with the transformed response. ACE provides a powerful method to detect underlying relationships between the variables and use them in a regression framework to predict the response variable. A case study is presented that illustrates the approach using a set of grids derived from geophysical attributes (gravity, magnetic, electromagnetic) and interpreted grids (isopach, total organic carbon) as predictor variables to estimate early hydrocarbon production.

Introduction

Data-driven approaches are being applied increasingly to geologic settings. This is particularly the case in a multivariate setting when multiple predictor variables have complex and often nonlinear relationships with the response variables. As opposed to "black-box" types of machine learning algorithms that can be difficult to interpret, alternating conditional expectations (ACE) provides more insight into the underlying relationships present in the data.

Multiple linear regression

In the classical multiple linear regression approach, the response variable y is modeled as a linear combination of the predictor variables $x_1, x_2, \ldots, x_p$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p . \qquad (1)$$

This imposes a strong assumption of linearity that is not appropriate when complex relationships exist.

In some cases, nonlinear relationships are present and known, in which case parametric transformations of the variables can be applied to build a model. This is the case in the Box-Cox family of transformations for continuous variables (Box and Cox, 1964). However, this also imposes a strong preconceived model on the data that is not valid in complex cases.

Alternating conditional expectations

Theory. A number of nonparametric regression techniques have been applied successfully when complex unknown relationships are present between predictor variables and a response variable (Kuhn and Johnson, 2013). In this family of techniques, the relationships between predictor and response variables are built from the data via optimal transformations. In some, only the predictor variables are transformed (generalized additive models). In others, both predictor and response variables are transformed. This is the case of the ACE technique (Breiman and Friedman, 1985). The equation takes the form:

$$\theta(y) = \varphi_1(x_1) + \varphi_2(x_2) + \ldots + \varphi_p(x_p) = \sum\nolimits_{i=1}^{p} \varphi_i(x_i) . \qquad (2)$$

Each variable (predictor variables x_1, x_2, … x_p, and response variable y) is transformed with optimal transformations φ_1, φ_2, … φ_p, and θ. The transformations are derived from the data using an iterative process aimed at maximizing the linear correlation between transformed predictor variables and transformed response variable. The approach works in an alternating fashion to minimize the error variance e^2 in equation 3 with respect to one function while keeping the other functions constant.

$$e^2 = \mathrm{E}\left\{\left[\theta(y) - \sum\nolimits_{i=1}^{p} \varphi_i(x_i)\right]^2\right\} . \qquad (3)$$

Some constraints can be imposed on the shape of the transformations, such as monotone or linear. The approach also allows the use of categorical variables as predictors.

In the prediction step, the transformations φ_i are applied to the predictor variables at locations to be estimated. The sum of transformed values is back transformed (transformation θ) to estimate the value of the response variable. The transformations are applied directly, without having to model a function as in parametric approaches or applying smoothing as in generalized additive models.

ACE has been shown to be able to model complex nonlinear relationships as illustrated in Wang and Murphy (2004). Barnett and Deutsch (2013) looked at an application to a geometallurgical example in a nickel laterite deposit where Ni was predicted from five predictor variables: Fe, SiO_2, MgO, Co, and Al_2O_3. The application shows good results from the approach at finding underlying relationships, although the spatial distribution was not presented.

Application to spatial data. Regression techniques can be applied in a spatial setting: down a borehole in 1D, over a 2D area, or in a full 3D setting. In a 2D case, one can consider predicting a response variable away from data locations where the variable is known (typically wells), using a set of exhaustive data sets covering the area of interest.

The general workflow illustrated in Figure 1 follows the steps:

- select the response variable to be predicted and gather a consistent data set (training points);
- select a set of relevant exhaustive predictor variables on a 2D grid (data layers) and extract values of the predictor variables at the response variable locations (wells);
- build optimal transformations for response and predictor variables from values at the training locations; and
- apply the transformations to the full 2D grid to predict the response variable over the area.

[1] NEOS.

http://dx.doi.org/10.1190/tle36030227.1.

Application to the Denver-Julesburg Basin

Overview. ACE is applied to a real case in an area of 3000 square miles located in the western part of the Denver-Julesburg Basin in Colorado. Major oil-producing areas shown on Figure 2b in the Niobrara formation include the Hereford in the northwest, the East Pony in the northeast, and the Wattenberg field in the center.

