



SPE 83446

## Identification of Contaminated Data in Hydraulic Fracturing Databases: Application to the Codell Formation in the DJ Basin

Andrei S. Popa, SPE West Virginia University; Shahab D. Mohaghegh, SPE, West Virginia University; Razi Gaskari, SPE, West Virginia University and Samuel Ameri, SPE, West Virginia University.

Copyright 2003, Society of Petroleum Engineers Inc.

This paper was prepared for presentation at the SPE Western Regional/AAPG Pacific Section Joint Meeting held in Long Beach, California, USA 19–24 May 2003.

This paper was selected for presentation by an SPE Program Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the Society of Petroleum Engineers and are subject to correction by the author(s). The material, as presented, does not necessarily reflect any position of the Society of Petroleum Engineers, its officers, or members. Papers presented at SPE meetings are subject to publication review by Editorial Committees of the Society of Petroleum Engineers. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of the Society of Petroleum Engineers is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of where and by whom the paper was presented. Write Librarian, SPE, P.O. Box 833836, Richardson, TX 75083-3836, U.S.A., fax 01-972-952-9435.

### Abstract

With the advance of computer technologies, digitized data is becoming increasingly available. Currently, many companies are in possession of oil or gas-field databases that contain large amounts of information related to hydraulic fracturing, reservoir characterization, production, drilling, etc. However, not all the records are completely accurate or reflect reality. Errors in stored data can be subjective or objective and can be the result of improper or incomplete data collection, errors in data entry, lack of proper interpretation and others. These errors can later lead to poor, erroneous, or even impossible interpretation of the data. This leads to the question: how much of the data is reliable and how can the contaminated data be identified?

The paper presents a new methodology for identification of contaminated data in hydraulic fracturing databases. The methodology combines a set of artificial intelligence tools, which integrates clustering techniques, neural networks modeling and an iterative process to achieve its convergence goal. The result can be seen as a separation of data into three categories: good data, slightly contaminated data, and “bad” data.

This new methodology is applied to two database cases. In the first case, random records in a sample database were intentionally corrupted. Then, using the new methodology presented in this paper, these records were identified as being contaminated. The second case involves a real situation in a Patina Oil and Gas Corporation hydraulic fracturing database, where using classic approaches, the data was unusable for any artificial intelligence processes. After applying this methodol-

ogy to identify contaminated data, successful fracturing interpretation was performed.

### Background

The motivation of this work is part of a comprehensive study performed for the identification of successful practices in hydraulic fracturing in the Codell Formation in the DJ Basin.

Patina Oil and Gas Corporation operates over 3,400 producing wells in DJ Basin. Most of the wells are completed and produced from Niobrara and Codell formations. In many cases there is no clear delimitation between formation and production is considered commingled. Tight gas sand formations made the fracturing stimulation program very successful with double, triple or even more production return after treatment. Until the date of this study (summer 2001), a total number of 230 wells were restimulated.

The success of the restimulation program prompted a study of best practices identification that would increase the chances of success of the ongoing restimulation program. The study consists of a complex data mining analysis as presented by Mohaghegh et al<sup>1</sup>. The five major steps outlined in that work are as follows: Data Quality Control, Fuzzy Combinatorial Analysis, Intelligent Production Data Analysis, Neural Network Modeling, and Successful Practices Analysis.

However, this paper presents only part of step one (Data Quality Control) by detailing the methodology developed for identification of erroneous data records in the data set. The second part of step one, filling in the missing data and “repairing” the data set, is presented as part of a different study.

In a previous paper “Identification of Successful Practices in Hydraulic Fracturing Using Intelligent Data Mining Tools; Application to Codell Formation in DJ-Basin”<sup>1</sup> a complex data mining process was presented.

The overall goal of the work was to identify and eliminate the contaminated data from the database in order to continue the identification of successful practices.

### Statement of the Problem

Unfortunately, in some cases, datasets provided for analysis are not always accurate and useful. The two most common causes are missing data and incorrect data. The authors refer to

incorrect data as data that is contaminated and/or flawed due to different types of errors. The errors can be subjective or objective and can be the result of improper or incomplete data collection, errors in data entry, lack of proper interpretation and others. Possible root causes are: data collection not performed by trained/professional personnel, inattention, gauge recording errors, automated collection, data collection without quality checking and necessary data corrections, or simply random natural chaos.

### Artificial Neural Networks

Artificial neural networks (ANNs) are one of the most widely used artificial intelligence tools in many disciplines. ANNs are an analog, adaptive, distributive, and highly parallel system capable of extracting information and storing knowledge to be used in pattern recognition problems. ANNs provide a powerful tool to perform non-linear, multi-dimensional interpolation. This feature makes it possible to capture the existing non-linear relationships between the input parameters and the output of the system. For successful training, the neural networks must be exposed to sufficient and representative data in order to gain knowledge to accurately predict new situations.

Substantial applicability for artificial neural networks has been found in the petroleum and natural gas industry, particularly in areas such as reservoir characterization, hydraulic fracturing, and optimization and drill bit diagnostics.

