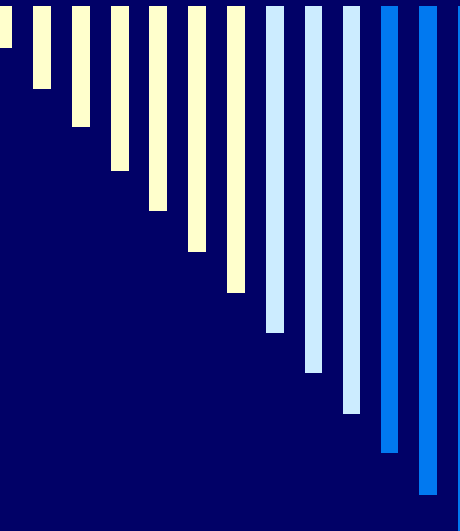


SPE 83446



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

**Andrei Popa, Shahab Mohaghegh,
Razi Gaskari and Sam Ameri**

WEST VIRGINIA UNIVERSITY



Society of Petroleum Engineers



OUTLINE

- Objective**
- Introduction**
- Background**
- Methodology**
- Results & Discussion**
- Conclusions**



OBJECTIVE

- **To introduce a new methodology for identification of contaminated data in hydraulic fracturing databases.**
 - **This methodology Integrates clustering techniques, neural network modeling, and an iterative process to achieve the convergence goal.**



BACKGROUND

- **Part of a comprehensive study of best practices identification for restimulation.**
- **SPE 77597 “Identification of Successful Practices in Hydraulic Fracturing Using Intelligent Data Mining Tools: Application to the Codell Formation in the DJ Basin”.**
- **Incorporates state-of-the-art in data mining, knowledge discovery and data-knowledge fusion techniques.**



BACKGROUND

- **Methodology includes five steps:**
 1. **Data Quality Control**
 2. **Fuzzy Combinatorial Analysis**
 3. **Production Data Analysis**
 4. **Neural Model Building**
 5. **Successful Practices Analysis**



INTRODUCTION

- **Databases are not always accurate and useful!**
 - **Missing data**
 - **Incorrect data or contaminated data**
 - **Improper / incomplete data collection**
 - **Data entry errors**
 - **Lack of proper interpretation**
 - **Gauge reading errors**
 - **Data collection not done by trained personnel**
 - **Inattention**
 - **Automatic data collection without verification / quality control**
 - **Simply random natural chaos**



INTRODUCTION

- **Artificial Neural Networks**
 - **Distributive, parallel processing system capable of solving non-linear problems.**
- **Fuzzy Clustering**
 - **Grouping of similar objects, classifying elements of a data set into groups using similarity criterion.**
- **Entropy**
 - **Measure of the lack of order in a system or the degree of lack of similarity between elements.**



INTRODUCTION

□ Goal:

To train a neural network capable of predicting Post-Restimulation Peak Production.



INTRODUCTION

□ Table of parameters used in the study

Rank	Feature	FCA Value	Rank	Feature	FCA Value
1	Flowback Volbbl	0	22	Frac Type	2.2848
2	CO -Phi-H	0.5811	23	No-CO-Perfs	2.303
3	Bicarbonate ppm	0.6666	24	Chloride ppm	2.3298
4	Peak Visc	0.7486	25	NI- Perfed-H	2.3302
5	Lat	0.7734	26	Water pHLab	2.3665
6	Orig20/40 Sand-Mlbs	0.9214	27	Pre-Refrac Mcfd	2.3956
7	Long	1.1	28	Cum MMcf	2.4009
8	Refrac Date	1.1934	29	Water Source	2.4018
9	ViscShear 100-30Min	1.3324	30	Iron ppm	2.4351
10	TotHardness ppm	1.518	31	MGAL	2.496
11	Calcium ppm	1.6692	32	TotalPerfs	2.5045
12	AvgRate BPM	1.7415	33	Sulfate ppm	2.5164
13	Est-Ult- GOR	1.7706	34	New Perfs	2.552
14	No-NI -Perfs	1.7863	35	Sodium ppm	2.6039
15	AvgPsi	1.8438	36	Magnesium ppm	2.6108
16	ViscShear 100-5Min	1.9401	37	ViscShear 100-0Min	2.6649
17	Top CO Perf	1.9819	38	Pre- FracISDP	2.7127
18	TDSolid ppm	2.0084	39	TestedPH	2.8066
19	MMcf	2.0777	40	Post- FracISDP	2.8256
20	Orig Fluid-Mgal	2.0855	41	Mlb20-40	2.8907
21	DOFP	2.2451	42	Communication	2.9554



METHODOLOGY

- **First attempt at analysis led to impossible interpretation of data.**
- **Data quality control then initiated.**



METHODOLOGY

- **Identification of outliers**
 - **Using Fuzzy Curves together with Regression plots**
 - **12 wells were eliminated as belonging to naturally fractured part of field.**