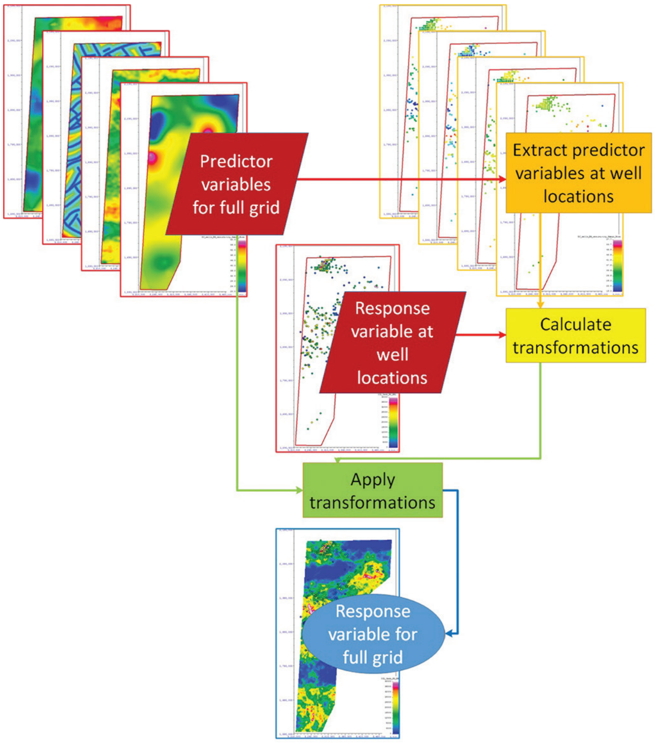

Predictor variables for full grid
Extract predictor variables at well locations
Response variable at well locations
Calculate transformations
Apply transformations
Response variable for full grid

Figure 1. Workflow for ACE application to spatial data.

Input data. The available data include production information and a set of geophysically and geologically derived data "layers."

Response variable. The response variable selected is the early cumulative oil production, calculated as the sum of the production of the best three months in the first year of production. Because historical production includes wells drilled over several decades using different technologies, it is necessary to subset the data to a consistent set of wells; in this case, only horizontal wells were used. This controls for part of the variability between wells, although some further filtering could control for additional variables, in particular engineering factors. Exploratory data analysis shows that in this case the subset of horizontal wells is sufficiently homogeneous. Figure 3 illustrates the spatial distributions of horizontal wells with corresponding early production in barrels of oil (BBL).

Predictor variables. A set of nine predictor variables available for the full area of interest and deemed related to productivity is compiled. Figure 4 shows an overview of the data layers. In the top row, left to right, the figure shows the Precambrian basement depth, the distance to interpreted faults, and the average Niobrara resistivity as interpolated from well logs. The middle row shows the Bouguer gravity anomaly, the reduced-to-pole (RTP) magnetic field, and the Niobrara isopach. In the bottom row are displayed the total organic carbon (TOC), the temperature at which the maximum rate of hydrocarbon generation occurs in a kerogen sample during pyrolysis analysis (T_{max}), and the vitrinite reflectance. Below we explain how the different layers were derived or where they were obtained, and why we think these layers are important for our analysis.

Of the nine different data layers, four directly involved the use of the airborne potential field data collected during the survey. The first two of these are the processed gravity and magnetic data

