A framework for oil well production data validation

Javier Herrera Vega[1], Felipe Orihuela-Espina[1], Eduardo F. Morales[1], and Luis Enrique Sucar[1]

National Institute for Astrophysics, Optics and Electronics, Puebla, Mexico,
{vega,f.orihuela-espina,emorales,esucar}@ccc.inaoep.mx

Abstract. Production data in petroleum engineering is often affected by errors occurring during data acquisition and recording. As interventions in the well alter the natural exponential decay of the production curve, the errors made during the data acquisition and recording are concealed. Automatic data validation techniques can help in cleaning production data. In this paper we propose solutions for three common problems that can be found in oil well production data; (i) detection of outliers in non-stationary signals, (ii) detection of sudden changes altering the natural trend of the signal and (iii) detection of rogue values disrupting signal trend in the light of statistically related variables. The solutions proposed make use of advanced computational solutions such as wavelets and Bayesian networks. The algorithms proposed are applied to an exemplary real well production dataset for illustration of the concepts.

Keywords: well modelling, Bayesian networks, wavelets

1 Introduction

Data validation is concerned with finding erroneous data in a time series and when appropriate, suggesting a plausible alternative value [11]. Data validation can be defined as a systematic process in which data is compared with a set of acceptance rules defining its validity. Often, the validation process is domain specific [6, 4]. In petroleum engineering causes for erroneous data include noise, sensor failures, data manipulation mistakes, etc.

In this paper we address three common types of errors that can be found in oil well production data, namely; (i) the detection of outliers in non-stationary signals, (ii) the detection of sudden changes altering the natural trend of the signal, and (iii) the detection of rogue values disrupting signal trend in the light of statistically related variables. For each of these problems we propose a generic solution and apply this solution to real production data.

Outliers are observations numerically distant from the rest of data. Surprisingly there is not a standard method for identifying them. Often data is assumed to comply with a Gaussian distribution and a distance criterion e.g. deviation, from the distribution descriptor determines the outlier condition of a data sample. Oil well production data is a non-stationary process, and thus the naive approach does not suffice. However, upon looking at a sample neighbourhood, stationarity can be assumed. Here we propose a local solution for outlier identification.

Atypical sudden changes deviating from the natural trend of the signal often correspond to noise, or failures in data recording. Noise in the context of oil well production can often be associated to well interventions. There already exist a number of approaches for the detection of sudden changes as for instance the use of the Laplacian of a Gaussian operator [7]. A discussion of existing approaches to detect signal discontinuities is beyond the scope of this paper. Here we use Haar wavelets for the detection of sudden changes in the signal proposing a variant from an existing approach developed for neuroimaging data [9].

The final validation problem addressed here is the detection of suspicious values which may be in range and agree with the signal trend but that contradict the trend in statistically dependent variables. In order to catch these rogue values we present an approach based on Bayesian networks. The use of a Bayesian network for validating data by related variables capitalises on the following idea; the trend of statistically related variables must grossly follow each other. When this premise is violated, the observation is likely to be a rogue value.

The paper first introduces the technique to address each of the validation challenges in Section 2. Then, in Section 3 the proposed techniques are applied to exemplary oil production data from a real well. Finally, the paper closes summarizing the conclusions in Section 4.

2 The validation framework: Techniques

2.1 Local outlier detection

Perhaps the easiest form of outlier detection consist of imposing a valid data range within which variable data is allowed, and labelling values outside the range as *outliers*. Often this range is established from the data distribution as defined by equations 1 and 2 :

$$\text{lower limit} = m - 3\sigma_m \tag{1}$$

$$\text{upper limit} = m + 3\sigma_m \tag{2}$$

where m is the distribution median and σ_m is the deviation from the median.

If stationarity does not hold, the above solution is not satisfactory. Notwithstanding, upon accepting that the decay of the oil well production curve is slow, local stationarity holds and the above solution can be reused. A local outlier detection can be constructed upon windowing the data. The basic idea is then to shift the window along the data and compute the lower and upper limits of the data range only for the visible data within the window.