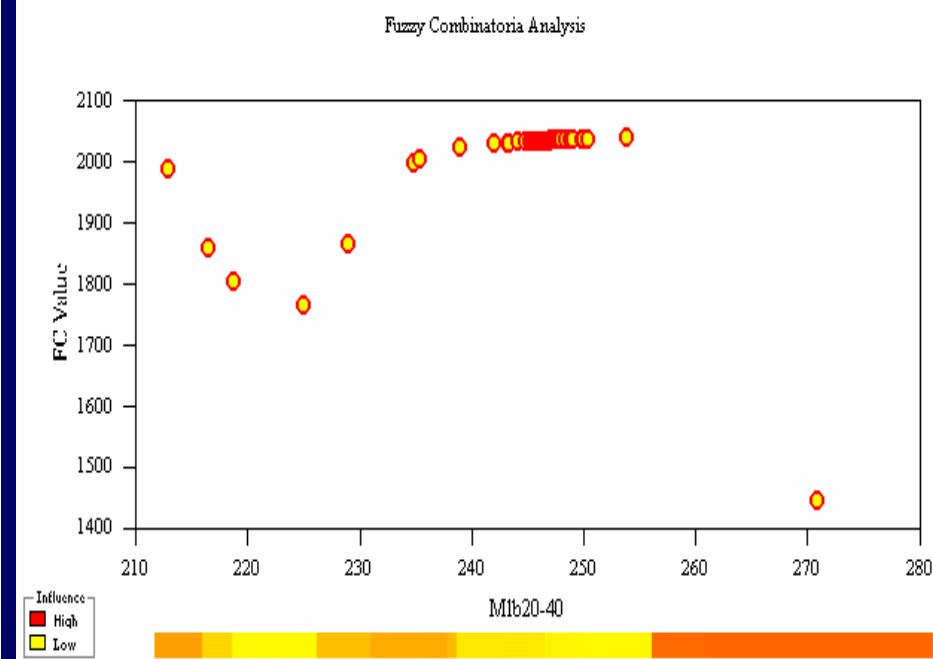
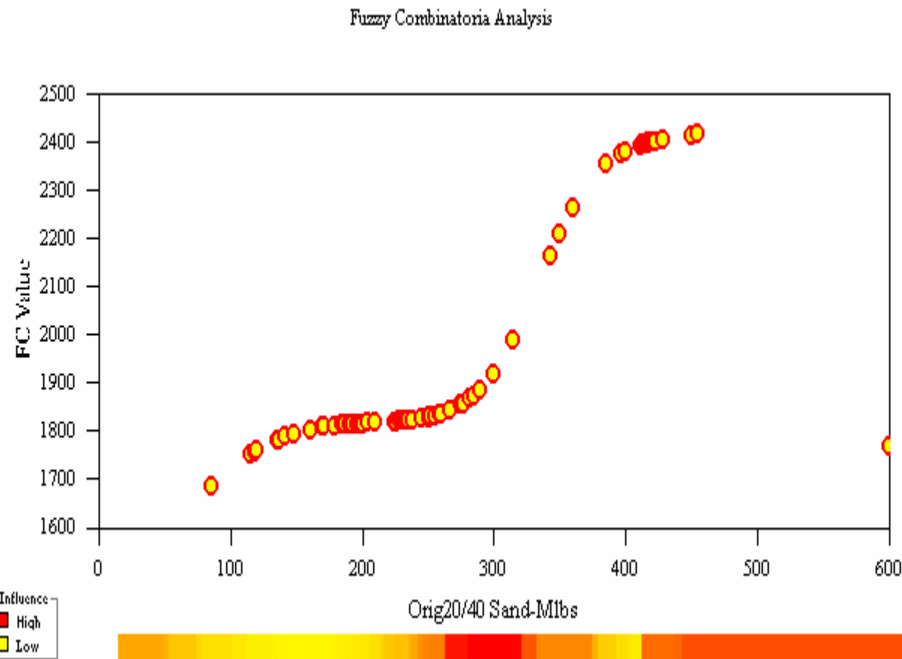


METHODOLOGY

- **Identification of outliers**
 - **Using Fuzzy Curves together with Regression plots**
 - **12 wells were eliminated as belonging to naturally fractured part of field.**

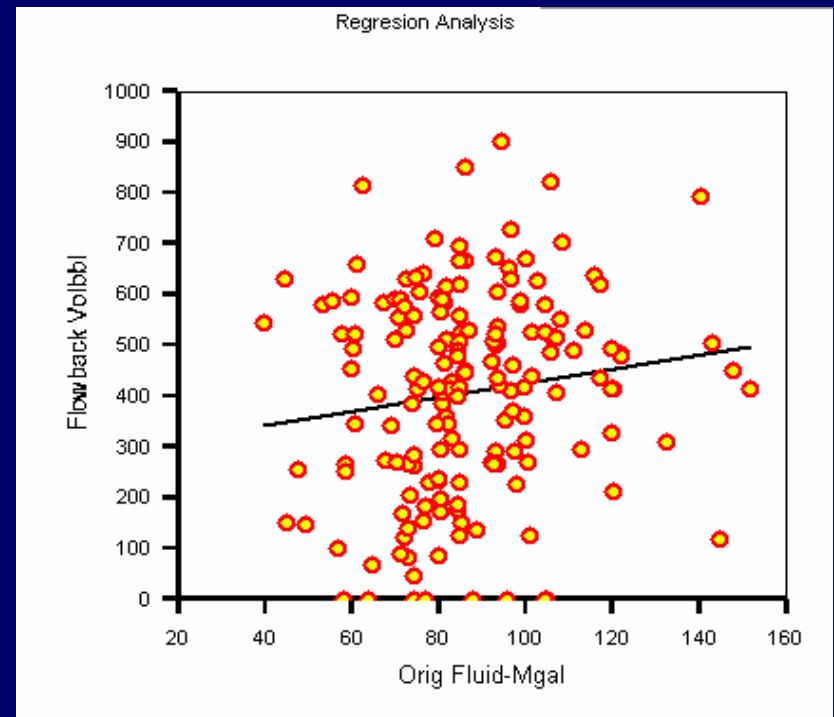
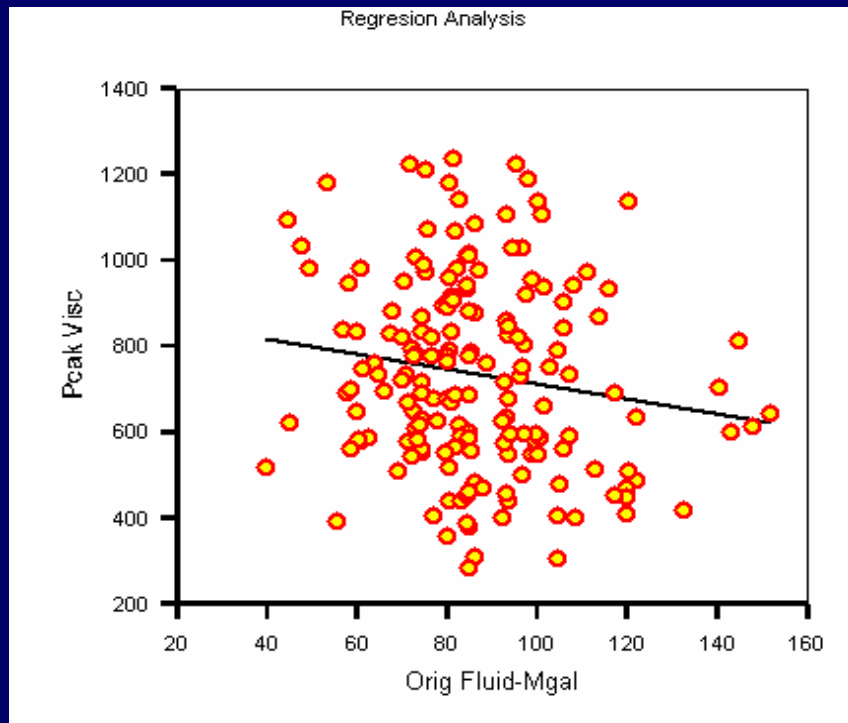
METHODOLOGY