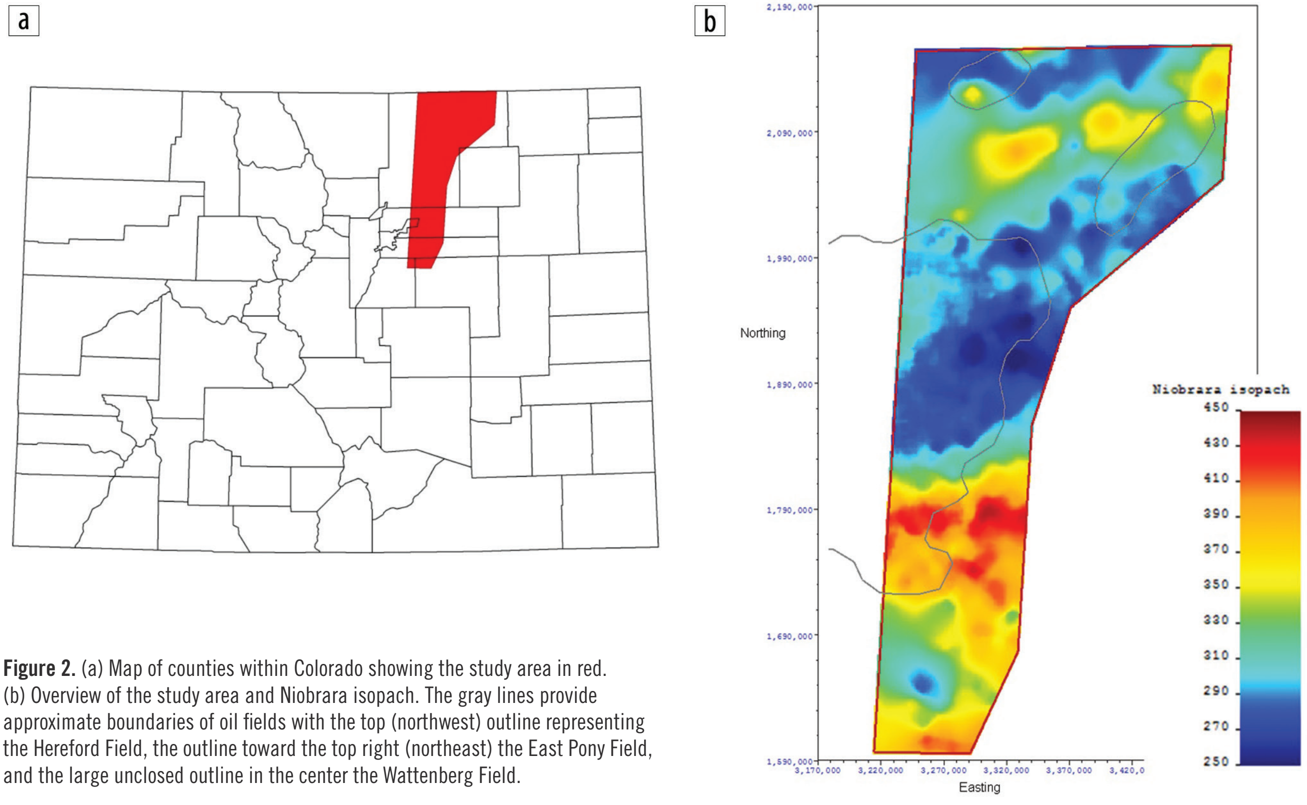


Figure 2. (a) Map of counties within Colorado showing the study area in red. (b) Overview of the study area and Niobrara isopach. The gray lines provide approximate boundaries of oil fields with the top (northwest) outline representing the Hereford Field, the outline toward the top right (northeast) the East Pony Field, and the large unclosed outline in the center the Wattenberg Field.

(the first two plots of the middle row in Figure 4). These data layers are employed to reflect the impact that basin structure and composition, and the presence of basement intrusions, might have on production. Sander Geophysics Limited acquired the gravity data for NEOS. For the predictive analytics we employed the complete Bouguer anomaly processed using a reference density of 2.60 g/cm^3. Terraquest LTD collected the magnetic data. Several different magnetic field attributes were calculated and inspected for use in the interpretation. In the end, it was decided to use the tilt derivative of the reduced-to-pole (RTP) magnetic anomaly as this derivative was thought to enhance the signature of faults.

The depth-to-basement map shown as the first figure in the upper row of Figure 4 was derived from the gravity and magnetic data through a 3D constrained modeling and inversion analysis in which constraints were provided via well tops in wells that penetrated the basement. First, a 3D model consisting of seven layers was constructed, and several inversions for the density with the layers were completed. Note that during the inversion process, no lateral density changes were allowed in each of the layers, and the minimum and maximum density allowed in each layer were tightly bounded. Next a structural inversion was completed where the layer depths were allowed to vary locally by a small amount. The last step was to allow the basement density to change laterally. The final result shown here matches the measured basement depth at the well locations within ±50ft.

The last data layer derived from the geophysical data was a distance-to-faults map, which was incorporated to reflect the impact that faulting can have on production. Faulting can lead to enhanced fracturing within the Niobrara and thus better production (e.g., Sonnenberg and Underwood, 2012), and it also can provide enhanced fluid and heat flow up along faults, thus affecting the thermal maturation of the source rock. The map was derived by first generating an integrated fault map from interpretation of the tilt derivatives of both the gravity and magnetic data. These interpreted faults were then combined with the fault map outlined in Weimer (1996). The distance to faults was calculated as the radial distance from any point in the survey area to the closest mapped fault.