The question that remains is how large should the window be. The answer depends on how quickly the signal change, and we make here no attempt at providing an automatic window size determination. Instead this remains a free parameter of the method.

2.2 Abrupt change detection

The wavelet transform [1] decomposes a signal in its time-scale components re-expressing the original function in terms of the wavelet family basis. The continuous wavelet transform (CWT) of a signal $x(t)$ is defined by equation 3.

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \Psi^{*}_{a,b}(\frac{t-b}{a}) dt \tag{3}$$

where a and b are the scale and shift parameters respectively, and $\Psi(t)$ is the wavelet function used for the decomposition of $x(t)$. Among the wavelets functions families, Haar wavelets [1, 9] are especially suitable for the detection of discontinuities.

For each a and b i.e. time-scale pair, a wavelet coefficient captures the similarity if the signal $x(t)$ and an stretched and shifted version of $\Psi(t)$. It is from these coefficients that it is possible to discriminate sudden changes in the signal. Application of the median filter to the coefficients independently at each scale emphasises the characteristics of the sudden changes as well as reduces the impact of white noise. To establish the limit between acceptable changes and inappropriate changes a threshold T is imposed in the matrix of coefficients. Here we chose to set the threshold automatically using the *Universal Threshold* [3] according to equation 4:

$$T = \sigma \cdot \sqrt{2 \cdot \ln n} \tag{4}$$

where σ is the absolute deviation over the median and n is the number of coefficients.

2.3 Rogue values detection with related variables

A Bayesian network is a probabilistic graphical model in which domain entities form the nodes of a directed graph and the conditional dependence assumptions between the variables are represented in the arcs.

Before validation can take place the Bayesian network must be constructed in a process known as structure learning. For this and abstracting of the training algorithm details, the joint probability distribution of each pair of variables must be found. The discretization of the variables ranges into a set of discrete intervals facilitates the determination of such joint probability distribution and paves the way for inference. The determination of the discrete intervals is a non-trivial problem and some common approaches include [5]; *equi-distance* where the variable data range is split in a predetermined number of equally distant intervals, and *equi-frequency* where the splitting of the intervals ensure that each interval holds the same number of samples. Here we propose a more sophisticated interval discretization approach based on a Gaussian mixture model. Under this interval discretization approach the data is assumed to be generated by a mixture of Gaussian distributions, each one characterized by its mean μ and its variance σ^2 as illustrated in Figure 1. The model is formally defined by equation 5.

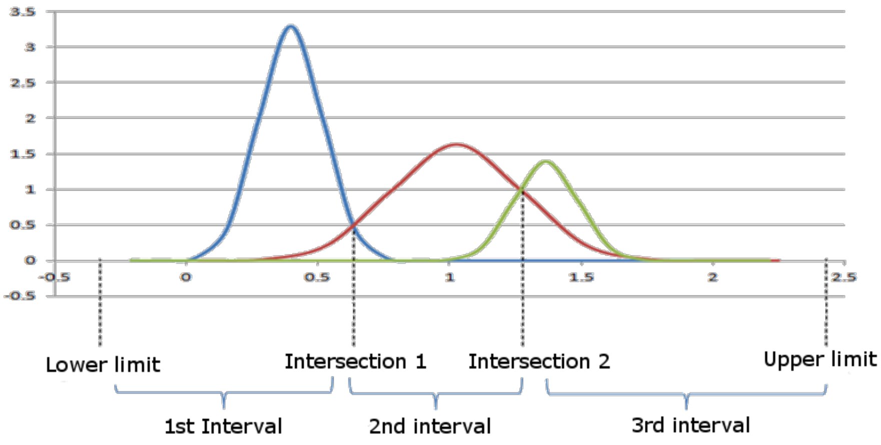


Fig. 1: Exemplification of the interval discretization by 3 Guassian distributions.

