SPE 163671

Selecting Representative Models from a Large Set of Models

Pallav Sarma, Wen H. Chen and Jiang Xie; Chevron ETC

This paper was prepared for presentation at the SPE Reservoir Simulation Symposium held in The Woodlands, Texas USA, 18–20 February 2013.

Abstract

In order to make the field development decision making and planning process tractable, the decision-makers usually need a few representative models (for example, P10, P50, P90 models) selected from a large ensemble of reservoir models. This ensemble of models may have been obtained from a static and/or dynamic modeling process involving uncertainty quantification (ED), history matching, optimization or other workflows. The usual approach to select a few models is using various variants of clustering. This selection process is not only suboptimal, but it could also be quite difficult to do if multiple output responses and/or percentiles are required and the number of models is large. The current approach in most oil companies is even more naïve, wherein, such models are chosen manually or using Excel spreadsheets. Thus, due to the unavailability of good approaches, representative models are usually chosen based on one or two criteria. Use of such models in the decision making process can lead to sub-optimal decisions. As such, there is a need to automatically select a small set of statistically representative models from a larger set based on multiple decision criteria.

We propose a new model selection approach, namely *minimax* approach, which can simultaneously and efficiently select a few reservoir models from a large ensemble of models by matching target percentiles of multiple output responses (for example, matching P10, P50 and P90 of OPC, WPC and OOIP), while also obtaining maximally different models in the input uncertainty space. The approach requires the simultaneous solution of two *minimax* combinatorial optimization problems. Since this requires the solution a complex multi-objective optimization problem, we instead convert the problem to the solution of a single constrained *minimax* optimization problem. We propose the solution of this optimization problem using a global exhaustive search (for small problems), and a very efficient greedy method wherein a simpler optimization problem can be solved directly by enumeration or by Markov chain Monte Carlo methods for larger problems with many models, target percentiles and variables. The new approach is implemented in Chevron's in-house uncertainty quantification software called genOpt and tested with multiple synthetic examples and field cases. The results demonstrate that the proposed approach is much more efficient than clustering and solution quality is generally better than clustering. For some models, *minimax* was orders of magnitude faster than clustering. The new approach could help business units select P10, P50 and P90 models efficiently for decision making and planning.

Introduction

Reservoir management workflows such as those related to uncertainty quantification, history matching and optimization are slowly but surely moving towards ensemble based approaches, wherein tens to hundreds (or even more if meta-models are used) of models are used to quantify uncertainty in model predictions. Good examples of such workflows are experimental design [Faidi et al., 1996, van Elk et al., 2000, Venkataraman, 2000] and the ensemble Kalman filter [Naevdal et al, 2003, Wen and Chen, 2005, Aanonsen et al, 2009]. However, analysis of such a large number of models in the decision making and planning process is generally intractable, thus, decision-makers may still require a subset of representative models (for example, P10, P50, P90 models) for the decision making and planning process. As such, an approach is required to obtain a small set of representative models from a large ensemble of models. This problem can be formally written as: given N independent input random variables $x_1,...,x_N$ (for example, WOC, permeability multiplier, etc.) and M output random variables of interest $y_1,...,y_M$ (such as NPV, cumulative oil/water/gas production, IOIP, IGIP, etc.), and a large but finite set of K models $\Omega = \{c_1,...,c_K\}$, the problem is to determine a small subset of P models $\tilde{\Omega} = \{c_1,...,c_P\} \subset \Omega$, that are "statistically representative" of Ω, and $P \ll K$.

Since the decision making problem and decision analysis using these models are usually based on the probability distributions of one or more of the output random variables [Smith, 1988], "statistically representative" implies that the models selected in $\tilde{\square}$ are somehow representative of the statistical properties of all the output random variables $Y_1,...,Y_M$ obtained from $\square$. While there can be various ways to define statistical representativeness, our approach is based on selecting models that are close to certain target percentiles (example P10, P50, P90) of all the output variables $y_1,...,y_M$ obtained from $\square$. The reasoning behind such an approach is that these percentiles are the main inputs to decision analysis algorithms, and models selected in this manner are consistent with their use in the decision making process. At the same time, the models selected in $\tilde{\square}$ should be maximally "different" from each other so that they span the N dimensional input uncertainty space as widely as possible. This is necessary for a robust model selection process in the face of changing decision processes, decision variables etc. in the future. It is also imperative that all the decision (output) variables $y_1,...,y_M$ be used in this process, because using just a few can lead to suboptimal model selection as will be demonstrated later.

The usual approach for model selection, at least in the petroleum industry, is by the application of some variant of clustering [MacQueen, 1967], such as distance-based clustering [Scheidt and Caers, 2008], where k-means clustering is used with distances between models defined based on some of the input and output variables. Clustering based approaches essentially try to find models that are different from each other based on some definition of distance between them, as such, the objective of obtaining "statistical representativeness" as defined here is not directly tackled. Thus, the selected models may not be consistent with the percentiles of the output variables $y_1,...,y_M$ used in decision analysis. Further, because clustering (k-means) is an n-p hard problem [Aloise et al., 2009], it can also be quite difficult to apply practically if the number of input and output variables and the number of models is large. The current approach in most oil companies can be even more naïve, where the P10, P50 and P90 models used in the decision making process are usually based on just one or two variables like IOIP, and obtained manually or using Excel spreadsheets.

In this work, we propose a new model selection approach aimed at directly solving the problem of statistical representativeness as defined here. The algorithm, called *minimax*, can simultaneously and efficiently select a few reservoir models from a large ensemble of models by matching target percentiles of multiple output responses (for example, matching P10, P50 and P90 of OPC, WPC and OOIP), while also obtaining maximally different models in the input uncertainty space. The approach requires the simultaneous solution of two *minimax* combinatorial optimization problems, hence the name [Russell and Norvig, 2003]. Since this requires the solution a complex multi-objective combinatorial optimization problem which can easily become intractable for a moderate number of target percentiles and large number of models, we instead convert the problem to the solution of a single constrained *minimax* optimization problem. We also propose the solution of this optimization problem using a global exhaustive search (for small problems), and a greedy method based on Markov chain Monte Carlo for larger problems with many target percentiles and variables. The new approach is implemented in Chevron's in-house uncertainty quantification software called genOpt. The last section demonstrates the application of *minimax* to a field case and comparison to distance-based clustering [Scheidt and Caers, 2008].