Two of the data layers, the Niobrara resistivity and isopach, were constructed from the interpolation of well-log data. In the case of the resistivity, this involved first calculating the harmonic mean of resistivity logs collected through the Niobrara formation and then interpolating this mean value between the wells. The isopach map was constructed in a similar manner by using well logs to pick the top and bottom of the Niobrara from which the thickness could be calculated and then interpolating to form a map.

The remaining three data layers input to the prediction workflow (last row in Figure 4) were pulled from the literature and provide valuable information about the amount of carbon and the thermal history of the source rock in the Niobrara. The TOC of the Niobrara (Curtis et al., 2012) provides information about the source-rock quality. The two other maps provide estimates of thermal maturity of the source rock. The T_{max} map (Thul and Sonnenberg, 2013) indicates the maximum temperature that the Niobrara has been subjected to, while the vitrinite reflectance map (Sonnenberg, 2011) provides a second measure of thermal maturity of the organics within the Niobrara formation across the study area.

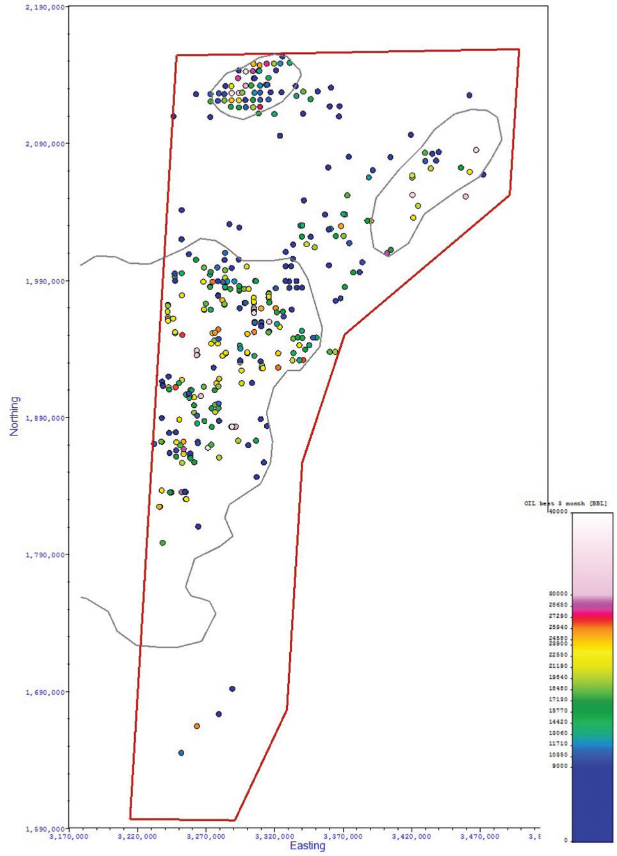


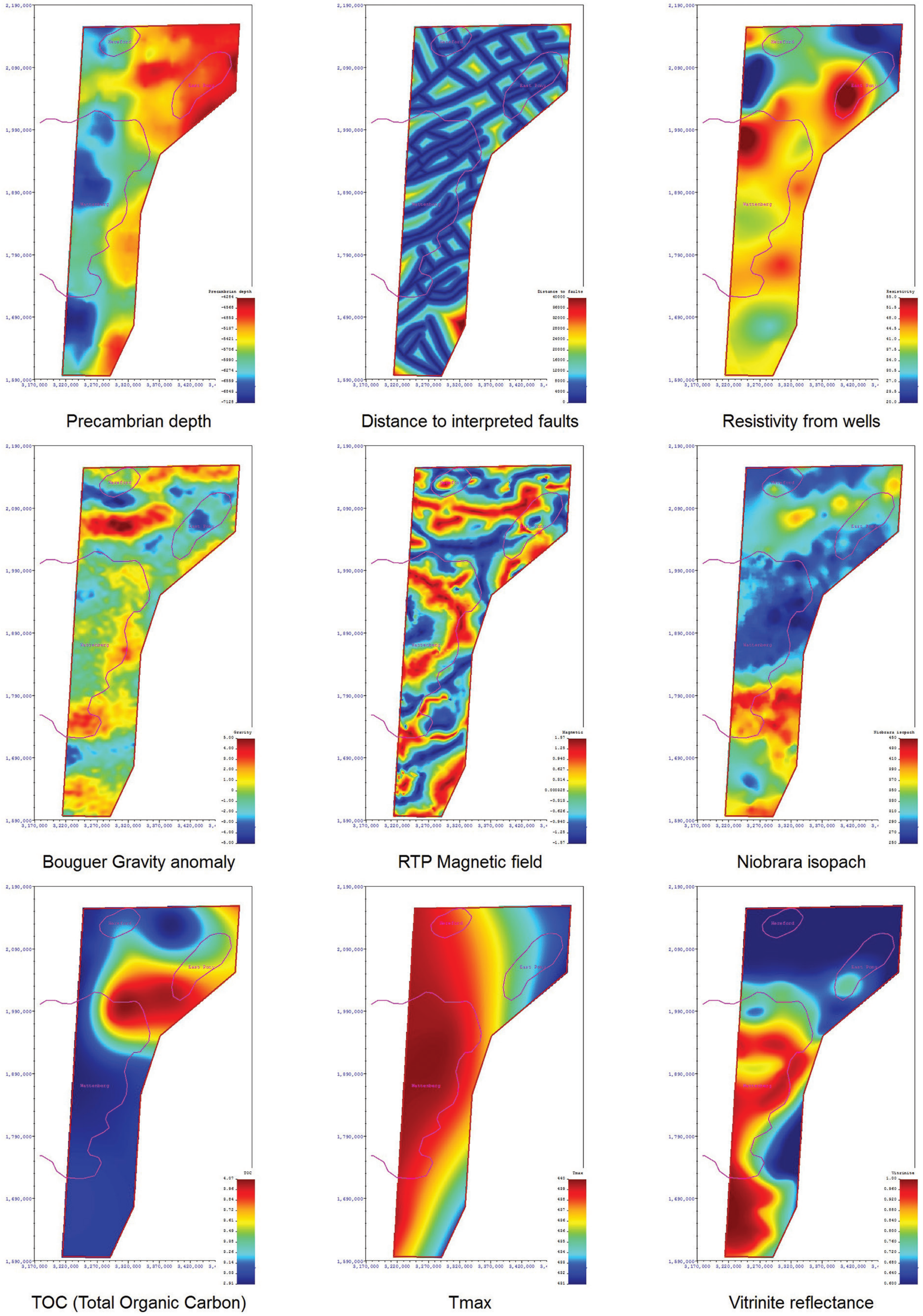
Figure 3. The response variable; spatial distribution of early oil production (BBL).

ACE result

The result of the prediction (early oil prediction in BBL) from the alternating conditional expectations is shown in Figure 5. The three major oil-producing areas are highlighted in the prediction, in part driven by the data. Additional areas with limited data control are predicted as high producing locations, in particular in the northeast, south of the East Pony field, and in the south where only a few training data points are available.