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sigma_k^2) \qquad (5)$$

where K is the number of Gaussians considered, $N(x|\mu_k, \sigma_k^2)$ represents a Gaussian with mean μ_k and variance σ_k^2 and π_k are the mixing coefficients, i.e. weights for the Gaussians. The algorithm has K as a single parameter, and the classical Expectation-Maximization algorithm [2] is used to optimize the fitting of the distributions.

After the intervals have been established for every variable in the model, the one being validated and the related variables, a Bayesian network is learned using a training algorithm, e.g. PC [10].

The learned Bayesian network capturing relations among the variables permits identification of rogue values using a two step process [8]:

1. **Identification of error candidates**: Every node (variable) in the net is analyzed, instantiating its *Markov Blanket (MB)* (the parents, the children and the parents of his children) and propagating evidence on the net. The posterior probability distribution of the variable is obtained and the probability of the analyzed value is determined to be a corrupted data if its probability is less than a threshold (p_value). A set of *possible errors* is obtained after all nodes have been analyzed.

2. **Isolation of real errors**: A new causal network is built with two levels. Both levels contains all the nodes in the original network. The upper level represents true errors in the given variable and the lower level represent possible errors. The relation between upper and lower nodes is given by the Extended Markov Blanket of each node [8]. Every node in lower layer is instantiated as true if this is a possible error (detected in the previous phase), the rest of the nodes is instantiated as false. With this evidence propagation is performed over the network and posterior probability distribution for every real node is read. If the true state of a real node is higher than a given threshold (pF) then we conclude that this variable represent corrupted data given the evidence in his relational variables.

3 Results and discussion

For results illustration purpose, data from an oil well from the Jujo field in Mexico has been used. The data contains three time series; Oil_Net, $Water\_Net$ and RGA describing the oil, water and gas production respectively, the latter express as the oil to gas relation.

3.1 Local outlier detection

For the purpose of this paper, examples were tested with a window worth 10% of samples. Figure 2 illustrates the difference in applying global and local outlier detection to a real oil well production curve.

3.2 Abrupt change detection

Figure 3 shows the matrix of Haar wavelets coefficients.

Figure 4 summarises the results from detecting sudden changes using the Haar wavelets with universal threshold in a given production data series. It can be appreciated how the method nicely agrees with the intuitive visual inspection of the data.

3.3 Rogue values detection with related variables

A Bayesian Network has been constructed to model the relations between production's variables: *Oil_Net*, *Water_Net* and *RGA* (Gas-Oil relation). We have chosen a randomly pick subset of 20% of samples of production data constraint to those free of other errors for training. Figure 5 summarise the intervals found for the production variable *Oil_Net* using the Gaussian mixture model based discretization. Analogous intervals were also calculated for $Water\_Net$ and RGA. Figure 6 shows the learned Bayesian network and dependencies found between the three variables. The 80% rest of the production data were validated using this model. Detection thresholds were set as: $p\_value = 0.01$ and $pF = 0.7$. These thresholds values were chosen based upon experience of the researcher. Figure 7 illustrates an example of the detection process.