Fuzzy Curve Analysis for outlier identification



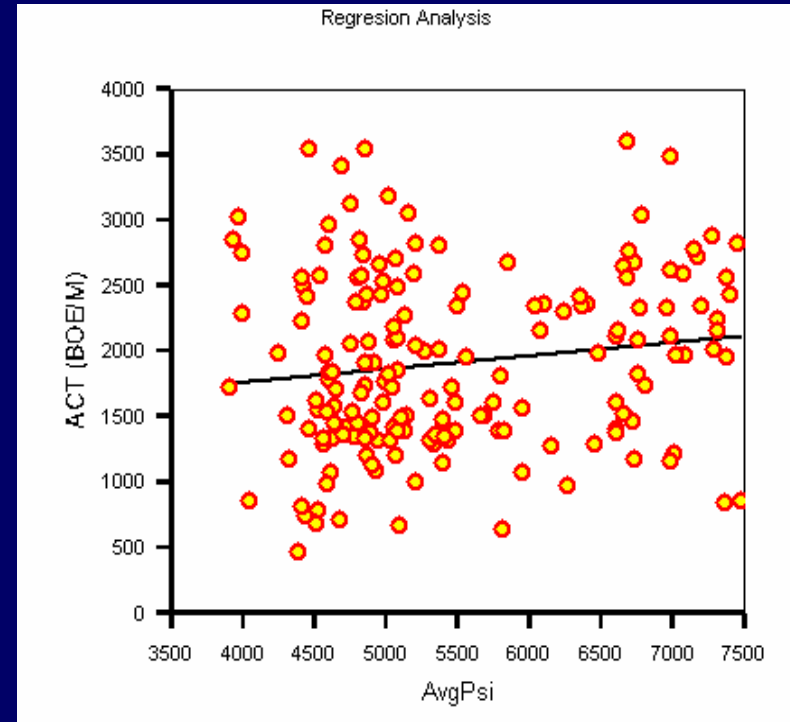
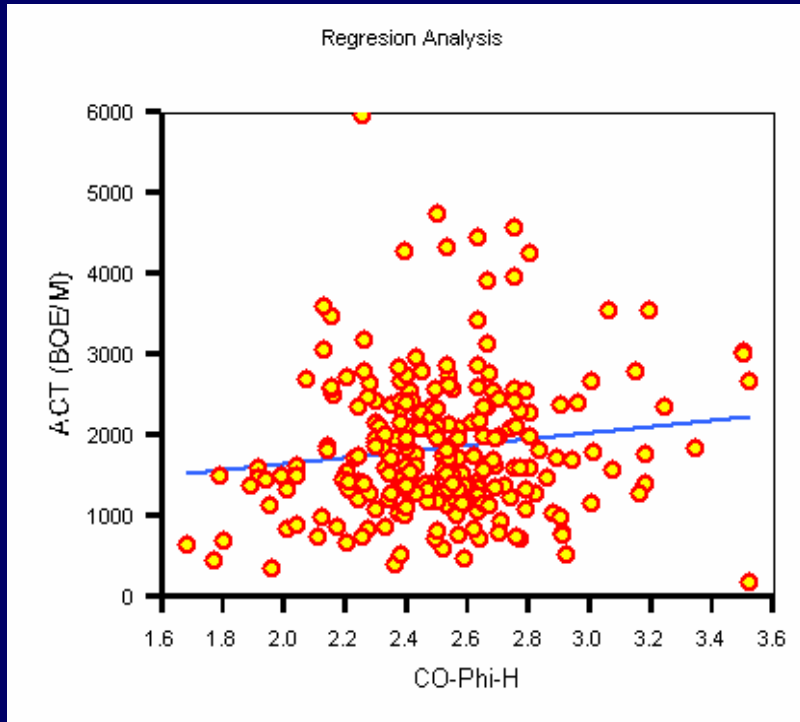
METHODOLOGY

□ Regression plots resemble shotgun.