### Fuzzy Clustering

Clustering is the grouping of similar objects<sup>2</sup>. In other words clustering is the process of classifying elements of a data set into groups (or classes) according to some similarity criterion. There are two well-known clustering methods: hard C-means (HCM) and fuzzy C-means (FCM) clustering.

This study uses fuzzy C-means clustering for data classification. Fuzzy C-mean or fuzzy clustering is a technique used to group a set of data into clusters such that elements with the same cluster have a high degree of similarity, while elements belonging to different clusters have a high degree of dissimilarity<sup>3</sup>.

Fuzzy C-means clustering presents the advantage that it allows a fuzzy classification in which an element or data can have partial membership in more than one class. Each data will have a number of fuzzy memberships equal to the number of fuzzy clusters chosen in the problem. However the highest membership value defines the cluster to which the element or data belongs the most. This is very important as the fuzzy entropy of the element or data can be calculated.

Fuzzy clustering plays an important role in the solution of problems in the areas of pattern recognition and fuzzy model identification.

The algorithm for fuzzy C-means clustering is presented in detail in many publications<sup>3,4</sup> and will not be repeated here.

### Entropy

By definition entropy is a measure of the lack of order in a system. However in fuzzy clustering, the entropy measures the degree of lack of similarity between elements. The entropy of

an element or data in fuzzy clustering is defined as the ratio of the minimum and the maximum values of the fuzzy memberships belonging to that element or data.

### Methodology

The Patina Oil and Gas Corporation fracturing database is an example where a first attempt of analysis using the contaminated/flawed data led to impossible interpretation of the data. Among the goals of the study was to train a neural network capable of predicting the "Post-Restimulation Peak Production". As the name suggests this is the peak production after stimulation (also referred to in this paper as Actual Peak).

**Input data.** The Patina Oil and Gas dataset consists of 42 parameters. Some of the most important parameters are: Well Name, Latitude, Longitude, Porosity, Net Pay, Average Pressure, Number of Perforations, Original stimulation Sand Volume, Original stimulation Fluid Volume, Restimulation or ReFrac Date, Flowback Volume, Refrac Sand Volume, Refrac Fluid Volume, Average Treating Rate, Peak Viscosity, Calcium ppm, Iron ppm, Chloride ppm, Sulfate ppm and others. The output of the system is Actual Peak.

The entire list of parameters provided in the database ranked using Fuzzy Curve Analysis is given in a previous publication<sup>1</sup>.

**Data quality control.** The original data set consists of 186 records, with each record corresponding to a well that has been restimulated. A first attempt in training the system was completely unsuccessful. At this point, data quality control was initiated. The first step of the quality control process was the identification of outliers. This resulted in identification of 12 wells/records that behaved very different after fracture stimulation than the rest of the wells. These wells returned more than five times the post-peak production after stimulation. These findings were communicated with the Patina engineers who, after further investigation, confirmed that these wells were drilled in a portion of the field known to have strong natural systems of microfractures and should not have been included in the dataset. Therefore, these records were removed and not considered in the analysis.

Included in the data quality control was the generation of regression plots (one parameter versus the other) and frequency distribution plots (Figures 3,4) for each parameter. It is interesting to note that most of the regression plots showed a shotgun distribution (Figures 1,2). From this it can be concluded that no simple or visible correlations were possible.

After the 12 wells in the natural system of microfractures were eliminated, the working dataset consisted of 174 wells/records. Using these records, a second attempt to train a neural network for post-peak production prediction was again unsuccessful. Several types of neural networks were tested. Hundreds of approaches were considered, using different combinations of parameters with different neural networks architectures. The best results, from a neural network training standpoint, had a correlation coefficient of 0.17 and an  $R^2$  of 0.07. These results are similar to that of training a system of

entirely random data. After a solid month of attempting different approaches, it was concluded that a neural network could not be successfully training using this dataset.

Corroborating the regression distribution charts with the unsuccessful neural network training attempts and with previous work performed on this data, it was concluded that there were problems with the data. It is also the understanding of the authors that many other types of correlations had been attempted using these data in the past with no successful results.

The conclusion from this investigation was that some of the records contain contaminated and/or flawed data. The problem posed here was how to find and eliminate these contaminated data.

**Neuro-Cluster Data Classification System.** The methodology described here uses a set of artificial intelligence tools, which integrates clustering techniques, neural networks modeling and an iterative process to achieve its convergence goal. The result can be seen as a separation of data into three categories, good data, slightly contaminated data, and “bad” data. In a previous study<sup>1</sup> the parameters were ranked for their influence on the process outcome.

The most influential ranked parameters were used in this system. First the fuzzy C-means clustering technique is used for data classification. The output of the system, post-peak restimulation production, was included in the clustering dataset. Data was classified in three fuzzy clusters. Based on clustering information, the entropy of each record is calculated. The clustering information together with entropy is added to the dataset and a neural network is trained with relatively good performance. The key in this process that contributes to the neuro-model with good performance is the presence of the output in the clustering process.