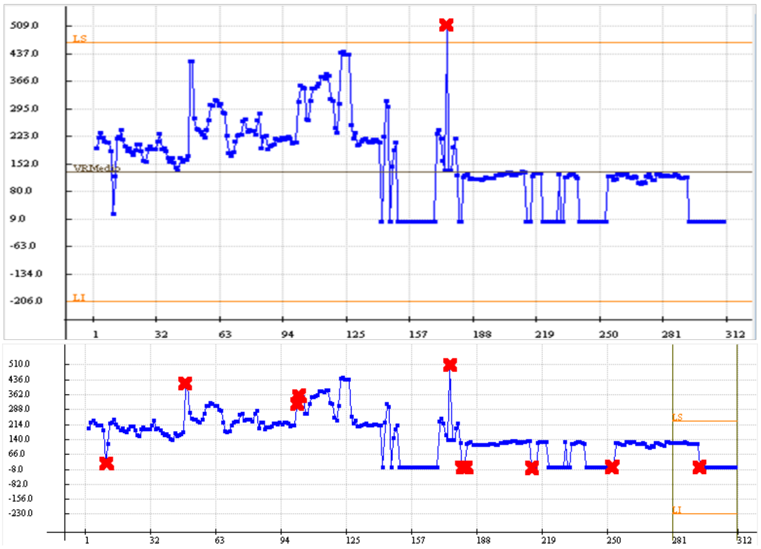


Fig. 2: Global (top) and local (bottom) approaches to outlier detection. The orange lines indicate the upper and lower limits respectively (whether global or local). Red crosses correspond to samples marked as outliers. For the local outlier detection a window sized 10% of the signal length was chosen.

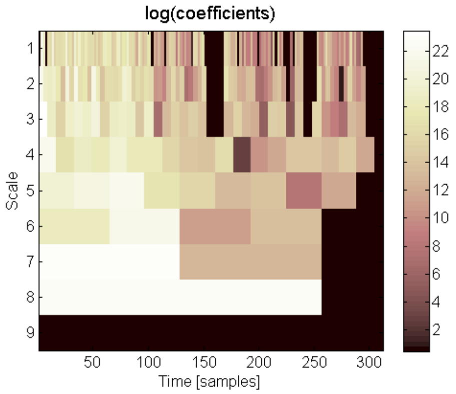


Fig. 3: The matrix of wavelet coefficients, represented in logarithmic scale for better appreciation.

4 Conclusions

We presented a framework for oil well production data validation underpinned by advanced computational solutions such as wavelets, Bayesian networks and

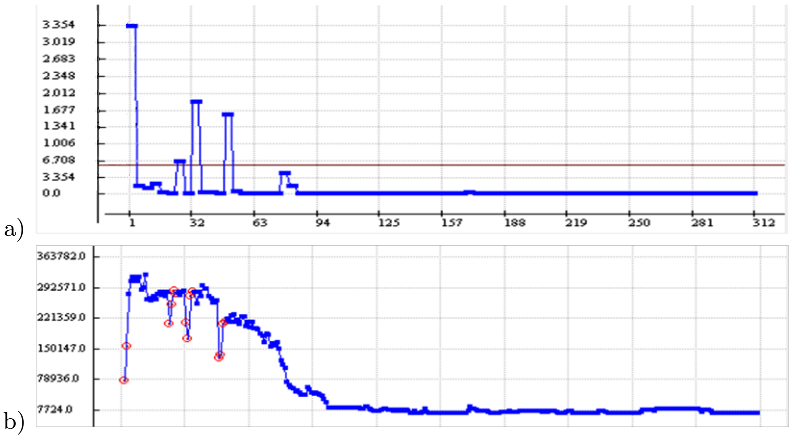
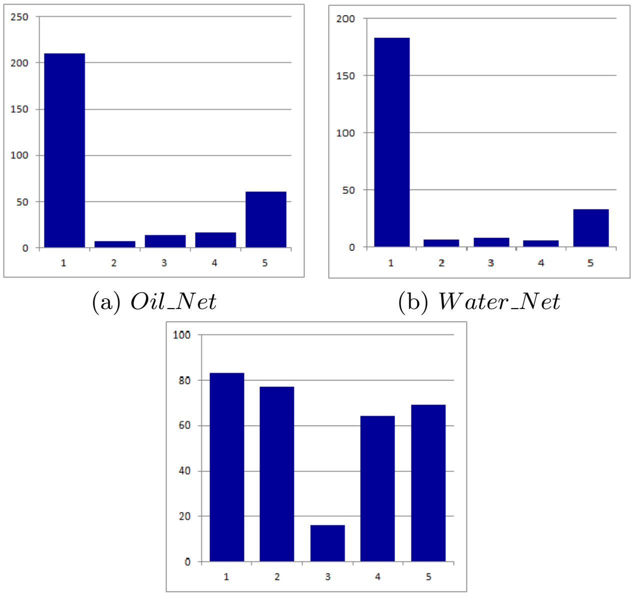


Fig. 4: a): The wavelets coefficients at the discrete scale (2) smoothed with the median filter and the threshold selected to discriminate sudden changes. b): The oil well production data signal with the sudden changes as captured with the Haar wavelets approach afore described.