The *MiniMax* Algorithm for Model Selection

As explained above, model selection based on statistical representativeness as defined above while also maximizing spread in the input uncertainty space can be approached directly by simultaneously solving two *minimax* optimization problems. However, before the optimization problems can be defined, the decision-makers have to decide how many models are required for the decision analysis problem and what should these models statistically represent. In other words, if *P* models are required, the decision-makers have to determine which percentiles of the *cdfs* of the output variables these models should approximate. For example, in case 3 models are required, the decision-makers have to determine whether they want the 3 models to represent the P10, P50 and P90 or the P25, P50 and P75 etc. of the distributions of $y_1,...,y_M$. Once the goal is defined, the optimizations problems can be written formally as:

$$\min_{X^1,...,X^P \in \square} \sum_{j=1}^{P} \left\{ \max_k \min_l \left| \frac{y_k^j - \bar{y}_k^l}{\bar{y}_k^l} \right| \right\} \qquad \forall\ k = 1,...,M;\ l = 1,...,P$$

sub to:

$$unique\left(\left[prctile\left(y_k^1\right),..., prctile\left(y_k^P\right) \right] \right) = P \qquad \forall\ k = 1,...,M \tag{1}$$

where:

$unique(X)$ = number of unique elements in vector X

$prctile(x)$ = index $(1,...,P)$ of closest target percentile to *x*

$$\max_{X^1,...,X^P \in \square} \left\{ \min_{j,l;\, j \neq l} \left(\min_i \left| x_i^j - x_i^l \right| \right) \right\} \qquad \forall\ j,l = 1,...,P;\ i = 1,...,N \tag{2}$$

Here, $X = [x_1, ..., x_N]$ is a model from the set Ω and $\bar{y}_k^l$ is the target percentile of y_k corresponding to index *l*. Equation (1) defines the problem of finding *P* models that are close to *P* target percentiles of each of the *N* input variables $x_1, ..., x_N$ such that no two models match the same percentile of any of the *M* output variables $y_1, ..., y_M$. This constraint makes this a very hard combinatorial optimization problem to solve. Equation (2) defines the problem of maximizing spread between the selected models in the input uncertainty space, and this problem can be solved relatively easily. However, both optimization problems have to be solved simultaneously, therefore, overall this results in a complex multi-objective combinatorial optimization problem. In order to make this problem tractable, we can instead solve a simpler one-objective combinatorial optimization problem, defined below:

$$\max_{X^1,...,X^P \in \Omega} \left\{ \min_{j,l;\, j \neq l} \left(\min_i \left| x_i^j - x_i^l \right| \right) \right\} \qquad \forall\ j,l = 1,...,P;\ i = 1,...,N$$

sub to: (3)

$$\sum_{j=1}^{P} \left\{ \max_k \min_l \left| \frac{y_k^j - \bar{y}_k^l}{\bar{y}_k^l} \right| \right\} \leq \varepsilon \qquad \forall\ k = 1,...,M;\ l = 1,...,P$$

$$unique\left(\left[prctile\left(y_k^1\right), ..., prctile\left(y_k^P\right) \right] \right) = P \qquad \forall\ k = 1,...,M$$

Equation (3) essentially states that, among a set of models that are closer than a predefined tolerance ε to the *P* target percentiles, find *P* models that maximize spread in the input uncertainty space such that no two models match the same percentile of any of the *M* output variables $y_1, ..., y_M$. By defining the optimization problem in this manner, instead of solving two optimization problems simultaneously, we are now solving them sequentially, wherein, Equation (1) is solved first to obtain a subset of models $\hat{\Omega}$ that satisfy the tolerance constraint, followed by solution of Equation (2), where the search space is constrained to $\hat{\Omega}$. Note that Equations (1) and (2) can also be combined the other way round, wherein, Equation (1) is the objective function and Equation (2) becomes an additional constraint. However, we prefer Equation (3), as this gives more control and importance to the target percentile matching problem, which is usually the more important one of the two.

For small problems with few hundred models, few input and out variables (less than 10) and a few target percentiles (less than 5), Equation (3) can be solved by exhaustive enumeration to obtain the globally best set of models satisfying Equation (3). However for larger problems with thousands of models and many target percentiles, such an approach is not feasible due to the combinatorial nature of Equation (1). The most common approach for solution of such combinatorial optimization problems are Markov Chain Monte Carlo (MCMC) [Press et al., 2007] approaches such as the well known Simulated Annealing (SA) [Kirkpatrick et al, 1983]. SA was applied to a few real and synthetic datasets but did not perform very well compared to a "greedy" algorithm we propose below.

The basic idea behind the greedy algorithm is that instead of solving for *P* models simultaneously as in Equation (1), the *P* models are solved for sequentially, while ensuring that the j^{th} model found does not match the same target percentiles as the models found before (*j*-1 to 1). If Equation (4) is implemented such that after the j^{th} model is found, the target percentiles that the *jth* model matches are removed from the next optimization, then the non-zero constraint (similar to the uniqueness constraint) is not necessary. Equation (4) implemented in this manner can be solved quite easily even for a large number of models and target percentiles, as it is no longer a combinatorial optimization problem, and scales linearly with the number of models, variables and percentiles. It is however, a greedy local algorithm, and it can easily get trapped in local minimum. Thus, it is necessary to evaluate Equation (4) many times (thousands) with an additional randomization step and select the best solutions (subset $\hat{\Omega}$) for constraining Equation (3). This is still very efficient even for large problems.

for $j = 1:P$

$$X^j = \arg\min_{X^j \in \Omega} \max_k \min_l \left| \frac{y_k^j - \bar{y}_k^l}{\bar{y}_k^l} \right| \qquad \forall\ k = 1,...,M;\ l = 1,...,P - j + 1$$

sub to:

$$nnz\left(prctile\left(Y^j\right) == prctile\left(Y^q\right) \right) = 0 \qquad \forall\ q = j-1,...,1 \tag{4}$$

end

where:

$nnz(X == Y)$ = number of non zero elements in vector X==Y

$prctile(X)$ = vector of indices $(1,...,P)$ of closest target percentiles to *X*

The modified version of Equation (4) with randomization is below. Here, instead of selecting the best X^j that is closest to a set of target percentiles, we randomly choose a model X^j from models that are close to a set of target percentiles within a specified tolerance ε.