Discretizing the output range, the model is run for all the discrete values of the output from minimum to maximum, essentially sweeping the entire interval.

In order to conclude whenever a data record is a good or a contaminated record, three curves are developed. These curves identify the following information for each point in the curve. Each point in the curve is a discretized value of the output.

Curve #1 identifies the cluster (A,B,C or 1,2,3) to which that point belongs to. Curve #2 identifies the entropy of the data record within the cluster it belongs to. Curve #3 identifies the difference between the actual output and the neural network prediction.

Therefore curve #1 will always have a discrete value of either 1, 2 or 3 and is represented by a step function. Curve #2 is a continuous function that always has a value greater than zero. The closeness of this curve (entropy) to zero identifies the higher information content of the data record when the output of the record is replaced with value in the X-axis.

The third curve identifies how close the network prediction is to the actual value of the output in the data record. This curve may assume negative as well as positive values since network predictions can be over-estimations as well as under-estimations.

The key is that the actual output in the data record has been indirectly used in the system input (through fuzzy cluster analysis results) thus making this mainly a bootstrapping technique.

**Algorithm.** The steps involved in data classification are:

1. Cluster data using the output of the system.
2. Train a neural network using the clustering information (membership functions and entropy).
3. Iterative process
  - Select a step for the range of the output system.
  - Sweep the output range while firing the neural network.
  - For each value of the assumed output:
    - Calculate cluster number for the assumed value.
    - Calculate the entropy.
    - Fire the neural network.
    - Calculate Error = | NNoutput – Assumed Output |
  - Display all three curves on a single graph or separate graphs.
  - Identify the position where data error is reaching zero and entropy is at the minimum for a constant cluster name representation.
4. Repeat process starting with step 3 for each of the records in the dataset.

Criteria and constraints are set to classify the records into good records, slightly contaminated records and “bad records”.

A record is considered “good” data when the values of curve #3 (error curve), and curve #2 (entropy) are close to zero at the same value of output (X-axis) while this value of output belongs to a dominant cluster identified by curve #1. (Figure 5). The blue vertical line represents the original output and the red vertical line represents the neural network output for the system. For a good case the blue line, the red line and the error curve (curve #2) at value zero, should be very close or overlapping (Figures 5 and 6).

A record is considered slightly contaminated when either the original output (vertical blue line) or neural network output (vertical red line) is closer to the curve #3 (error curve) at value zero, and either one of the two lines corresponds to minimum entropy (curve #2). Notice that there is considerable difference between the values represented by the vertical red and blue lines (Figures 7 and 8).

A record is considered “bad” when no correlation is observed between the original output (vertical blue line), neural network (vertical red line) and error curve (curve #3). Significant difference is present between the lines and there is no correlation with the entropy curve (curve #2) (Figures 9 and 10).

The above iteration procedure has been optimized in a computer program that uses both the neural network and the fuzzy clustering procedure together to generate the plots presented in figures 5 through 10. The wells/records are run one at a time and the identification is made manually/visually by the user.

In order to test the applicability of this methodology, a new database was created based on a commercial software package (FRACPRO-PT). The database was comprised of ten input parameters and three outputs. The input parameters included formation depth, pad volume, slurry volume, proppant concentration, pumping rate, reservoir pressure, stress profile, formations thickness, formation permeabilities, and reservoir type. The outputs included the fracture geometry, as defined by fracture length, fracture width, fracture height. However in this study only one output of the system was used, namely fracture length. Upon completing the construction of the database several data records were intentionally corrupted. The idea was to see if the developed methodology is capable of identifying the corrupted data records.

This test of the methodology showed that if this methodology is carefully applied to a database that includes several corrupted data records, it can successfully identify the corrupted data records.

### Results and Discussion

The methodology presented above was applied to the database collected from the Patina Oil and Gas field hydraulic fracturing program. Upon the completion of the exercise, the database was classified into 88 good data records, 26 slightly contaminated data (data that could possibly be used), and 60 "bad" data records.

The 88 good data records were used in training a successful neural network that was used for the identification of best practices in the Patina Oil and Gas field. Data was divided into training, calibration and verification sets. The performance of the neural network trained using the 88 cases is presented in Figures 11, 12, and 13. The neural network resulted in an  $R^2$  of 0.82 and a correlation coefficient of 0.876 for the training set.

The most important measure of the neural network accuracy was the correlation coefficient of the verification set. The verification set, which contains the set of data records, not used during the training and calibration process showed an  $R^2$  of 0.806 that is quite satisfactory for our analysis. The results of the verification dataset are shown in Figure 13.

### Conclusions

1. A neuro-cluster system was developed for identification of contaminated data. The system uses a specialized neural network as an optimization tool to predict the output of the system.
2. The applicability of the newly developed methodology was verified using a synthetic dataset developed using commercial fracture stimulation software.
3. The application of this methodology can be extended to any type of database for the identification of contaminated data.
4. The flexibility of the method allows fast identification of corrupted data in datasets.
5. The combination of these two intelligent tools, neural networks and fuzzy C-mean clustering is innovative and

provides a simple solution to problems like data classification, as presented in this paper.