Gaussian mixture models. Although the application is for petroleum engineering the methods are generic and thus generalizable to other domains. The current work has focus in the detection of errors but has obviated the suggestion of alternative values for the suspicious samples. While of course interpolation is a clear candidate, in the presence of statistically dependent information the use of Bayesian networks may surpass the interpolation capabilities for suggesting plausible values. We are now investigating this potential use of Bayesian networks and preliminary results are promising.

References

1. Daubechies, I. Ten lectures on wavelets. SIAM: Society for Industrial and Applied Mathematics; 1st edition, 377 pgs (1992)
2. Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39(1), pp. 1–38 (1977)
3. Donoho, D. L.; Johnstone, I. M. Ideal spatial adaptation by wavelet shrinkage. Biometrika 81, 425–455 (1994).
4. Gonzalez, R.; Huang, B.; Xu, F. ; Espejo, A. Dynamic bayesian approach to gross error detection and compensation with application toward an oil sands process. Chemical Engineering Science 67, 44–56 (2012)
5. Kotsiantis, S.; D. Kanellopoulos Discretization techniques: A recent survey GESTS International Transactions on Computer Science and Engineering 32(1), 47–58 (2006)



(a) *Oil_Net* (b) *Water_Net*

(c) *RGA*

Fig. 5: Intervals detected by the discretization with Gaussian mixture model interval discretization approach for the data in Figure 4 (b). An arbitrary 5 intervals were chosen

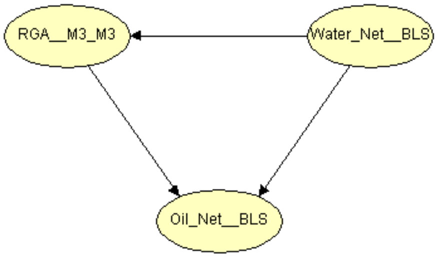
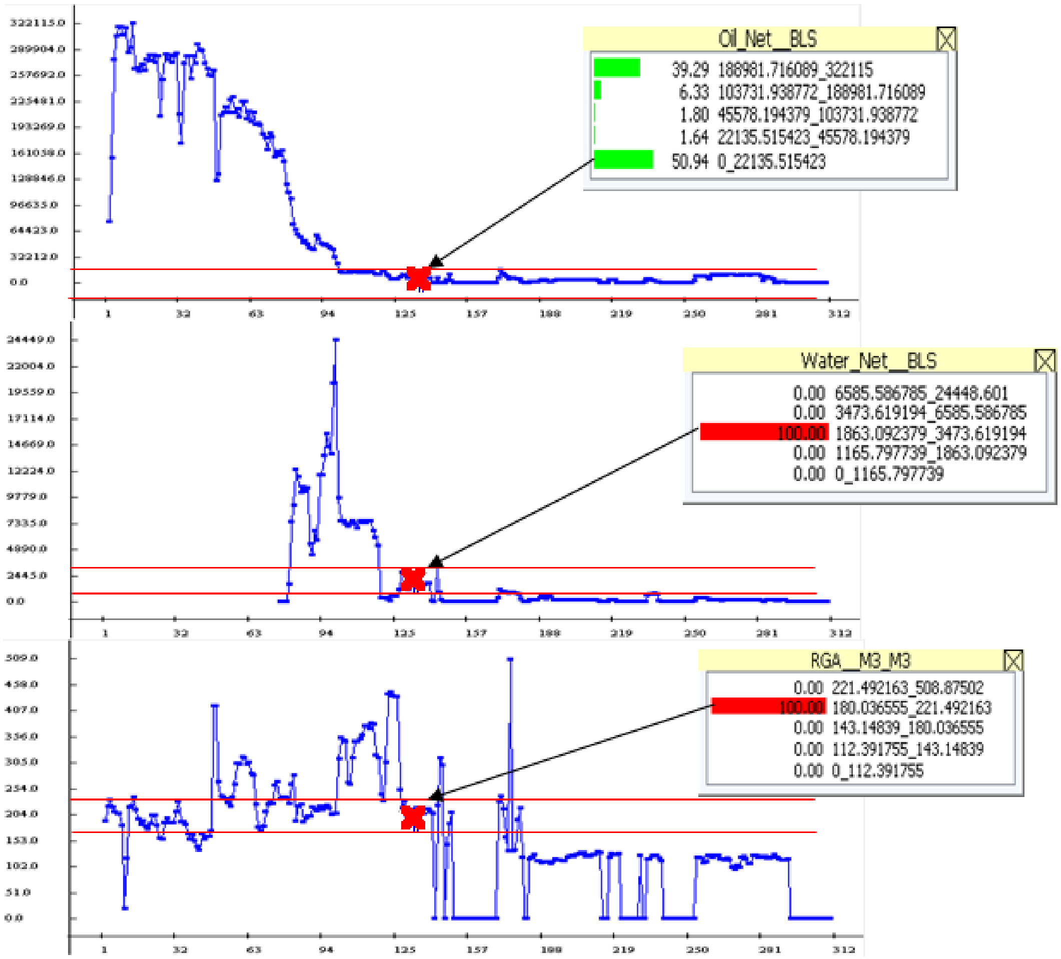


