Evaluation of Metamodeling Techniques on a CO2 Injection Simulation Study

Jared Schuetter  
*Battelle*, schuetterj@battelle.org

Srikanta Mishra  
*Battelle*, mishras@battelle.org

Doug Mooney  
*Battelle*, mooneyd@battelle.org

Follow this and additional works at: https://scholarsarchive.byu.edu/iemssconference

Part of the Civil Engineering Commons, Data Storage Systems Commons, Environmental Engineering Commons, Hydraulic Engineering Commons, and the Other Civil and Environmental Engineering Commons

https://scholarsarchive.byu.edu/iemssconference/2014/Stream-A/23

This Event is brought to you for free and open access by the Civil and Environmental Engineering at BYU ScholarsArchive. It has been accepted for inclusion in International Congress on Environmental Modelling and Software by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
Evaluation of Metamodeling Techniques on a CO$_2$ Injection Simulation Study

Jared Schuetter$^a$, Srikanta Mishra$^a$, and Doug Mooney$^a$

$^a$Battelle, 505 King Ave., Columbus, OH 43201 USA (schuetterj@battelle.org, mishras@battelle.org, mooneyd@battelle.org)

Abstract: Full-physics models for geologic carbon sequestration applications can be time-consuming to run. As a result, optimization of a particular response over a range of predictor values can be difficult. A common approach is to use a small number of simulation runs to develop a metamodel that can approximate the system response with much less computing time; this metamodel can then be used for optimization. Choosing a metamodeling approach is not always straightforward, and it can also be difficult to determine how well a model is fitting after it is trained. Here, we present a case study for a CO$_2$ injection problem in the ARCHES province in the American Midwest, wherein a cross-validation approach was used to evaluate the quality of several metamodel fits.

To begin, the STOMP-CO$_2$ simulator was used to generate 36 simulation runs by varying the values of three predictors. The modeling of CO$_2$ injection into a closed volume results in three responses describing the amount and extent of the CO$_2$ stored in the model domain, as well as the pressure at the injection site. These 36 runs were used to train several metamodels, which included quadratic polynomial regression, kriging, MARS, AVAS, and TPS. Cross-validation was then used to evaluate the root mean squared error (RMSE) of these metamodels over the design space. In all cases, the kriging model had the most favorable RMSE.

Keywords: Carbon Sequestration; Metamodel; Proxy Model; Response Surface Model

1 INTRODUCTION

1.1 Background

To understand the behavior of a response function with respect to multiple predictor values, one typically needs a large number of observations to adequately cover the input space. A naïve approach is to compute the response for all combinations of predictor values chosen on a suitably fine grid. Usually, this is not feasible. In physical experiments, some combinations of predictors may not be available to the experimenter, or may produce responses that are beyond the capability of the instrumentation to measure. In simulated experiments (e.g., finite element models), a large amount of computation may be required to collect each response. Therefore, computing responses over a grid of predictor values may take too long, or be too expensive to complete.

The standard method for avoiding costly data collection is to only observe the response at a subset of predictor values, and then fit a metamodel (also called a proxy model or response surface model) to those points. Metamodels approximate the response at unobserved combinations of predictor values using the available sampled data, and are typically designed for rapid prediction. In this way, an approximate response surface can be generated for the entire input space in a short amount of time, and it can be subsequently used to meet project-specific research goals.

In the oil and gas literature, metamodels are often used as proxies for the underlying simulation mod-

There are also examples of specific case studies in which metamodeling was used. Kalla and White [2005] compared a second order polynomial model and kriging model using an orthogonal array (OA) sample design in a gas coning case study. In this case, the second order polynomial outperformed kriging with a 36-run design in 14 variables. Anbar [2010] settled on first order polynomial models for fitting outputs of a CMG STARS simulation for CO₂ sequestration in deep saline carbonate aquifers. The models were fit using LHS designs of size 100 over 16 variables. Finally, Wriedt et al. [2014] used a Box-Behnken design and a stepwise quadratic regression model to develop probability distributions for responses related to CO₂ injection into deep saline reservoirs.

1.2 Data Description

The analysis described in this paper used data from a CO₂ injection simulation for the ARCHES province in the Midwest, which is described in Mishra et al. [2014]. In that study, single-well simulations of CO₂ injection into a closed volume (as would be the case in a network of wells employed for regional-scale CO₂ storage) were carried out using STOMP-CO₂ [White et al., 2012]. Stratigraphic columns corresponding to three different ratios of reservoir (Mount Simon sandstone, MS) and caprock (Eau Claire shale, EC) thickness were considered, with different depths to injection zone in each case.