### Acknowledgments

The authors would like to thank Patina Oil and Gas Corporation for supplying the data used in this work.

### References

1. Mohaghegh, S., Popa, A., Gaskari, R., Ameri, S., and Wolhart, S.: Identification of Successful Practices in Hydraulic Fracturing Using Intelligent Data Mining Tool; Application to the Codell Formation in the DJ-Basin", SPE 77597, Proceedings, 2002 SPE Annual Technical Conference and Exhibition, 29 September -2 October, San Antonio, Texas.
2. Ali, J.K.: "Neural Networks: A New Tool for the Petroleum Industry?," SPE 27561 presented at the 1994 European Petroleum Computer Conference, Aberdeen, U.K., March 15-17.
3. Hartigan, J. "Clustering Algorithms". Wiley, New York, 1975.
4. Ross, T.J.: Fuzzy Logic with Engineering Applications". McGraw Hill Inc., (1995).
5. Kroszynski, U. and Zhou, J. "Fuzzy Clustering" IKS, December 1998.
6. Mohaghegh, S., Gaskari, R., Popa, A., Ameri, S., and Wolhart, S.: Identifying Best Practices in Hydraulic Fracturing Using Virtual Intelligence Techniques", SPE 72385, Proceedings, 2001 SPE Eastern Regional Conference and Exhibition, October 17-19, North Canton, Ohio.