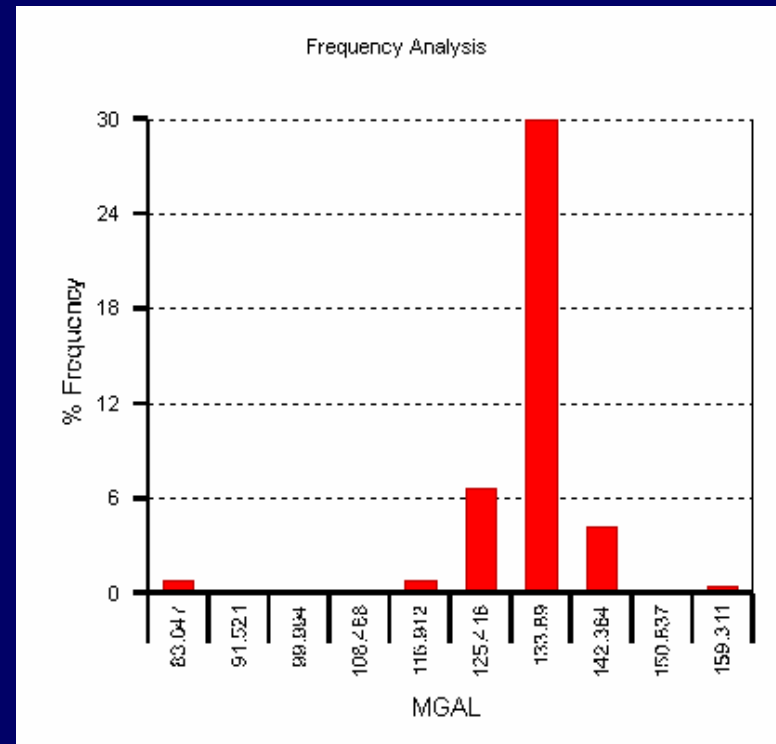
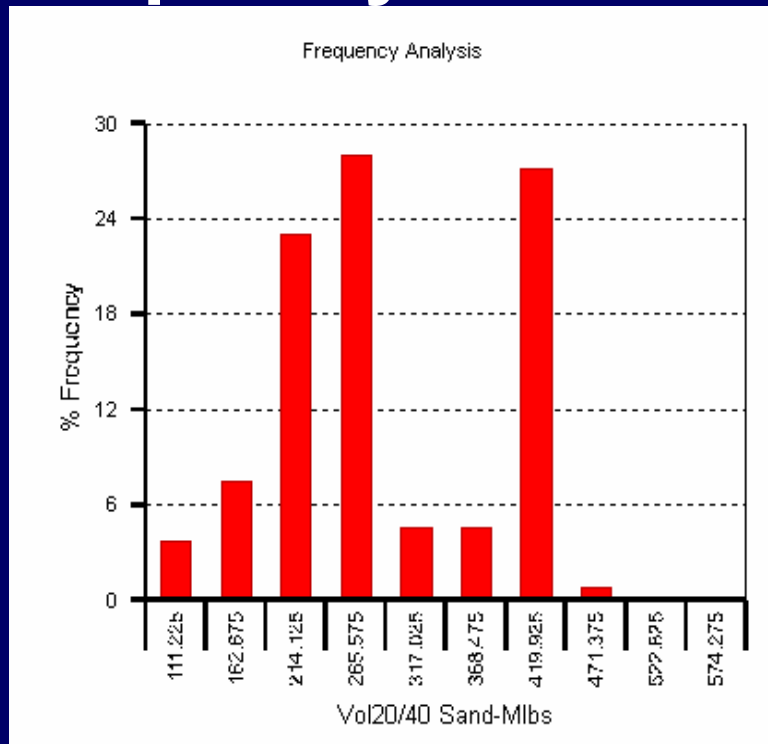
METHODOLOGY

□ Regression plots resemble shotgun.



METHODOLOGY

□ Frequency distribution have no trend.





METHODOLOGY

- **Again, unable to train a Neural Network Model.**
 - **Several types of neural networks.**
 - **Different neural network architectures.**
 - **Combinations of parameters.**
 - **Best results:**
Correlation coefficient = 0.17
 $R^2 = 0.07$



METHODOLOGY

- We concluded that Data can not be trained in its present form!**
- Intelligent system approach was considered for identification and clean up of the erroneous records in the dataset.**



METHODOLOGY

□ Neuro-Cluster Data Classification System

- Integrates clustering techniques, neural network modeling, and iterative process to achieve convergence.
- Separates data into 3 categories:
 - Good data
 - Slightly contaminated data
 - “Bad” data
- Uses most influential parameters.



METHODOLOGY

- 1. Cluster data using output of system.**
- 2. Train ANN using cluster info (membership functions and entropy).**
- 3. Iterative process.**
- 4. Repeat step 3 for each record in dataset.**



METHODOLOGY

- 1. Cluster data using output of system.**
 - Since the output is used during the cluster analysis, the information generated by this analysis carries the signature of the system behavior as a whole, input/output.**



METHODOLOGY

- 2. Train ANN using cluster info (membership functions and entropy).**
 - Results of step 1 is used during the network training. This means the system output is contributing to the input, albeit, indirectly.**
 - This constitutes a bootstrapping method where output is indirectly present in the system input.**



METHODOLOGY

3. Iterative process.

- Select a step for the range of output system.
- Sweep output range while firing neural network.
- For each value of assumed output:
 - Calculate cluster number for assumed value.
 - Calculate entropy.
 - Fire neural network.
 - Calculate error = | NN output – Assumed Output |
- Display all 3 curves on single graph.
- Identify position where data error reaching zero and entropy is minimum for constant cluster label representation.



METHODOLOGY

- **This methodology is based on the following hypothesis.**
 - **In a well behaved system, the output should be able to contribute to its own prediction and identification.**