For each synthetic site case, simulations with 4 different well patterns (3x3, 4x4, 5x5 and 6x6 well arrays), and 3 different permeability group variations (High $k$, Medium $k$ and Low $k$) were run. Each permeability group consists of a set of correlated variables: permeability of MS, permeability of EC, and the capillary entry pressure. This brings the total number of simulations to $3 \times 4 \times 3 = 36$. Simulations of pressure-constrained injection (at 90% of the fracture pressure) were carried out using a 2-D r-z model with 20 vertical rows x 100 radial columns. The predictors and response variables extracted from the STOMP-CO₂ simulations are described in Table 1.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>D Depth to injection (m), which affects the fracture pressure, and hence, the maximum pressure differential under which injection can be carried out</td>
<td></td>
</tr>
<tr>
<td>L Well spacing (m), which determines the volume of the closed system into which CO₂ is injected</td>
<td></td>
</tr>
<tr>
<td>$kh_{MS}$ Permeability-thickness product (md-ft) for the injection reservoir (Mount Simon sandstone), which controls the amount of CO₂ that can be injected for a given pressure differential</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Response</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cum_CO₂ Cumulative volume of CO₂ injected (millions of metric tons, MMT)</td>
<td></td>
</tr>
<tr>
<td>CO₂_R Radius of CO₂ plume (m)</td>
<td></td>
</tr>
<tr>
<td>PCT_CO₂ % Mass flux entering the caprock</td>
<td></td>
</tr>
</tbody>
</table>
### 2 Study Methodology

#### 2.1 Metamodels

In this analysis, several metamodels were created for each of the responses in the ARCHES dataset described in Section 1.2. These metamodels were trained using competing techniques, and the accuracies of the models were compared using a cross-validation approach. The goal of the study was to determine which metamodeling methods were most accurate for applications similar to the ones for which the ARCHES dataset was collected.

Suppose that a set of \( p \) predictors \( \mathbf{x} = (x_1, x_2, \ldots, x_p) \) yields a response \( f(x) \). The goal of response surface modeling is to create an approximating function \( \hat{f}(x) \) that minimizes the residual error over the input space (i.e., the range of possible values for \( x \)). To fit such a model, a sample \( \{x^1, x^2, \ldots, x^n\} \) of size \( n \) is drawn, where the \( i^{th} \) observation is given by \( x^i = (x^i_1, x^i_2, \ldots, x^i_p) \).

Five metamodeling strategies were considered for this study. In the oil and gas literature, second order (quadratic) polynomial models and kriging models are often used, so they are included here. Also included here are Multivariate Adaptive Regression Splines (MARS), Additivity and Variance Stabilization (AVAS), and Thin Plate Spline (TPS) models, which are other popular methods for response surface modeling. A description of these models is given below.

- The quadratic polynomial model expresses the response as a sum of all linear, quadratic, and pair-wise cross-product terms between predictors. That is,

\[
\hat{f}(x) = b_0 + \sum_{i=1}^{p} b_i x_i + \sum_{i=1}^{p} \sum_{j > i}^{p} b_{ij} x_i x_j + \sum_{i=1}^{p} b_{ii} (x_i)^2
\]  

(1)

- The kriging model [Cressie, 1993; Simpson et al., 1998] expresses a response at one location as a weighted average of neighboring observations, where the weights have a specified autocorrelation structure. That is,

\[
\hat{f}(x) = \mu(x) + Z(x),
\]

(2)

where \( \mu(x) \) is the overall trend and \( Z(x) \) is the autocorrelation term. In this study, two different trend functions were examined. The first (Ordinary Kriging) used a constant term \( \mu(x) = m \), while the second (Universal Kriging) used the quadratic trend shown in (1). In both cases, a Matérn\((5/2, \theta)\) correlation was used for \( Z(x) \).

- MARS [Friedman, 1991] approximates the response surface using a collection of simple step and hinge functions. Each function is only defined over a particular range of predictor values, and all of the functions collectively form a single piecewise function over the full input space.

- AVAS [Tibshirani, 1988] uses a non-parametric procedure to find transformations of the response and predictors such that the transformed response can be modeled as a linear combination of the transformed predictors with constant error variance. That is, it finds functions \( g_0, g_1, \ldots, g_p \) such that

\[
g_0(f(x)) = \sum_{i=1}^{p} g_i(x_i)
\]

(3)

- TPS [Duchon, 1977] are a generalization of splines in multiple dimensions. The name refers to the modeling of the response using a surface analogous to a thin semi-rigid sheet of metal. This surface can be deformed to fit the response, but at the expense of a penalty applied for non-smoothness.

Each of the metamodels was trained on the ARCHES dataset in the \( R \) statistical computing language [R Development Core Team, 2011]. Table 2 indicates the packages and methods used to fit the models. The three predictor variables were scaled before fitting the metamodels by recoding their ranges of values to the interval \([0, 1]\).
### Table 2. Methods and Packages for Fitting Metamodels in R.