for $i = 1 : nTrials$

for $j = 1 : P$

$$\mathrm{X}^j = \arg \underset{X^j \in \square}{rand} \left\{ \left(\max_k \min_l \left| \frac{y_k^j - \bar{y}_k^l}{\bar{y}_k^l} \right| \right) < \varepsilon \right\} \qquad \forall\ k = 1,...,M;\ l = 1,...,P - j + 1$$

sub to: (5)

$$nnz\left(prctile\left(Y^j\right) == prctile\left(Y^q\right) \right) = 0 \qquad \forall\ q = j - 1,...,1$$

end

end

Since subset $\hat{\square}$ is usually not very large (few hundred models), Equation (2) constrained to $\hat{\square}$ can be solved very efficiently. While the optimization problem in Equations (4) or (5) can be solved efficiently by enumeration for a reasonably large set of models, for very large problems (1e5+ models) it may be more efficient to solve it by the usual MCMC techniques like SA or other stochastic optimization techniques. Note that this SA problem is much simpler than the SA as applied directly to Equation (1). For the example dataset below, we applied a simple but efficient form of SA called Threshold Annealing [Dueck and Scheuer, 1989], where the transition probabilities between models was defined as the normalized inverse of the Euclidian distance between the models.

Application to a Real Dataset

The *minimax* approach is demonstrated on a real dataset from a gas condensate field, and the results are also compared to manual model selection and clustering. The dataset is the output of a history matching process, wherein, 841 history matched models were obtained by application of experimental design and genetic algorithms. The field has three reservoirs, and the the input variables of interest for this exercise are the porosity multipliers, permeability multipliers and critical gas saturations of the three reservoirs (9 variables). Figure 1 shows the histograms and *cdfs* of these input variables obtained from the 841 history matched models. The output variables of interest are the gas in places of reservoirs one and two and the full field, and the cumulative gas production of the field (4 variables). Figure 2 shows the histograms and *cdfs* of these output variables obtained from the 841 history matched models. The objective of this exercise is to select 3 and 9 models respectively that are representative of these 841 models. For the 3 model case, the decision makers are interested in the P10, P50 and P90 models of the output variables, and for the 9 model case, the target percentiles are P10, P20,…, P90. For the 3 model selection problem, both the exhaustive enumeration approach and greedy algorithm were used to obtain the best models satisfying Equation (3), however, for the 9 model selection problem, the exhaustive approach is not applicable, thus only the greedy approach was used.

Figure 3 shows the selection of 3 models manually, just based on the P10, P50 and P90 of the *cdf* of cumulative gas production. It is clear from Figure 3 that these 3 models are not representative at all of the P10, P50 and P90 values of the rest of the 3 output variables. This also demonstrates the potential pitfalls of selecting models just based on one or two variables, as is common practice. Figure 4 shows the models selected via *minimax* with exhaustive search, while Figure 5 shows the same obtained with k-means clustering. Clearly, the models chosen via *minimax* are superior compared to both manual and clustering approaches in terms of matching target percentiles. Figure 6 shows the models selected with *minimax* with the greedy algorithm. Although the selected models are not as good as *minimax* with exhaustive search, they are certainly acceptable. An interesting observation from Figure 4 is that the P10 model of IGIP1 is the P90 model of IGIP2, the P90 of IGIP1 is the P50 of IGIP2, and P50 of IGIP1 is the P10 of IGIP2. This demonstrates that due to complex nonlinear relationships between variables, a given model may be close to different target percentiles for different variables. Clearly, from Figure 7, there is somewhat of an inverse relationship between IGIP1 and IGIP2, resulting in this behavior. Physically, this may be due to the fact that the models are a result of history matching, thus, in order to match total gas in place, if IGIP1 is increased, IGIP2 (or IGIP3) has to be decreased to match observations. Thus, whenever a model is qualified in terms of percentiles (e.g. P50 model), it has to be with respect to some variable (e.g. P50 w.r.t. oil in place), as just stating "P50 model" has no meaning.

Figure 8, Figure 9 and Figure 10 shows how the selected models are spread in the input uncertainty space. Clearly, both *minimax* and clustering perform similarly as both algorithms are essentially trying to find models different from each other. However, the model spread from these approaches seems to be somewhat better compared to the manual approach.

Figure 11 and Figure 12 show the selected models using *minimax* and distance-based clustering [Scheidt and Caers, 2008] for the 9 model selection problem on the *cdfs* of the 4 output variables. Again, as in the 3 model case, the models selected by *minimax* are much closer to the target percentiles overall compared to the distance-based clustering approach, and are more evenly spread. Further, as seen in Figure 13 and Figure 14, the selected models are also better spread out in the input

uncertainty space with *minimax* compared to distance-based clustering. Note that the results shown here are using enumeration for the optimization problem in Equation (5), although the results with SA are comparable.

Figure 15 compares the computing time required by *minimax* vs. k-means clustering as implemented in Chevron's proprietary software genOpt. Clearly, at least compared to this particular implementation of clustering, *minimax* is orders of magnitude faster. However, computational efficiency compared to other implementations of clustering will need further investigation.

Conclusions

This paper discusses the development and application of the *minimax* algorithm for the problem of selecting a few representative models from a large set of models. The need of such an algorithm exists because many uncertainty quantification, history matching and optimization workflows result in a large number of models, and the analysis of such a large number of models in the decision making and planning process is generally intractable. As such, decision-makers may still require a subset of representative models (for example, P10, P50, P90 models) for the decision making and planning process. "Representative" here is defined as "statistically" representative, wherein models are sought that match the target percentiles of the output variables of interest. At the same time, the selected models should be maximally "different" from each other in the input uncertainty space. This is necessary for a robust model selection process in the face of changing decision processes, decision variables etc. in the future. The approach is demonstrated on a real dataset of a gas condensate field, and compared to manual and clustering based approaches. The results indicate that *minimax* provides a superior model selection both in terms of matching target percentiles of the output variables and maximizing spread in the input uncertainty space. Further, *minimax* is orders of magnitude faster compared to k-means clustering as implemented in genOpt.

References

Aanonsen, S.I., Naevdal, G., Oliver, D.S., Reynolds, A.C., and Valles, B., The Ensemble Kalman Filter in Reservoir Engineering – A Review, SPE Journal, 14(3), 393-412, 2009.

Aloise, D., Deshpande, A., Hansen, P., Popat, P, NP-hardness of Euclidean sum-of-squares clustering, Machine Learning, 75: 245–249, 2009.