Cross validation is performed where 90% of the data is used as training data for prediction, and the remaining 10% is used for validation. The 40 data points kept for validation are selected at random. Figure 6 shows a crossplot between real and predicted production values for the validation set of 40 points. It shows a good predicting power for the approach with a correlation of 0.67. The average of the predicted production values (18,815 BBL) is very close to the average of the true production values (18,867 BBL). However, for the validation set, the prediction seems to underestimate slightly the high values of production.

The crossplot between the sum of the transformed predicted variables (right term in equation 1) and the transformed response variable (production) is shown on Figure 7. The correlation coefficient is 0.74; this is the measure that the ACE approach aims



Precambrian depth

Distance to interpreted faults

Resistivity from wells

Bouguer Gravity anomaly

RTP Magnetic field

Niobrara isopach

TOC (Total Organic Carbon)

Tmax

Vitrinite reflectance

Figure 4. The predictor variables; data layers input to the ACE run as listed in the text.

to maximize through the calculation of the optimal transformations of the predictor and response variables.

For each predictor variable, the value it lends to the regression can be calculated. Predictor variables can be ranked according to their values, as shown in Table 1, with decreasing value from top to bottom. T_{max} and Precambrian depth contribute significantly to the regression. The following six variables show about the same lower contribution, while distance to faults has a very low contribution and could be removed from the regression without significantly affecting the prediction.