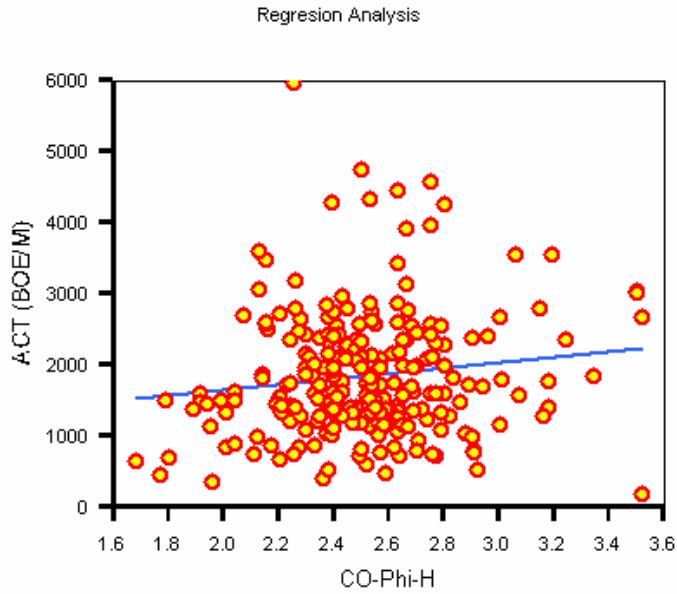


Figure 1. Regression Distribution

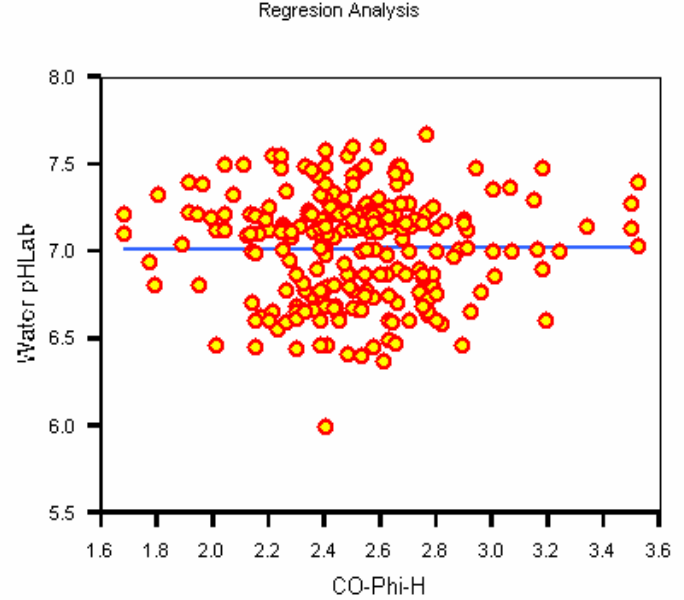


Figure 2. Regression Distribution

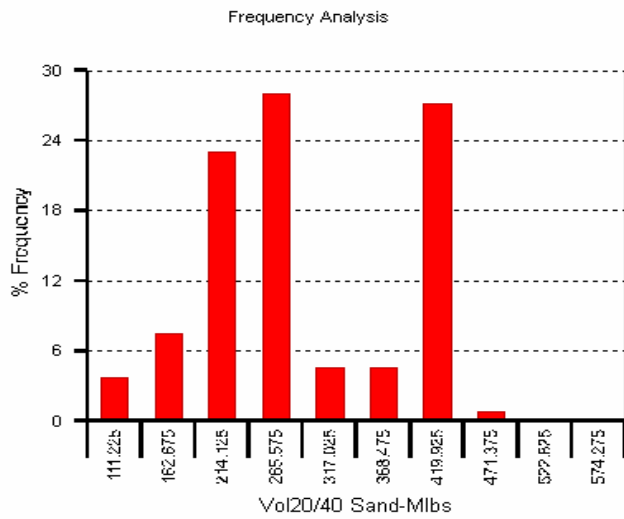


Figure 3. Frequency Distribution

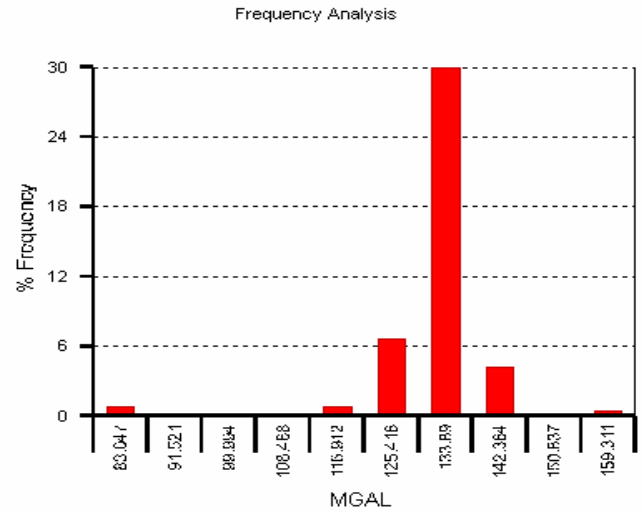


Figure 4. Frequency Distribution

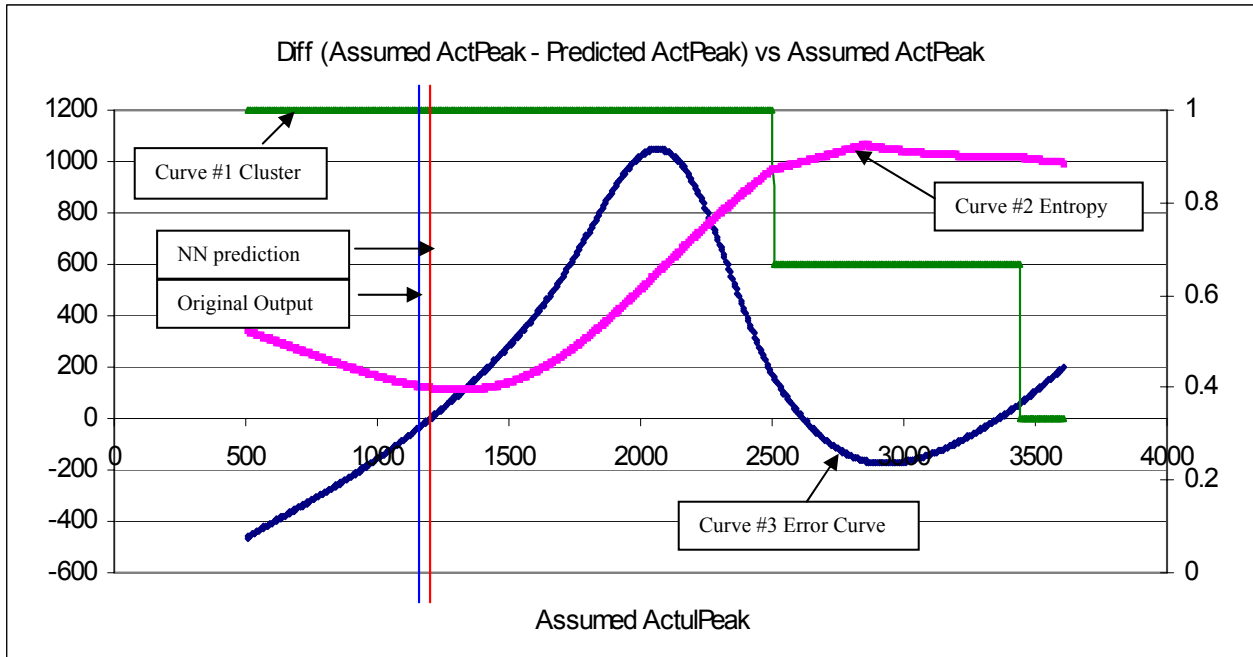


Figure 5 – “Good” data example.