Fig. 6: Bayesian network capturing the relation between three oil engineering variables of the production process.

6. Lamrini, B.; Lakhal, El-K.; Lann, M-V.; Wehenkel, L. Data validation and missing data reconstruction using self-organizing map for water treatment. Neural Computing and Applications 20, 575–588 (2011)

7. Marr, D.; Hildreth, E. Theory of edge detection. Proceedings of the Royal Society London B 207, 187–217 (1980)



Oil_Net__BLS

🟩	39.29	188981.716089_322115
🟩	6.33	103731.938772_188981.716089
	1.80	45578.194379_103731.938772
	1.64	22135.515423_45578.194379
🟩	50.94	0_22135.515423

Water_Net__BLS

0.00	6585.586785_24448.601
0.00	3473.619194_6585.586785
🟥 100.00	1863.092379_3473.619194
0.00	1165.797739_1863.092379
0.00	0_1165.797739

RGA__M3_M3

0.00	221.492163_508.87502
🟥 100.00	180.036555_221.492163
0.00	143.14839_180.036555
0.00	112.391755_143.14839
0.00	0_112.391755

Fig. 7: Timecourses of production variables *Oil_Net*, *Water_Net* and *RGA* The red cross points the zoomed data for exemplification. In *Water_Net* and *RGA* the red labelled interval indicated the fixated value. In *Oil_Net* the lower green squares indicate a probability above *p_value*. In this case, the validated value is considered correct.

8. Ibargüengoytia, P; Vadera, S.; Sucar, L.E. A probabilistic model for information validation. British Computer Journal 49(1), 113–126 (2006)
9. Sato, H.; Tanaka, N.; Uchida, M; Hirabayashi, Y; Kanai, M.; Ashida, T.; Konishi, I.; Maki, A. Wavelet analysis for detecting body-movement artifacts in optical topography signals. NeuroImage 33, 580–587 (2006)
10. Spirtes, P.; Glymour, C.; Scheines, R. Causation, Prediction, and Search. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Massachusetts, 2nd edition, 2000
11. Tamrapani, D.; Johnson, T. Exploratory Data Mining and Data Cleaning. John Wiley, Hoboken, NJ, 1st Edition. 224 pgs (2003)