Dueck, G., Scheuer, T., Threshold Accepting: A General Purpose Optimization Algorithm Appearing Superior to Simulated Annealing, Journal of Computational Physics, Volume 90, Issue 1, 1990.

Faidi, S.A., Ponting, D.K., Eagling, T.L., Experimental Design in Interactive Reservoir Simulation, presented at the Petroleum Computer Conference, Dallas, Texas , 1996.

Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P., Optimization by Simulated Annealing, Science 220 (4598), 1983.

MacQueen, J. B., Some Methods for classification and Analysis of Multivariate Observations, in the Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967.

Naevdal, G., Johnsen, L.M., Aanonsen, S.I., Vefring, E.H., Reservoir Monitoring and Continuous Model Updating Using Ensemble Kalman Filter, SPE paper 84372 presented at the SPE Annual Technical Conference and Exhibition, Denver, CO, 2003.

Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P., Numerical Recipes: The Art of Scientific Computing (3rd ed.), Cambridge University Press, 2007.

Russell, Stuart J.; Norvig, Peter, Artificial Intelligence: A Modern Approach (2nd ed.), Prentice Hall, pp. 163–171, 2003.

Scheidt C. and Caers J., A new method for uncertainty quantification using distances and kernel methods: application to a deepwater turbidite reservoir, SPE Journal, 14 (4): 680-692., 2008

Smith, J.Q., Decision Analysis: A Bayesian Approach, Chapman and Hall, 1988.

van Elk, J.F., Guerrera, L., Vijayan, K., Gupta, R., Improved Uncertainty Management in Field Development Studies through the Application of the Experimental Design Method to the Multiple Realisations Approach, presented at the SPE Asia Pacific Oil and Gas Conference and Exhibition, Brisbane, Australia, 2000.

Venkataraman, R., Application of the Method of Experimental Design to Quantify Uncertainty in Production Profiles, presented at the SPE Asia Pacific Conference on Integrated Modeling for Asset Management, Yokohama, Japan, 2000.

Wen, X.-H., Chen, W.H., Real-time Reservoir Model Updating Using Ensemble Kalman Filter, paper SPE 92991 presented at the SPE Reservoir Simulation Symposium, Houston, TX, 2005.