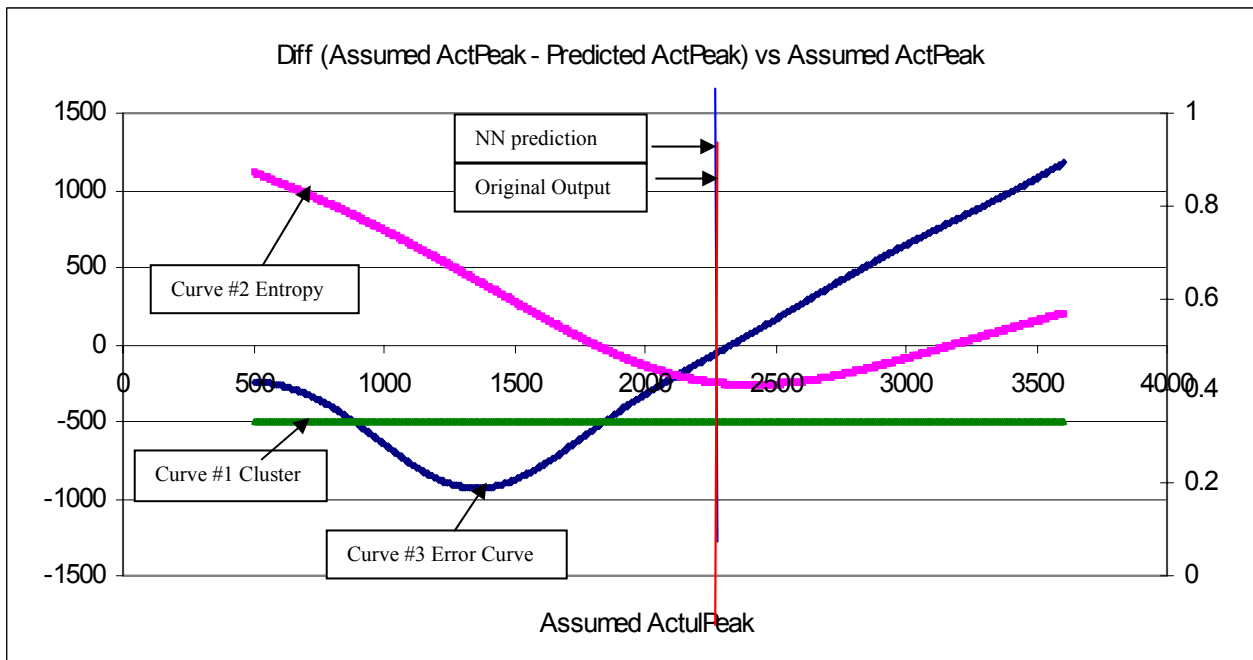


Figure 6 – Good data example

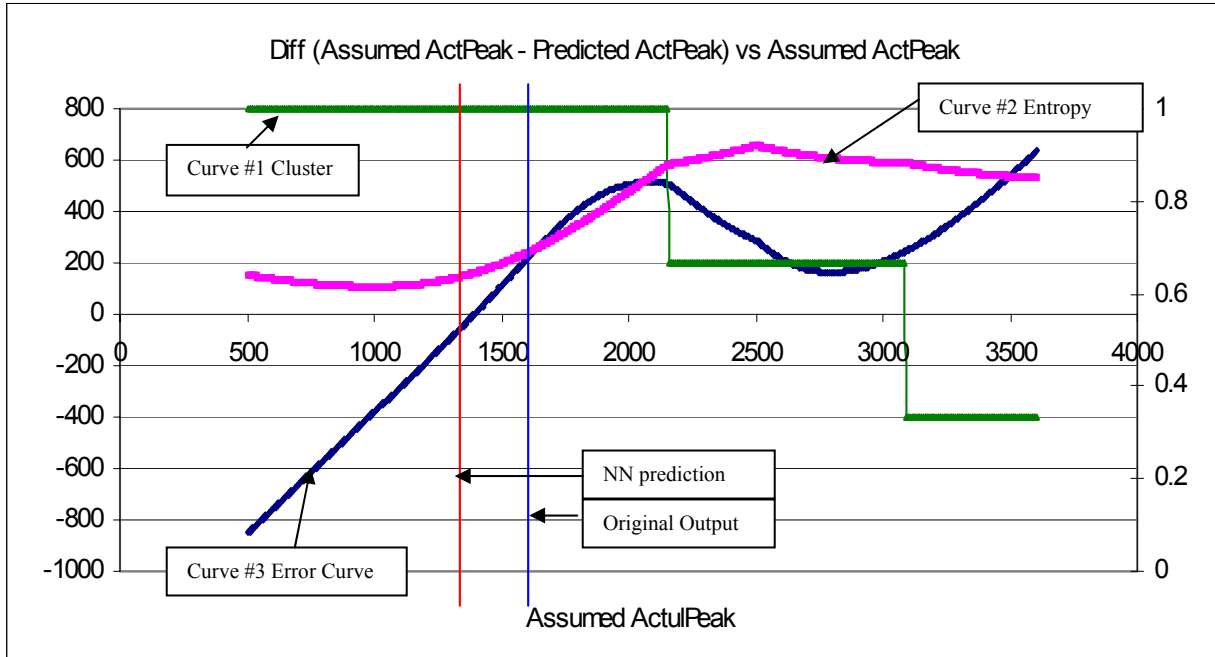


Figure 7 – Slightly Contaminated data example

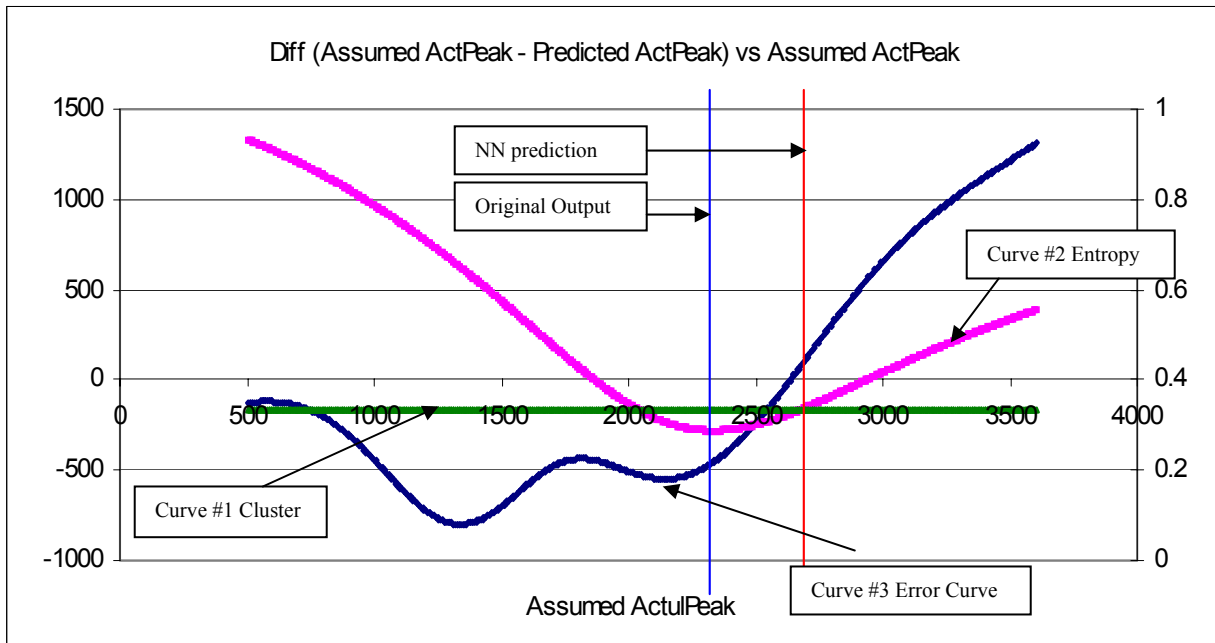