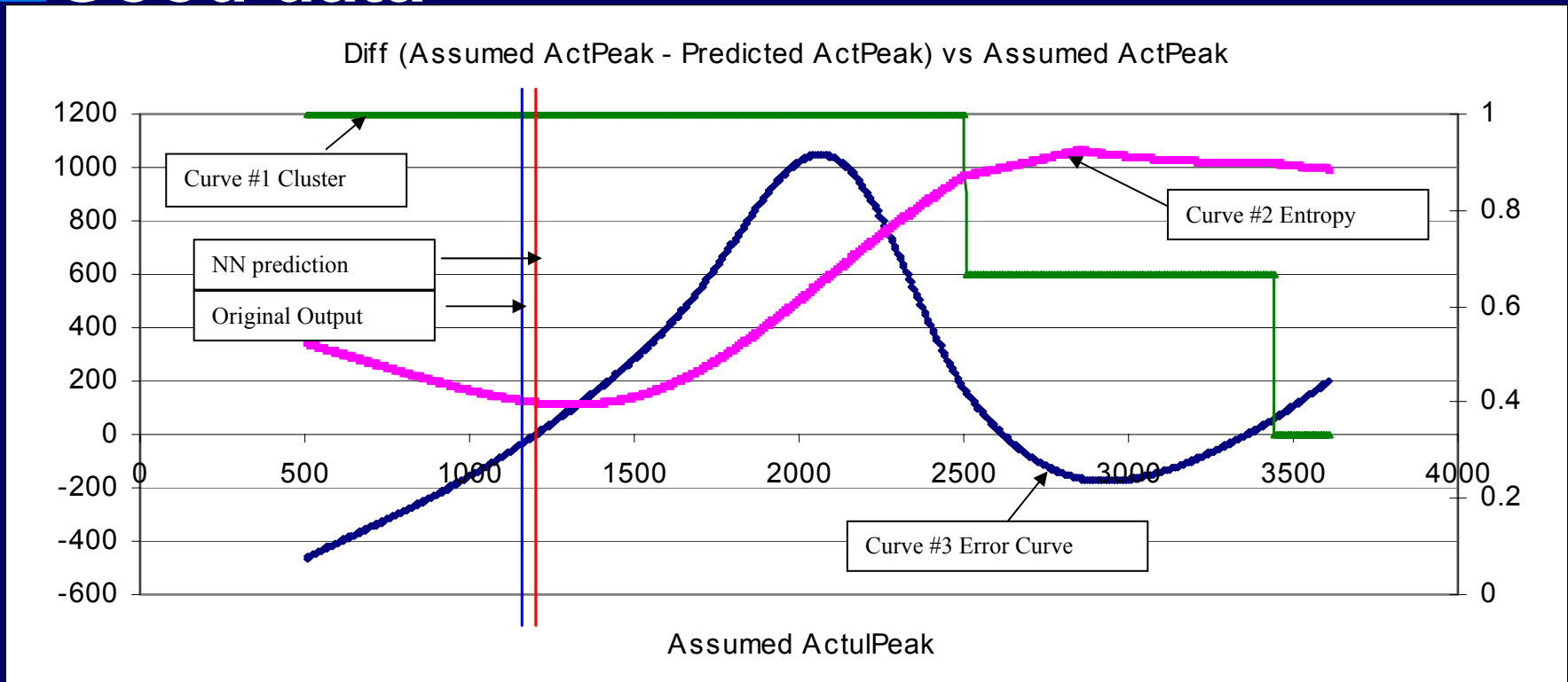


RESULTS & DISCUSSION

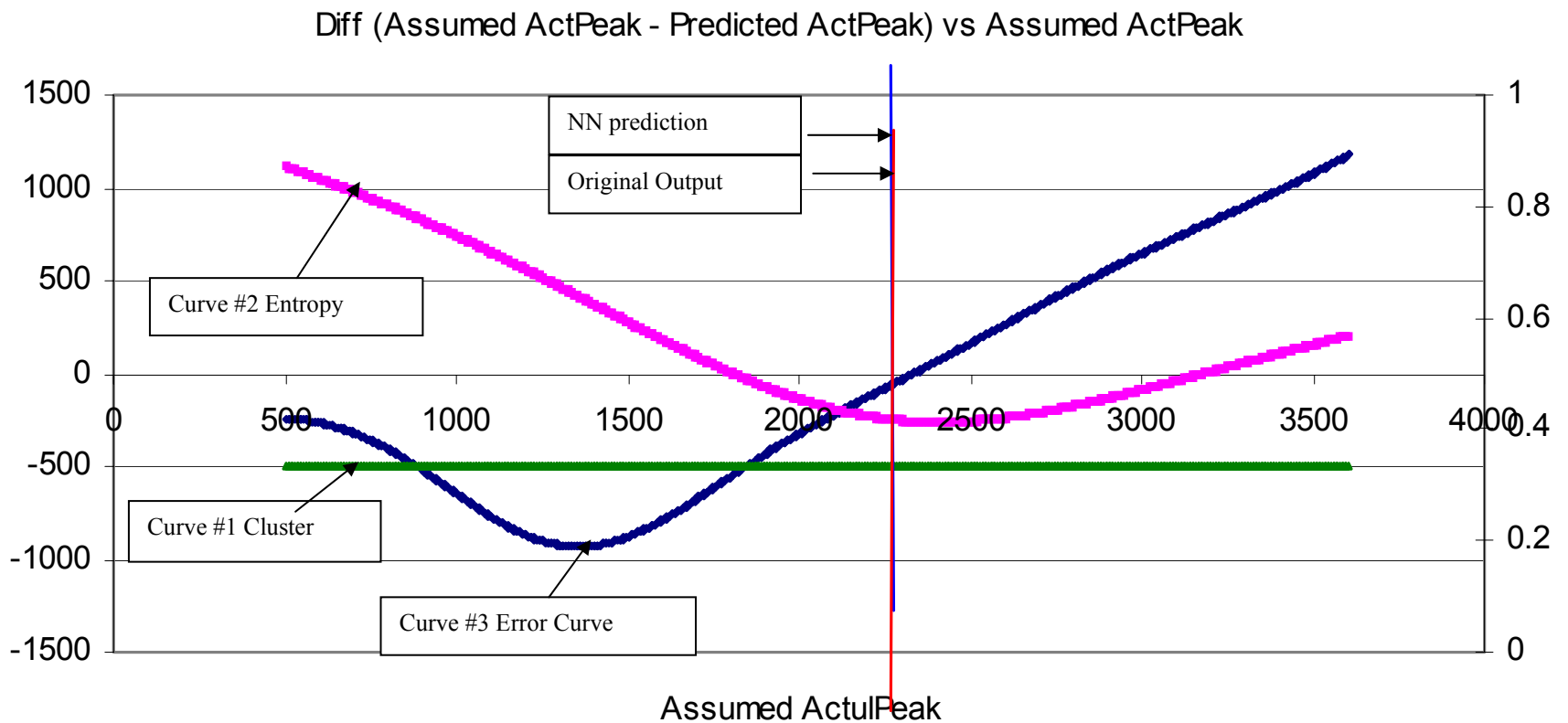
- ❑ **Output range for the output parameter during identification was between 500 – 2500 scf**
- ❑ **Three fuzzy clusters were considered**
- ❑ **Entropy was defined as the ratio between the smallest and largest membership function resulted after clustering. The range of the entropy was between 0 and 1.**
- ❑ **Error curve between – 1000 and +1500 scf.**