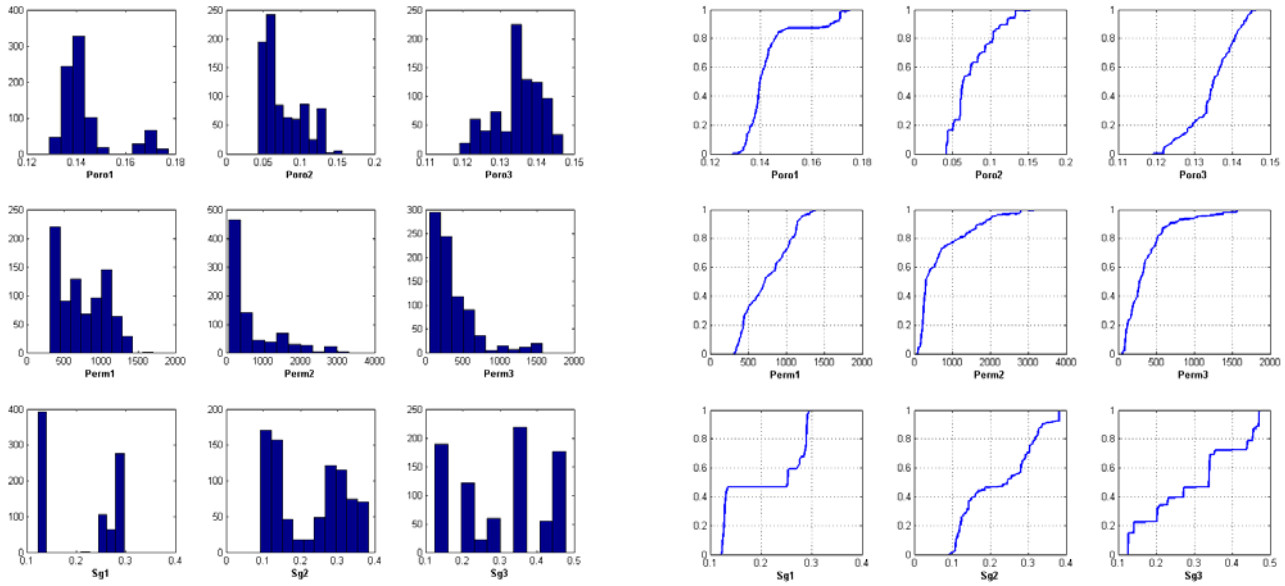

Figure 1 Histograms and *cdfs* of the 9 input variables from the 841 models

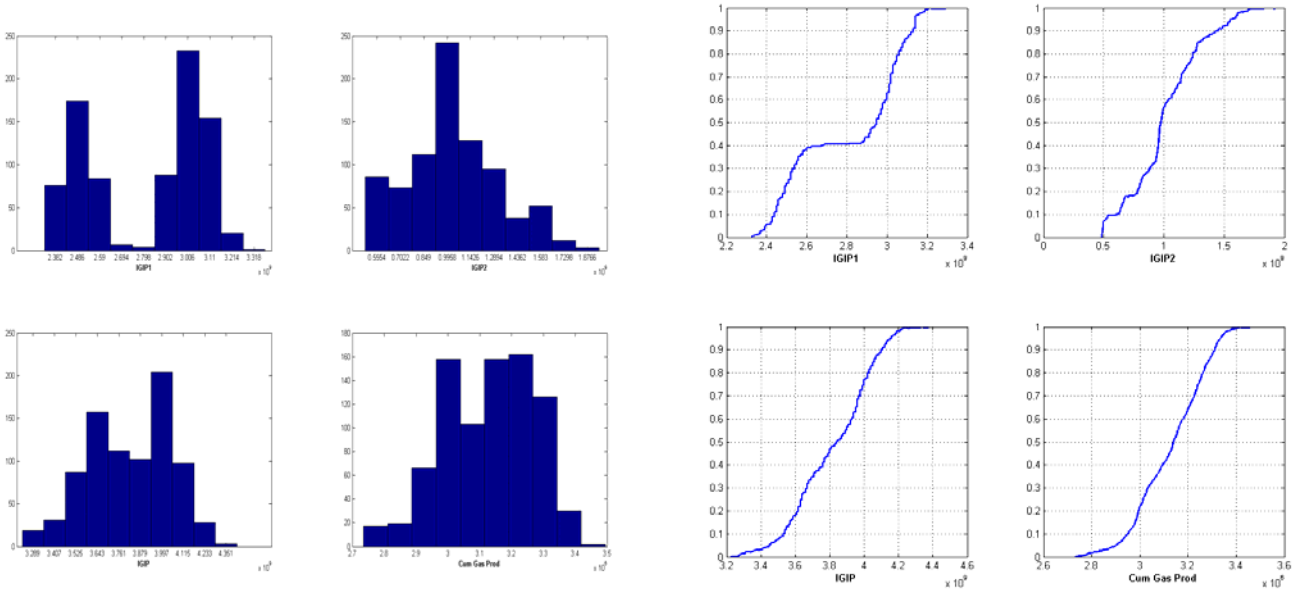

Figure 2 Histograms and *cdfs* of the 4 output variables from the 841 models

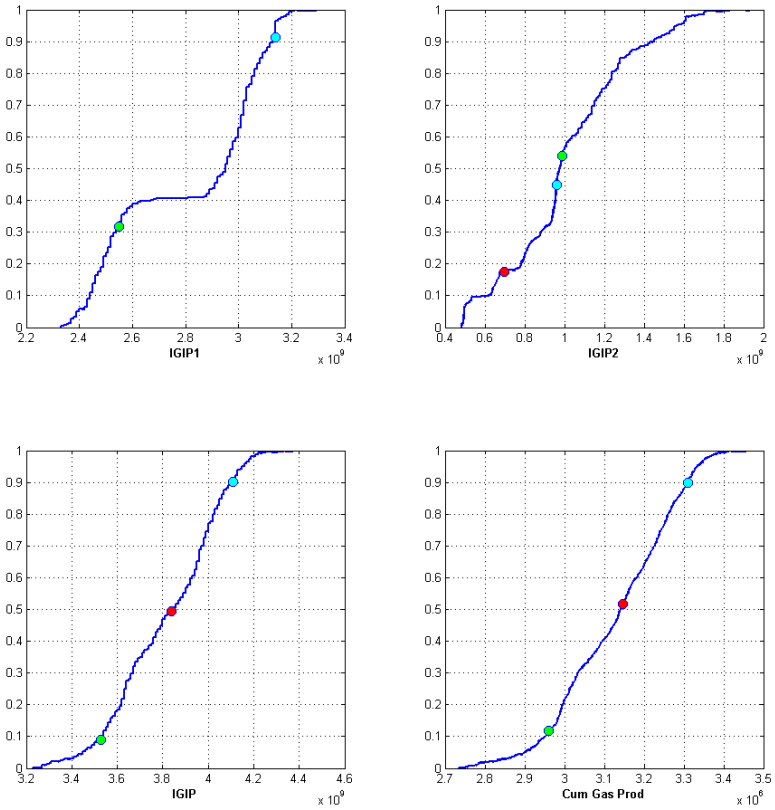

Figure 3 Three models selected using a manual approach

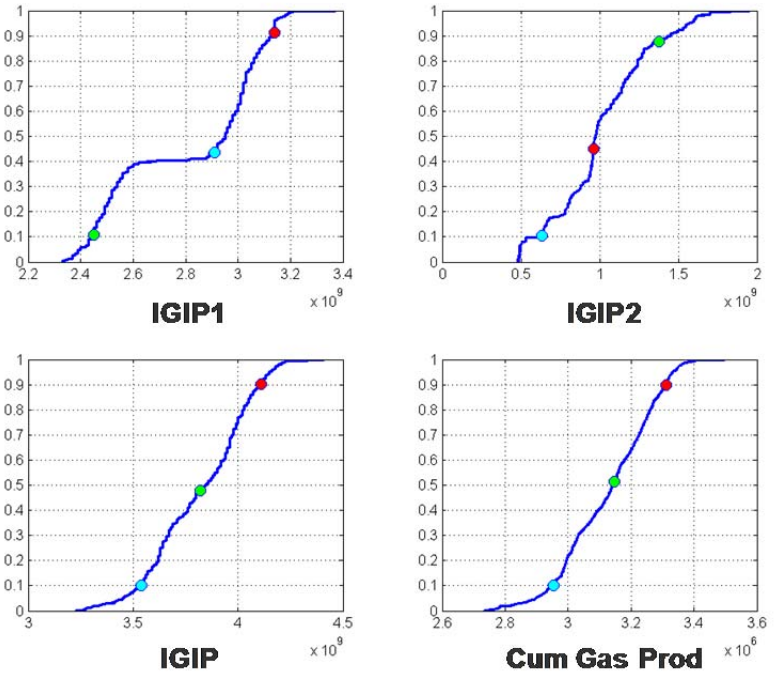

Figure 4 Three models selected using the *minimax* exhaustive search