Figure 8 – Slightly Contaminated data example

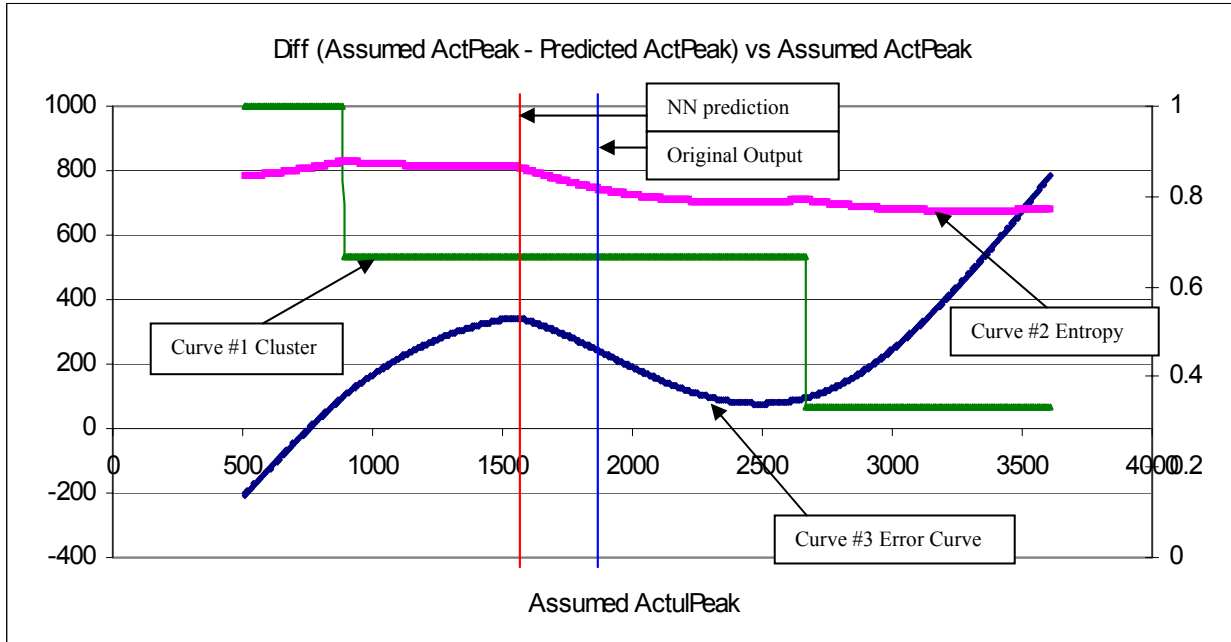


Figure 9 – “Bad” data example

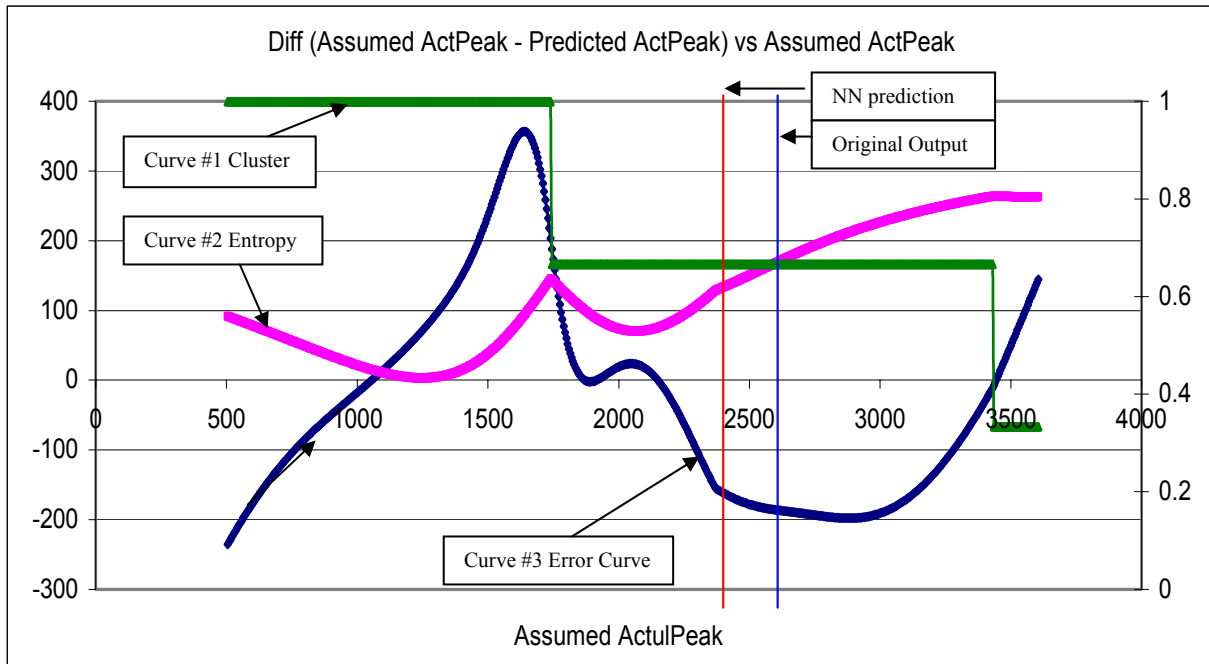


Figure 10 – “Bad” data example