RESULTS & DISCUSSION

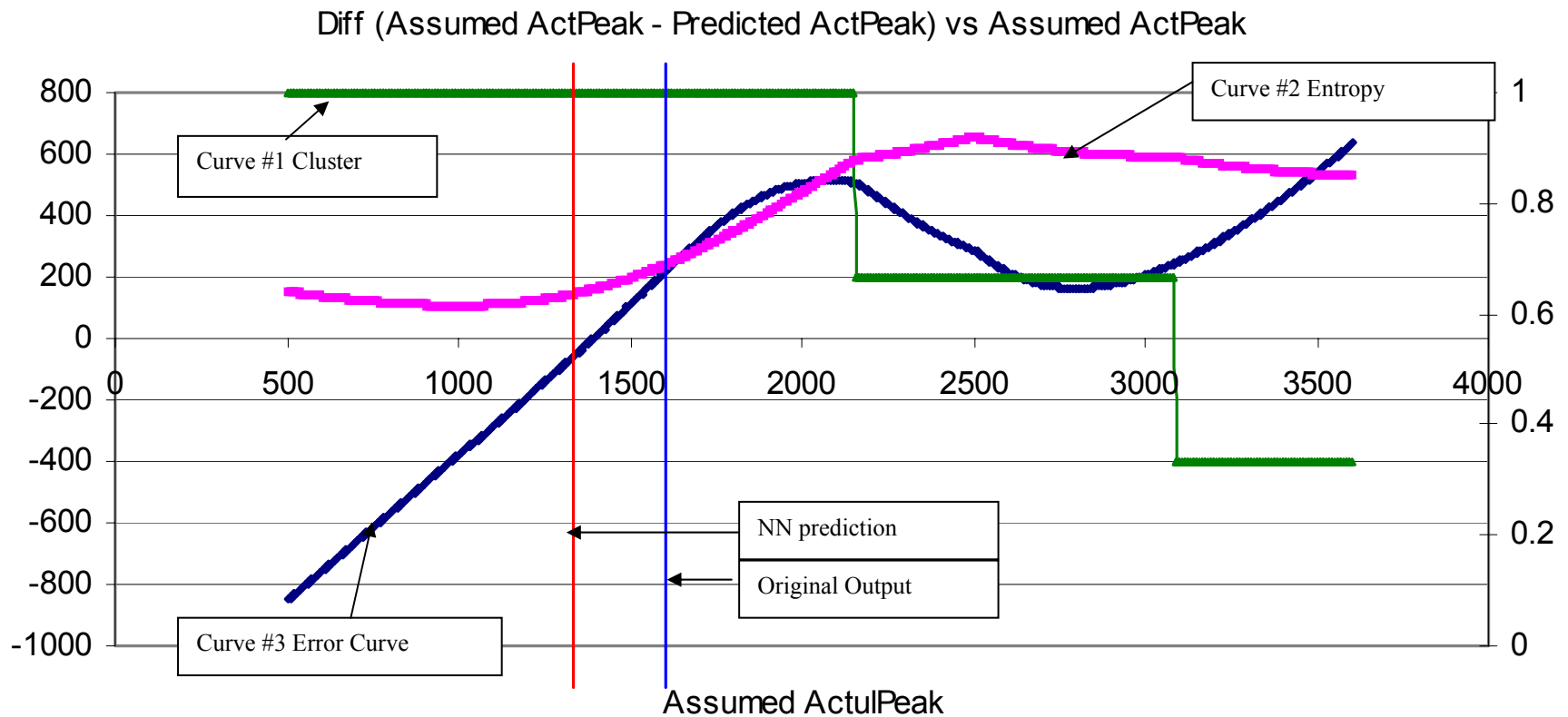
□ Good data



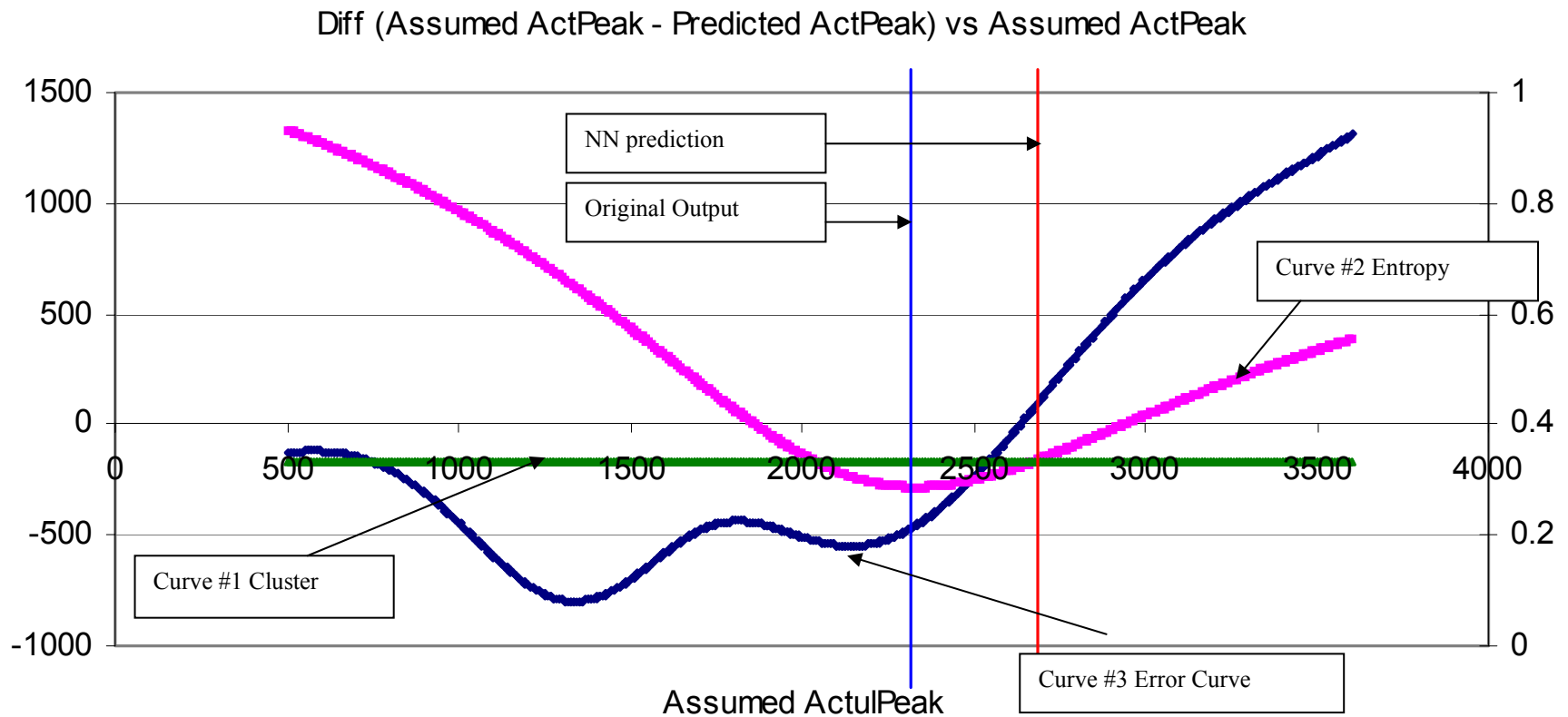
RESULTS & DISCUSSION



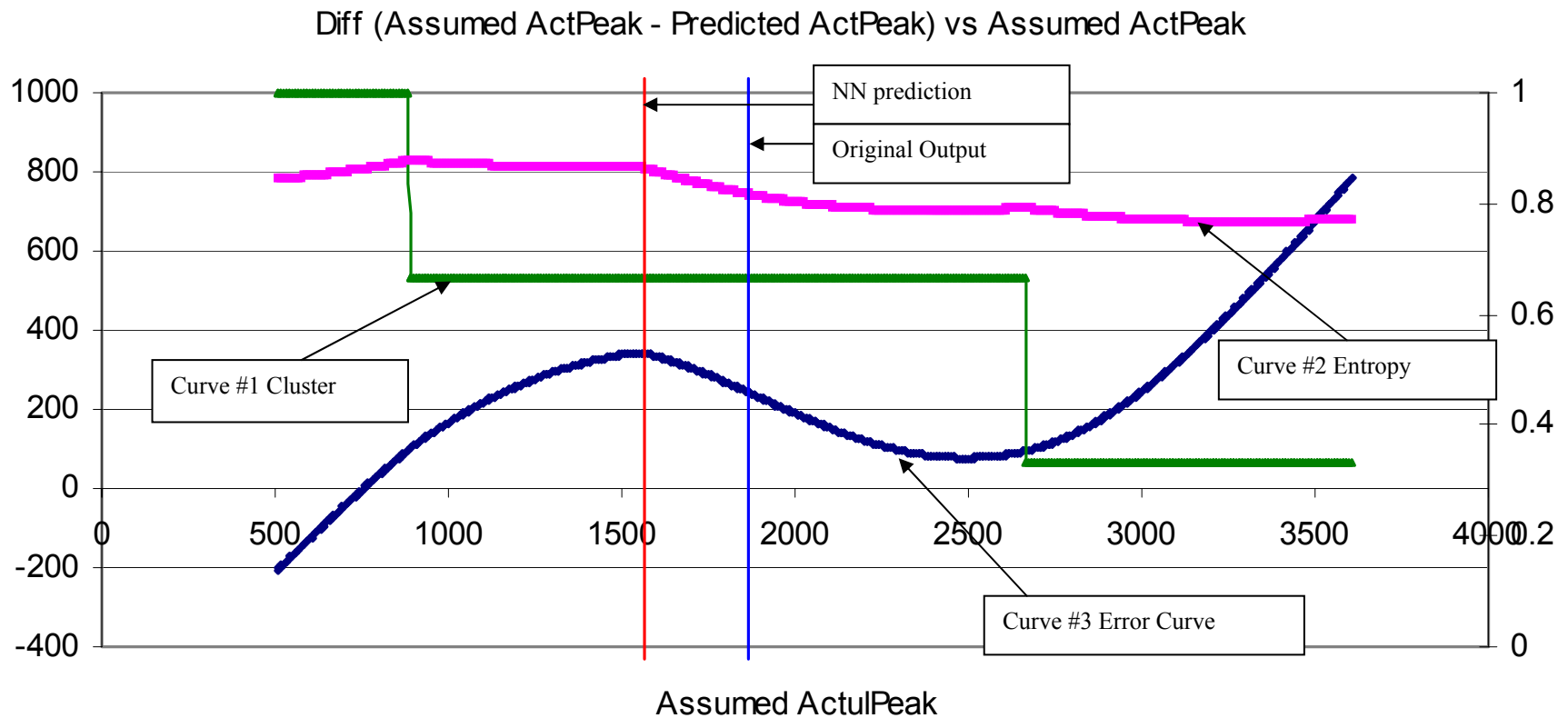
RESULTS & DISCUSSION



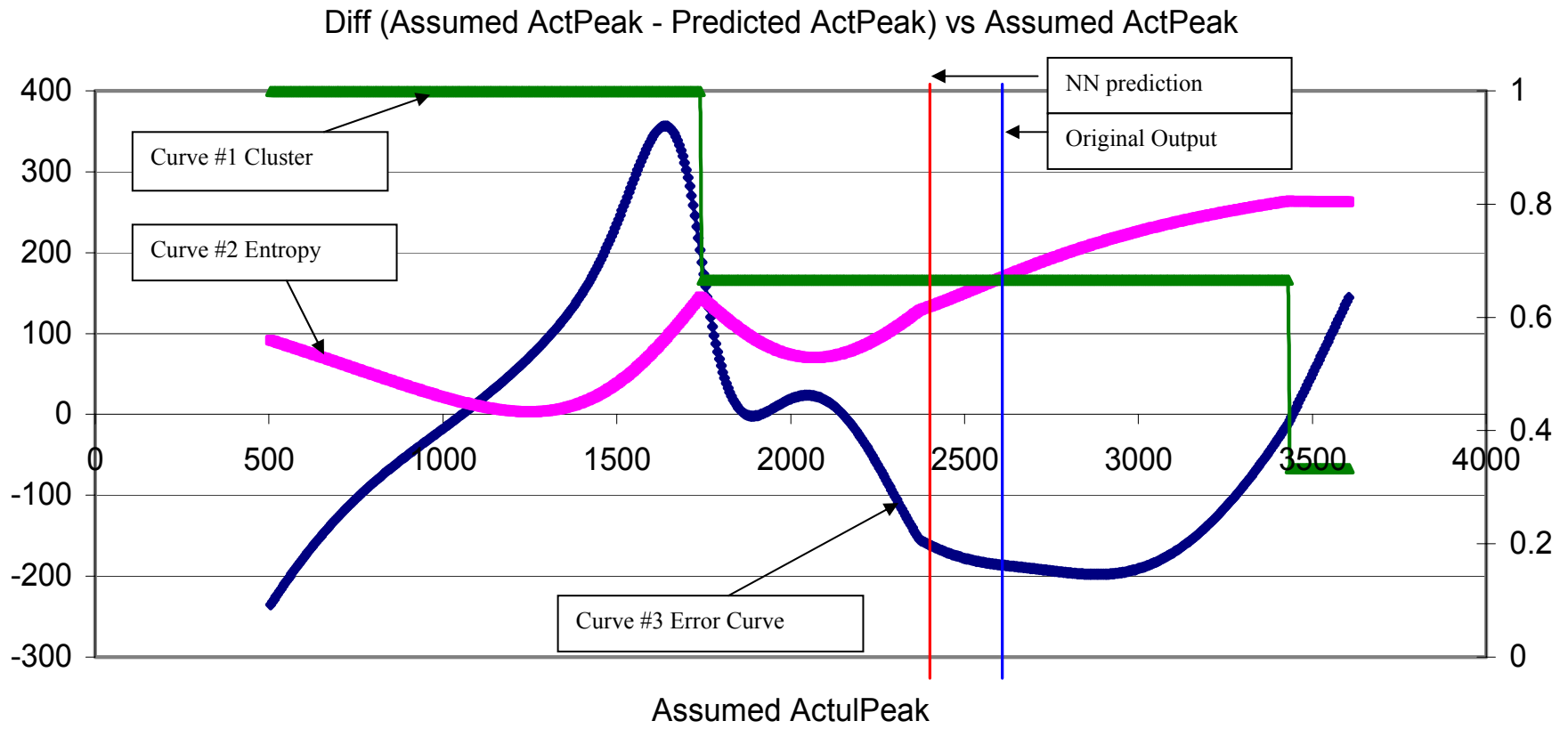
RESULTS & DISCUSSION



RESULTS & DISCUSSION



RESULTS & DISCUSSION





RESULTS & DISCUSSION

□ Data classified:

88 good

26 slightly contaminated