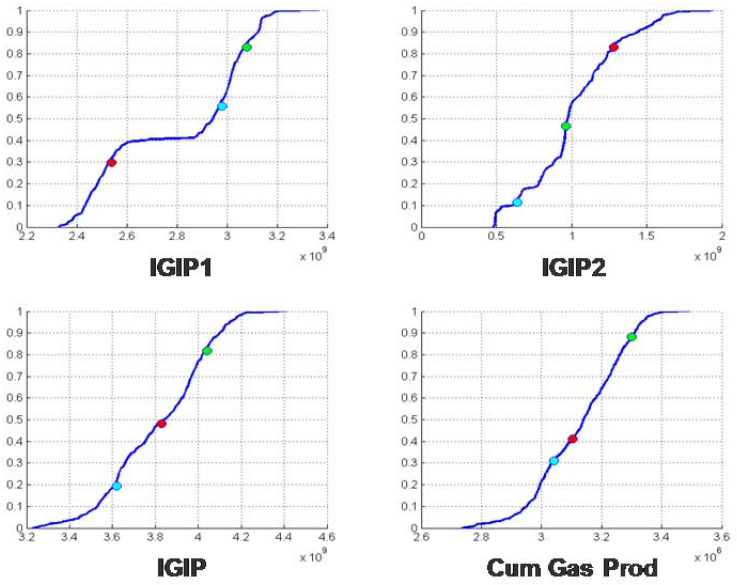

Figure 5 Three models selected using k-means clustering

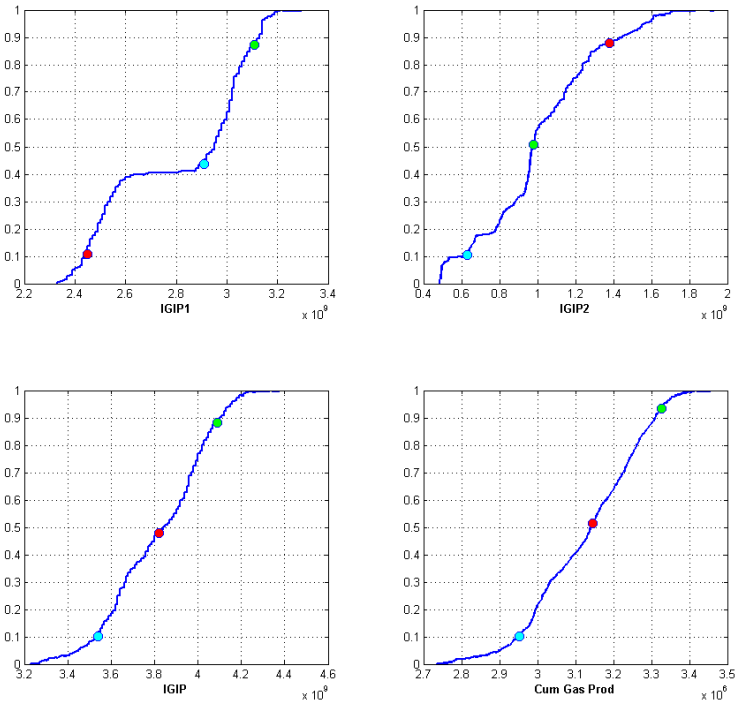

Figure 6 Three models selected using *minimax* greedy algorithm

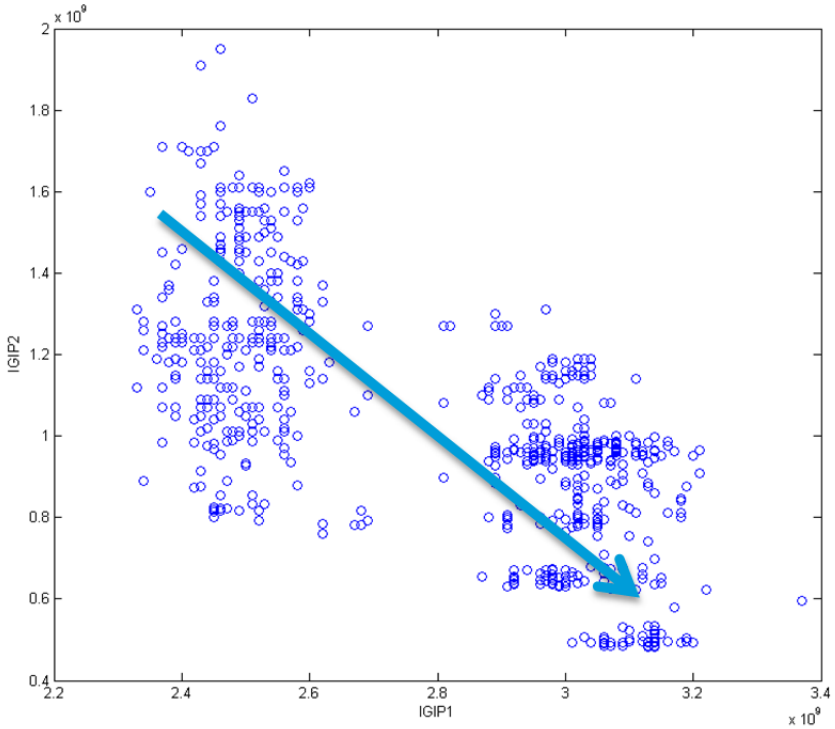

Figure 7 Inverse relationship between IGIP1 and IGIP2

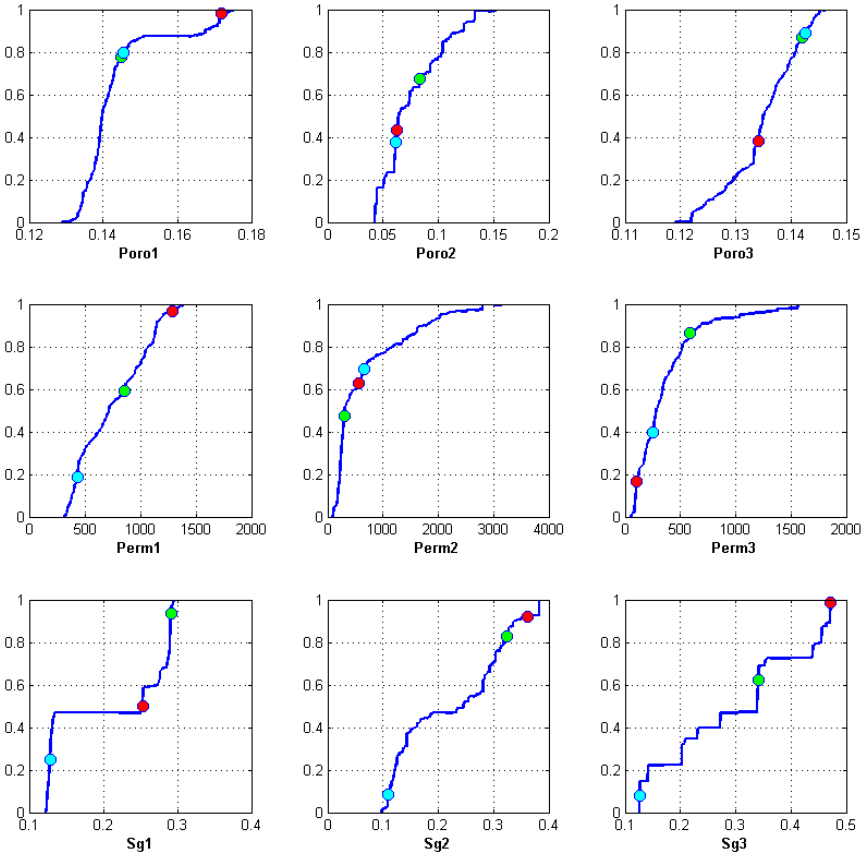

Figure 8 Three models selected using a manual approach a seen on the *cdfs* of the input variables