<table>
<thead>
<tr>
<th>Model</th>
<th>R Package</th>
<th>Method</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadratic</td>
<td>stats</td>
<td>lm</td>
<td>Base R</td>
</tr>
<tr>
<td>Kriging</td>
<td>DiceKriging</td>
<td>km</td>
<td>Roustant et al. [2011]</td>
</tr>
<tr>
<td>MARS</td>
<td>mda</td>
<td>mars</td>
<td>Hastie et al. [2011]</td>
</tr>
<tr>
<td>AVAS</td>
<td>acepack</td>
<td>avas</td>
<td>Spector et al. [2010]</td>
</tr>
<tr>
<td>TPS</td>
<td>fields</td>
<td>Tps</td>
<td>Furrer et al. [2011]</td>
</tr>
</tbody>
</table>

#### 2.2 Metamodel Evaluation

In this study, a \( k \)-fold cross-validation approach [see, e.g. Hastie et al., 2009, Chapter 7] was used to evaluate metamodel performance. Under this paradigm, the dataset is partitioned into \( k \) folds, which are mutually exclusive and exhaustive subsets of the observations. Each fold is systematically held out and the metamodel is fit to a dataset consisting of only the remaining \( k - 1 \) folds. This model is then used to make predictions on the observations that were left out. After repeating this process on all \( k \) folds, there are a total of \( k \) models that are constructed, each of which are used to predict the responses for observations in the single fold that was left out of the training set. Error estimates produced in this way do not require collection of additional validation data, which was not possible in this case, and are less biased than the error measured over the training data.

While the cross-validation approach does not specifically test the unique model that is created by using all \( n \) training observations together, it does test the algorithm that is used to construct the model. When each fold is held out of the training set, it will behave like independent test data as far as that particular model is concerned. Therefore, the error magnitudes from the cross-validation more accurately reflect error rates in the model fit over parts of the response surface that have not been sampled.

In general, selecting the number of folds \( k \) is a balance between the bias in estimating the model accuracy and the variance in model fits. If \( k \) is small (i.e., each fold contains many observations), the model fits will be made using a smaller subset of the dataset, with less overlap in training data between models. This can result in large variations in the models as successive folds are held out, and thus create highly variable estimates of model accuracy from fold to fold. If \( k \) is large (i.e. each fold contains only a few observations), the model fits will be less variable, since most of the training data will be in common between fits. However, estimation of model accuracy will tend to be more optimistically biased, since all of the models will begin to look similar to the full model trained using all \( n \) observations.

Since the ARCHES dataset contains only 36 observations and 3 predictors, there is clearly the potential for highly variable model fits if \( k \) is too small. For this study, \( k \) was chosen to be 12, which places 3 observations within each fold. Since model fits are being made with 33 observations, they are unlikely to vary wildly from fold to fold. Also, with such a small sample size, bias is unlikely to be a problem since all of the observations carry a lot of weight in the model fitting process. This guarantees that a model created with one fold removed will be different than the model trained using all of the data.

The 12-fold cross-validation procedure was repeated 100 times for each metamodel, with fold memberships being randomly selected each time. For each response, this produced 100 cross-validated predictions at every set of sampled predictor values. The overall measure of metamodel accuracy was the root mean squared error \((RMSE)\), which is the square root of the average squared difference between the true and predicted responses. In this case, let \( \hat{f}_j(x^i) \) be the prediction of the response in the \( j^{th} \) cross-validation replicate for the \( i^{th} \) observation. In similar fashion, let \( f_j(x^i) \) be the true response. Then define the metamodel accuracy for that response to be:

\[
RMSE = \sqrt{\frac{1}{36} \sum_{i=1}^{36} \frac{1}{100} \sum_{j=1}^{100} (\hat{f}_j(x^i) - f_j(x^i))^2} \tag{4}
\]
Table 3. Metamodel Performance, 12-Fold Cross-Validation

<table>
<thead>
<tr>
<th>Model Name</th>
<th>RMSE (SRMSE)</th>
<th>CUM_CO2</th>
<th>CO2_R</th>
<th>PCT_CO2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadratic</td>
<td>2.376 (0.108)</td>
<td>900.504 (0.095)</td>
<td>0.466 (0.216)</td>
<td></td>
</tr>
<tr>
<td>Ordinary Kriging</td>
<td>0.640 (0.029)</td>
<td>784.904 (0.083)</td>
<td>0.087 (0.040)</td>
<td></td>
</tr>
<tr>
<td>Universal Kriging</td>
<td>0.766 (0.035)</td>
<td><strong>536.496 (0.057)</strong></td>
<td><strong>0.080 (0.037)</strong></td>
<td></td>
</tr>
<tr>
<td>MARS</td>
<td>9.811 (0.445)</td>
<td>1322.644 (0.140)</td>
<td>1.092 (0.507)</td>
<td></td>
</tr>
<tr>
<td>AVAS</td>
<td>6.996 (0.317)</td>
<td>1393.492 (0.147)</td>
<td>0.443 (0.206)</td>
<td></td>
</tr>
<tr>
<td>TPS</td>
<td>3.924 (0.178)</td>
<td>1035.832 (0.109)</td>
<td>0.694 (0.322)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Metamodel Performance, 6-Fold Cross-Validation