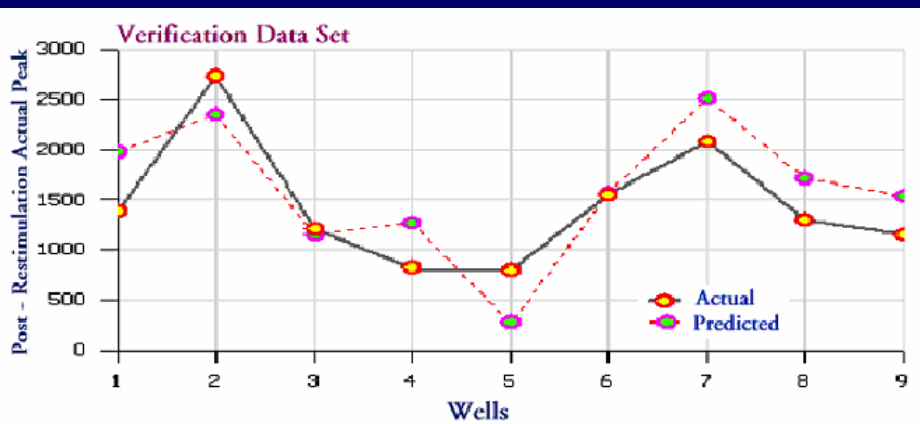
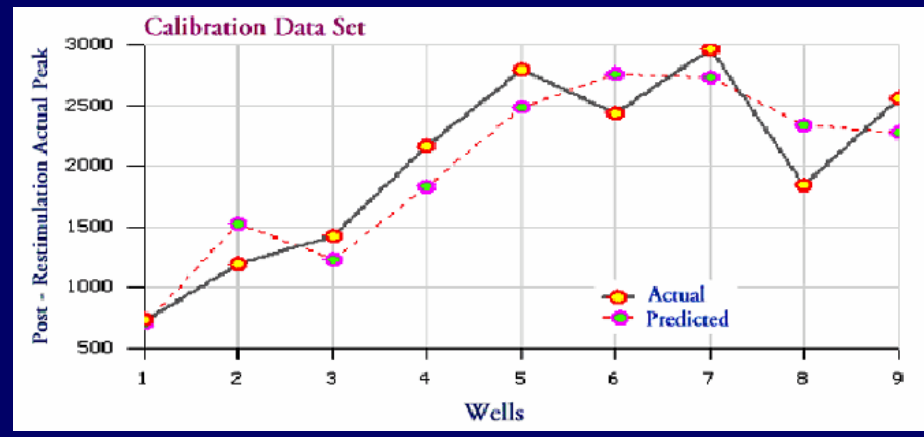
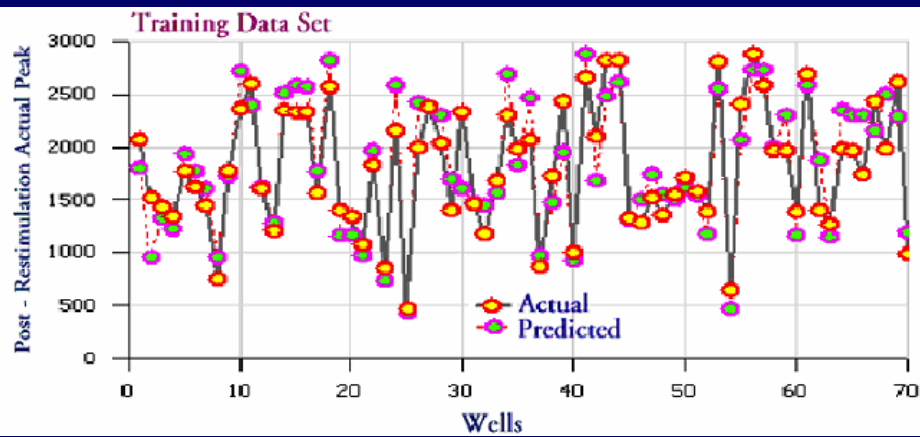
60 “bad”

□ Results:

Correlation coefficient = 0.876

$R^2 = 0.82$

RESULTS & DISCUSSION





CONCLUSIONS

- **A Neuro-Cluster Data Classification System was introduced to identify contaminated data.**
- **The combination of two intelligent tools, neural networks and fuzzy C-mean clustering provides a simple solution to problems like data classification, as presented in this paper.**



CONCLUSIONS

- **The applicability of this methodology was verified using a synthetic dataset developed using a commercial fracture simulator.**

- **The application of this methodology can be extended to any type of database for identification of contaminated data.**