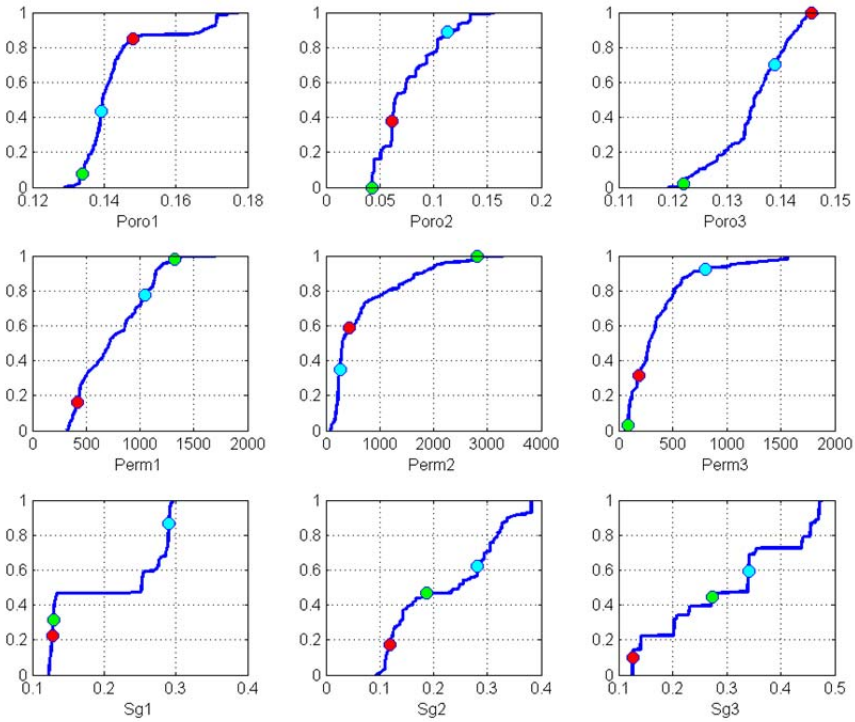

Figure 9 Three models selected using the *minimax* approach a seen on the *cdfs* of the input variables

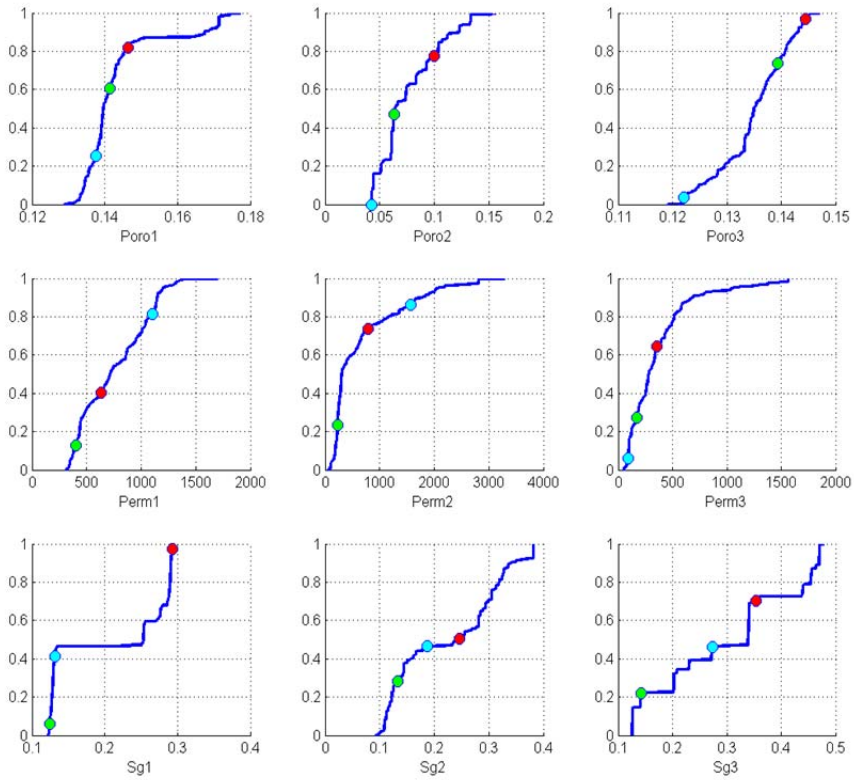

Figure 10 Three models selected using k-means clustering a seen on the *cdfs* of the input variables