The optimal transformations can be plotted and analyzed to understand the effect of each variable on the response variable. Figure 8 shows the calculated optimal transformation of one of the predictor variables (Precambrian depth) against the original values. This plot illustrates the effect of the variable on the response variable: the relatively flat first part of the graph between a depth of -7000 m and -6000 m indicates that a change of depth within that interval does not affect the production significantly. Between -6000 m and -4500 m however, the relationship between Precambrian depth and production is roughly linear.

The fluctuations between -7000 m and -6000 m visible on the transformation curve is likely indicative of overfitting that is common in machine learning methods and can warrant more detailed analysis.

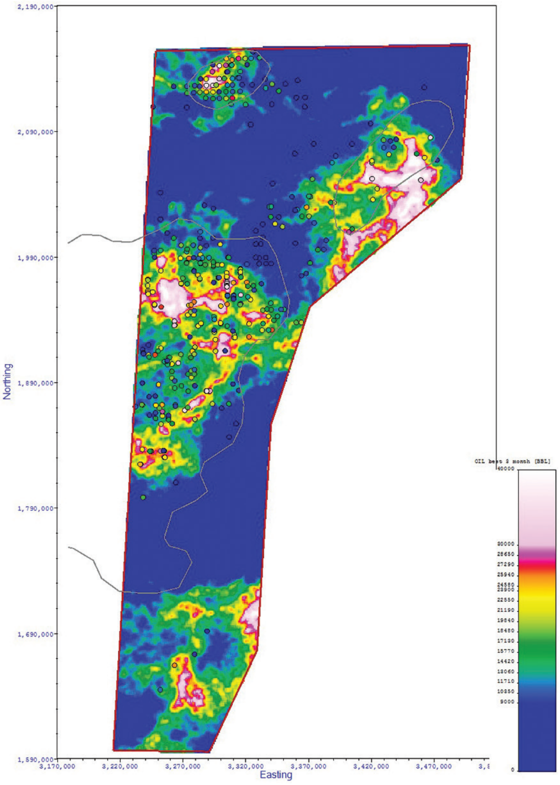


Figure 5. Production prediction output of the ACE run.

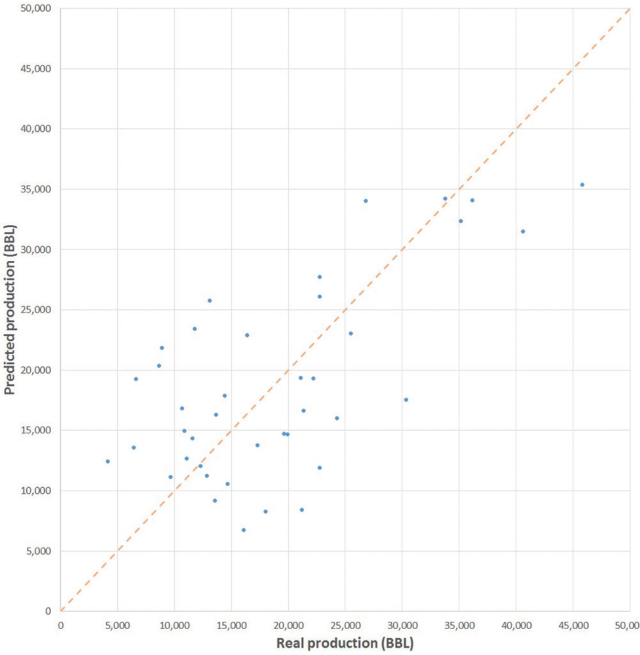


Figure 6. Real versus predicted production for the validation set of 10% of the complete data set.