<table>
<thead>
<tr>
<th>Model Name</th>
<th>RMSE (SRMSE)</th>
<th>CUM_CO2</th>
<th>CO2_R</th>
<th>PCT_CO2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadratic</td>
<td>2.455 (0.111)</td>
<td>933.624 (0.098)</td>
<td>0.480 (0.223)</td>
<td></td>
</tr>
<tr>
<td>Ordinary Kriging</td>
<td>1.305 (0.059)</td>
<td>743.243 (0.078)</td>
<td>0.137 (0.064)</td>
<td></td>
</tr>
<tr>
<td>Universal Kriging</td>
<td><strong>1.186 (0.054)</strong></td>
<td><strong>611.898 (0.065)</strong></td>
<td><strong>0.117 (0.054)</strong></td>
<td></td>
</tr>
<tr>
<td>MARS</td>
<td>10.061 (0.456)</td>
<td>1574.248 (0.166)</td>
<td>1.157 (0.537)</td>
<td></td>
</tr>
<tr>
<td>AVAS</td>
<td>7.083 (0.321)</td>
<td>1253.068 (0.132)</td>
<td>0.476 (0.221)</td>
<td></td>
</tr>
<tr>
<td>TPS</td>
<td>4.214 (0.191)</td>
<td>1141.412 (0.120)</td>
<td>0.763 (0.354)</td>
<td></td>
</tr>
</tbody>
</table>

3 Results

Table 3 shows a comparison of model accuracy across all metamodels and responses. Note that the raw \( RMSE \) values have different magnitudes from response to response. This is due to different original scales for the observed responses. To allow better comparison of the \( RMSE \) values, they may be scaled by the average response across the \( n = 36 \) observations, which is denoted here as \( SRMSE \). Both the \( RMSE \) and the \( SRMSE \) are given in Table 3 for each combination of metamodel and response.

For all three responses, the kriging models outperform the other four types of metamodels. In particular, the universal kriging model with a quadratic trend seems best overall. This can be seen, for example, in Figure 1, which shows plots of cross-validated metamodel performance on one of the responses, \( CO_2_R \). Note that the universal kriging model provides relatively stable results across the range of values for this response. The residuals appear to be largely uncorrelated as well, which indicates that there are not large systematic components to the response that are not being accounted for.

The entire cross-validation procedure was also repeated using 6 folds and 100 replicate runs. In this case, there are 6 observations in each fold, and the successive metamodel fits are made using fewer observations (30 instead of the 33 in the 12-fold case). Using a smaller number of folds should have the effect of reducing overfitting effects on the predictions and increasing model variability from run to run. Results for the 6-fold cross-validation are shown in Table 4. The kriging models are still the best performers for all responses, although now the universal kriging model is the top performer for the first response as well.

4 Conclusions

In this study, the cross-validation performance of five different metamodeling approaches was measured in a closed volume injection case study of the ARCHES dataset. The dataset contained three \( CO_2 \) predictors, three responses, and 36 observations. Results showed that the kriging metamodels outperformed the others for all three responses – in particular, the universal kriging model with a quadratic trend had the best overall cross-validated \( RMSE \). The quadratic polynomial model was second best, followed by TPS, AVAS, and MARS. These rankings held true for both the 6- and 12-fold
Figure 1. CO2,R metamodel performance in 12-fold cross-validation. Circles represent the median prediction over the 100 replicate runs. Vertical lines indicate various percentile ranges over the 100 replicate runs: Black = 25th - 75th, Dark Grey = 5th - 95th, Light Grey = Min - Max.
cross-validation.

Clearly, a single case study with 36 observations is not sufficient to make a robust comparison between metamodels. Performance is dependent not only on the number of observations, but also on what sampling design was used and how many predictors there are. However, in this particular case, the kriging metamodel was better at uncovering the structure of the response surface (at least at the 36 observed sample points) than the other models. Further testing on this conclusion is currently in progress in another simulation setting. That study uses a larger number of predictors, and is comparing different combinations of metamodels and sampling designs.

ACKNOWLEDGMENTS

This work was supported by U.S. Department of Energy National Energy Technology Laboratory award DE-FE0009051 and Ohio Department of Development grant D-13-02.

REFERENCES