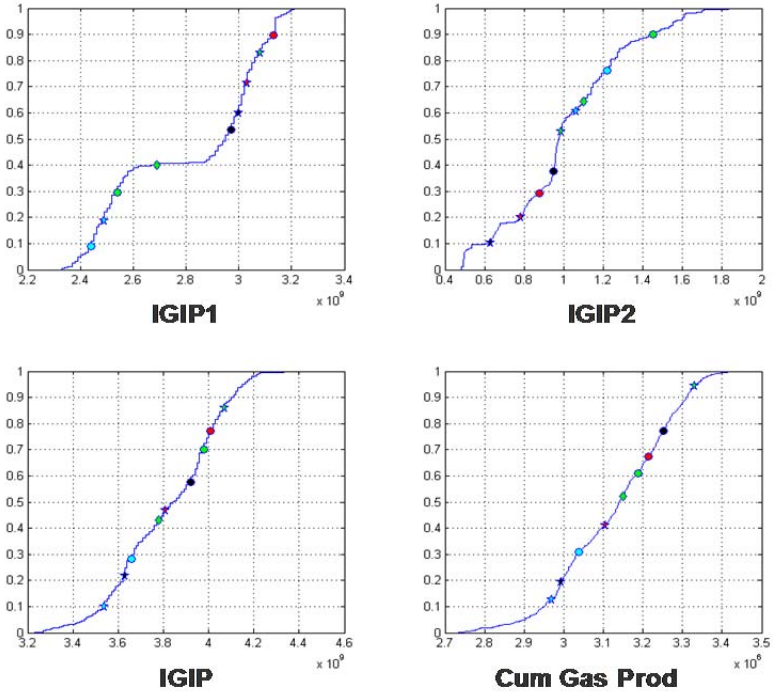

Figure 11 Nine models selected using the *minimax* approach

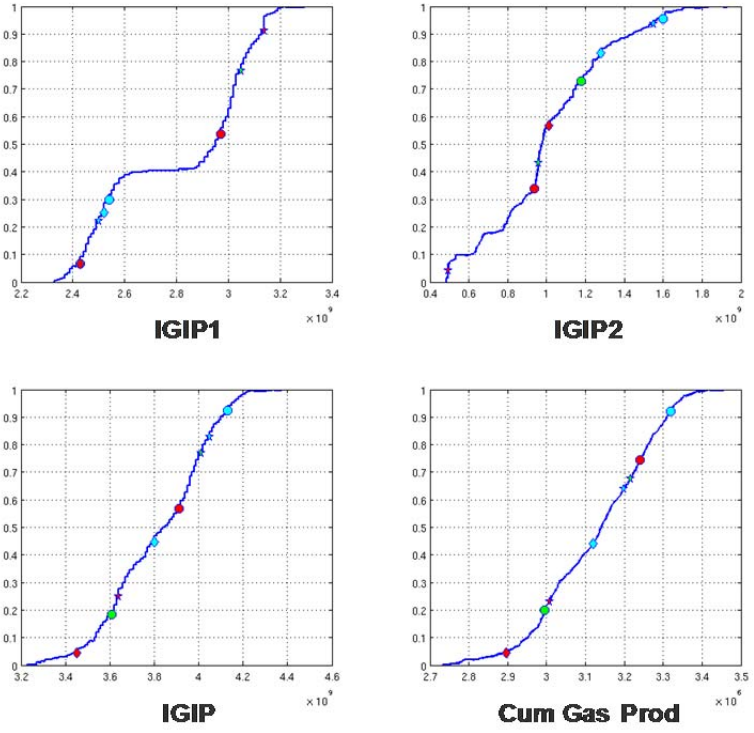

Figure 12 Nine models selected using distance-based clustering

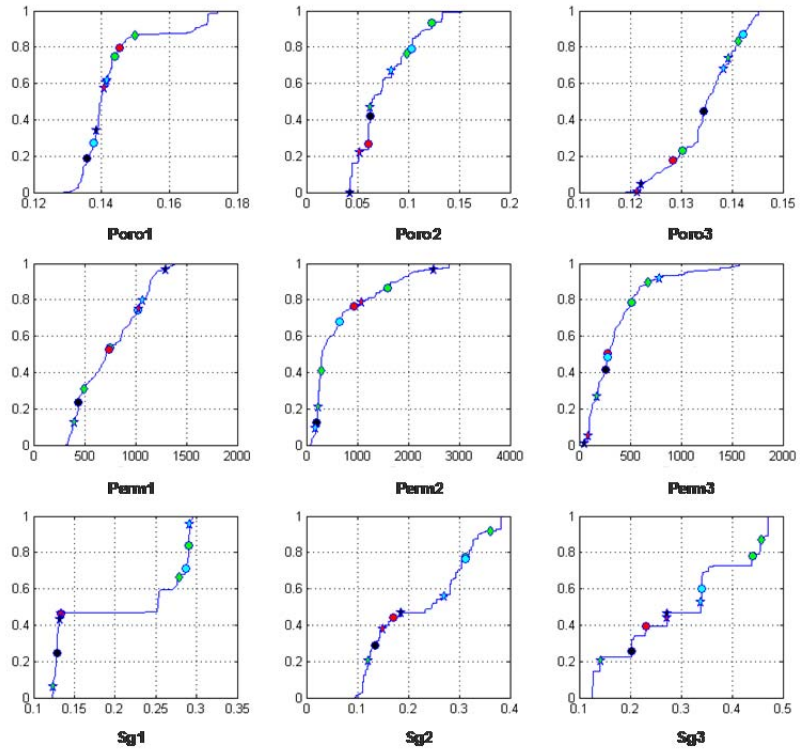

Figure 13 Nine models selected using the *minimax* approach

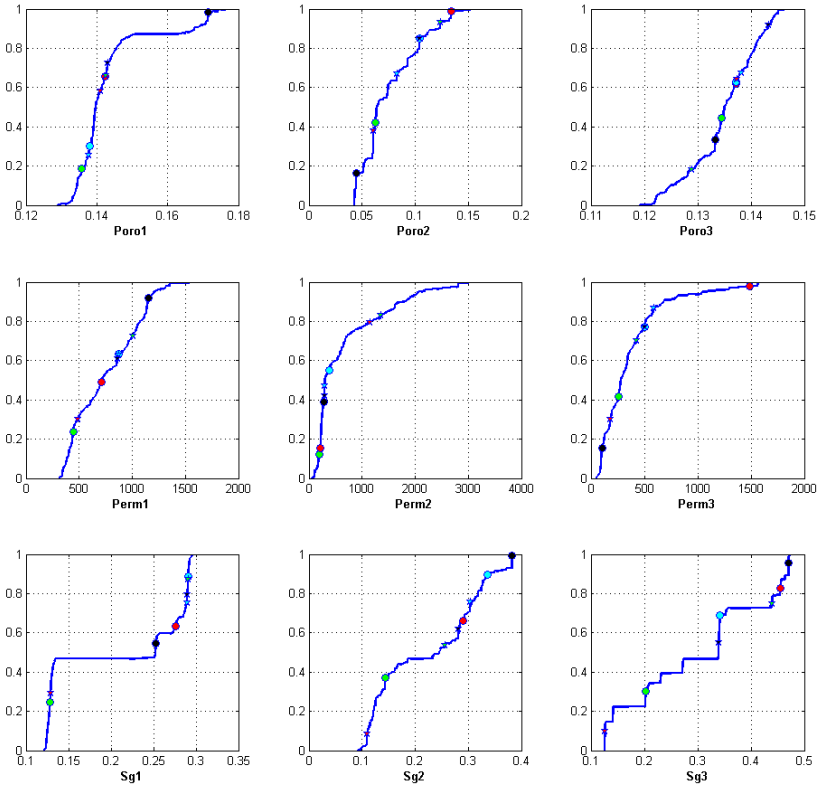

Figure 14 Nine models selected using distance-based clustering

	Minimax approach (seconds)	Clustering method (seconds)
Selecting **3** out of 841 models	1	1493
Selecting **9** out of 841 models	365	More than a day (killed)

Figure 15 Computational time for *minimax* and k-means clustering for the 3 and 9 model selection problems