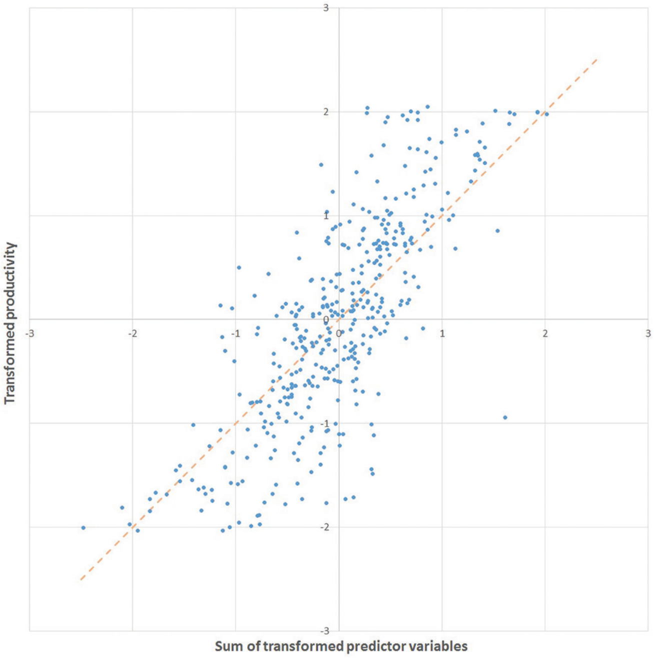


Figure 7. Crossplot between the sum of the transformed predictor variables and the transformed response variable (production).

Table 1. Ranked value that each predictor lends to the regression.

Variable	Value
T_{max}	0.20
Precambrian depth	0.18
Gravity	0.12
Resistivity	0.11
Vitrinite	0.10
Niobrara isopach	0.09
TOC	0.08
Magnetic	0.08
Distance to faults	0.04

Time-step analysis

As an additional validation of the procedure, a time-step analysis is performed. Starting from the current date, we move back six months at a time and perform the prediction with the data set that would have been available at each date step. Figure 9 shows the succession of three time steps and the final run with the full data set available. While the details change between the versions, the general patterns start emerging with the data available a year and a half early, with only half the number of wells drilled.

Conclusions

Many regression techniques can be used to predict a variable from a set of predictor variables. When relationships are complex and cannot be modeled parametrically either through linear model or more complex functions, a nonparametric approach is more appropriate.

Machine learning algorithms (support vector machines, neural networks, etc.) are powerful but tend to be difficult to interpret. Nonlinear, nonparametric approaches based on optimal transformations of the variables such as alternating conditional expectations and some modifications designed to address some limitations, additivity and variance stabilization (AVAS) for example (Tibshirani, 1988), provide the opportunity to inspect the optimal transformations and put them in a physical context for validation.

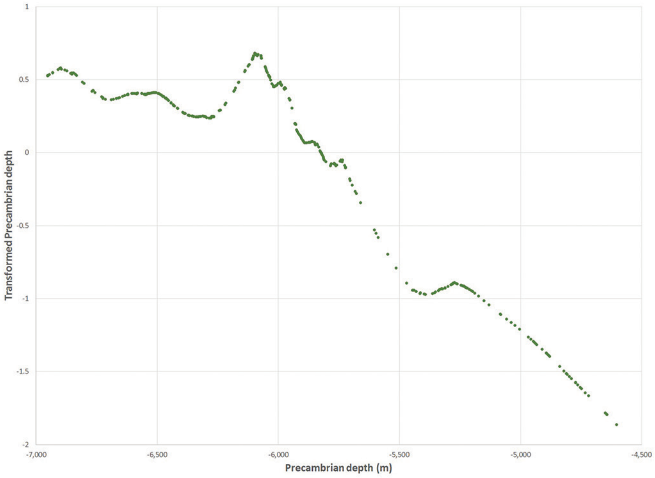


Figure 8. Optimal transformation of the Precambrian depth.

A number of potential pitfalls need to be kept in mind when applying any regression technique. Overfitting the training data is a common problem in predictive modeling that should be watched closely through cross validation. The predictor variables should be chosen with care, and it should be possible to formulate a reason why each is likely to be related to the response variable, even if the link cannot be defined expressly. The method relies on colocated relationships and does not take into account relations to neighbors, such as is the case in cokriging or cosimulation.

The example presented shows how ACE can be applied in a spatial setting to predict a grid of early oil production. The case study illustrates the improvement in correlation between transformed predictors and transformed response. Other variables such as petrophysical properties (e.g., porosity) could be considered for this approach, and while the case study is 2D, it is applicable in 3D. TLE

Acknowledgments

We would like to thank NEOS for their permission to present this work as well as the reviewers for their constructive comments on the paper. The Center for Computational Geostatistics at the University of Alberta in Edmonton, Canada, provided source code for the alternating conditional expectations algorithm.

Corresponding author: eschnetzler@neosgeo.com

References

Barnett, R. M., and C. V. Deutsch, 2013, Tutorial and tools for ACE regression and transformation: Centre for Computational Geostatistics (CCG) Annual Report, **15**, 401: University of Alberta.