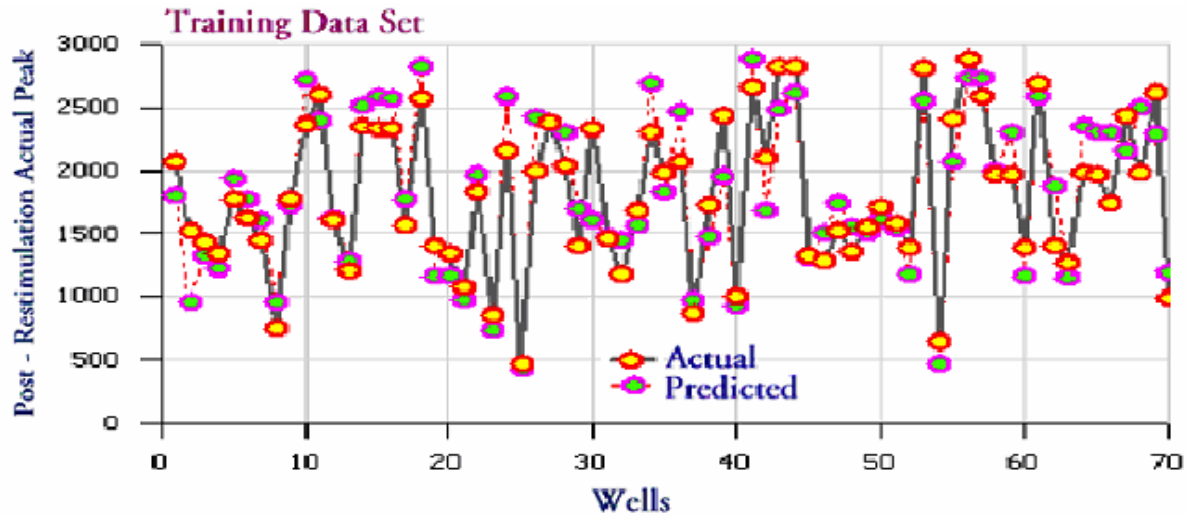


Figure 11. Neural Network - Trainind Data Set



Figure 12. Neural Network - Calibration Data Set



Figure 12. Neural Network - Verlibration Data Set