Box, E. G., and R. D. Cox, 1964, An analysis of transformations: Journal of the Royal Statistical Society. Series A (General), **26**, no. 2, 211–252.

Breiman, L., and J. H. Friedman, 1985, Estimating optimal transformations for multiple regression and correlation: Journal of the American Statistical Association, **80**, no. 391, 580–598, http://dx.doi.org/10.1080/01621459.1985.10478157.

Curtis, J. B., J. Zumberge, and S. Brown, 2012, Evaluation of Niobrara and Mowry Formation petroleum systems in the Powder

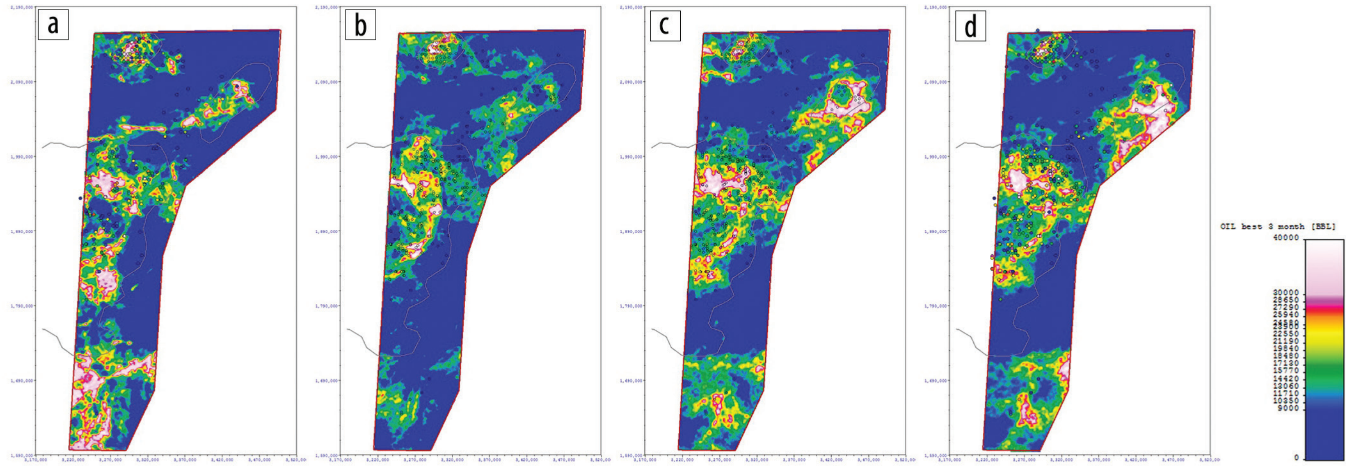


Figure 9. Simulated time-step production prediction with ACE. ACE results calculated with available production data (a) 1.5 years before the current time, (b) 1 year before the current time, (c) 0.5 years before the current time, and (d) the current time.

River, Denver and Central Basins of the Rocky Mountains, Colorado and Wyoming, USA: Presented at AAPG Annual Convention and Exhibition.

Kuhn, M., and K. Johnson, 2013, Applied predictive modeling: Springer.

Sonnenberg, S. A., 2011, The Niobrara petroleum system, a major tight resource play in the Rocky Mountain region: AAPG Search and Discovery Article #10355.

Sonnenberg, S. A., and D. Underwood, 2012, Polygonal fault systems: a new structural style for the Niobrara Formation, Denver Basin, CO: AAPG Search and Discovery Article #50624.

Thul, D. J., and S. Sonnenberg, 2013, Niobrara source rock maturity in the Denver Basin: a study of differential heating and tectonics on petroleum prospectivity using programmed pyrolysis: AAPG Search and Discovery Article #80341.

Tibshirani, R., 1988, Estimating transformations for regression via additivity and variance stabilization: Journal of the American Statistical Association, **83**, no. 402, 394–405, http://dx.doi.org/10.1080/01621459.1988.10478610.

Wang, D., and M. Murphy, 2004, Estimating optimal transformations for multiple regression using the ACE algorithm: Journal of Data Science: **2**, 329–346.

Weimer, R. J., 1996, Guide to the petroleum geology and Laramide orogeny, Denver Basin and Front Range, Colorado: Colorado Geological Survey Bulletin 51.