Structured Noise Detection: Application on Well Test Pressure Derivative Data

Farhan Asif Chowdhury
The University of New Mexico
Albuquerque, New Mexico
fasifchowdhury@unm.edu

Satomi Suzuki
ExxonMobil Upstream Integrated Solutions
Spring, Texas
satomi.suzuki@exxonmobil.com

Abdullah Mueen
The University of New Mexico
Albuquerque, New Mexico
mueen@unm.edu

ABSTRACT

Real-valued data sequences are often affected by *structured noise* in addition to random noise. For example, in pressure transient analysis (PTA), semi-log derivatives of log-log diagnostic plots show such contamination of structured noise; especially under multi-phase flow condition. In PTA data, structured noise refers to the response to some physical phenomena which is not originated at the reservoir, such as fluid segregation in wellbore or pressure leak due to a brief opening of a valve. Such noisy responses commonly appear to mix up with flow regimes, hindering further reservoir flow analysis.

In this paper, we use the Singular Spectrum Analysis (SSA) to decompose PTA data into additive components; subsequently we use the eigenvalues associated with the decomposed components to identify the components that contain most of the structured noise information. We develop a semi-supervised process that requires minimal expert supervision in tuning the solitary parameter of our algorithm using only one pressure buildup scenario. An empirical evaluation using real pressure data from oil and gas wells shows that our approach can detect a multitude of structured noise with 74.25% accuracy.

CCS CONCEPTS

• **Information systems** → **Data cleaning**; *Process control systems*; • **Computing methodologies** → *Modeling methodologies*;

KEYWORDS

Structured Noise Detection, Data Cleaning, Process Automation

ACM Reference Format:
Farhan Asif Chowdhury, Satomi Suzuki, and Abdullah Mueen. 2019. Structured Noise Detection: Application on Well Test Pressure Derivative Data. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330661

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330661

1 INTRODUCTION

Sensor data collection and analysis have become ubiquitous in production and manufacturing operations for continuous surveillance and monitoring [16]. These sensors gather and record data at a regular interval, which in turn, produces a vast amount of data, mostly in the form of time series [13]. Using data mining techniques, this massive data trove can be potentially transformed into explicit actionable knowledge. However, these datasets are often contaminated with noise due to inefficient calibration, measurement error, external interruption, etc. and this presence of noise is the main hindrance to further automated analysis and decision-making process.

In the oil and gas industry, permanent downhole pressure gauges are installed in wells drilled in oil and gas fields to monitor production [15]. The deployed pressure gauges continuously record pressure in well at regular time intervals producing a time series of the downhole pressure. This time series is utilized in *pressure transient analysis (PTA)* to determine reservoir and well characteristics [5], to continuously asses reservoir and well condition and to forecast about future production performance. The log-log pressure derivative plot during the shut-in period of well is widely used in PTA for well property evaluation [2].

Identifying different flow regimes in log-log pressure derivative data during the shut-in period reveals characteristics of the *hydrocarbon-bearing formation*, i.e., reservoir, and condition of the well. Currently, the task of identifying the flow regime is mostly being done by manual observation, although some automated PTA methods have been developed recently [6, 22, 24]. One of the main impediment of fully automated PTA is that flow regime identification in log-log pressure derivative data is misled due to the presence of *structured noise*.

There is no unified definition of noise in general; instead, it is highly domain and problem specific as they relate to different events in different observation and measurement systems. We use the terminology *structured noise*, in the context of pressure data recorded by downhole gauge, to distinguish the usual pressure response from the pressure response to *non-reservoir origin physical phenomenon*, for example, fluid segregation in wellbore or pressure leak due to a brief opening of a valve. Mostly they are deviations from the usual response for a brief amount of time positions in the original data affecting the overall trend. These structured noises often maintain a similar pattern between different observation which implies that there must be a common underlying physical event that may have generated them.

In Figure 1, we show two pressure derivative data, often also referred to as pressure buildup (PBU) data. In Figure 1(left), the PBU

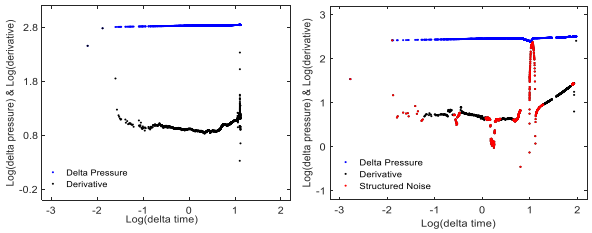

Figure 1: (left) An ideal pressure derivative data without any structured noise; (right) A structured noise corrupted pressure derivative data where red labeling indicates the structured noise (both data from Well-A [4]).

data is free from structured noise, and in Figure 1(right), the PBU data is contaminated by structured noises which are marked in red (automatically labeled using our method, showing an accurate detection case-study). These structured noises often mix-up with flow regimes; and in those cases flow regime identification algorithms generate erroneous result preventing further down the road analysis. Hence, an efficient and robust structured noise detection algorithm will open up new frontiers of data mining application on PBU data. In this paper, we propose a structured noise detection method for well test pressure derivative data using *singular spectrum analysis (SSA)*.

1.1 Why the problem is challenging?

The pressure derivative data is non-stationary, has variable duration and amplitude offset; thus the use of Fourier Transform based filtering techniques are not feasible. The associated random noise in the log-log pressure data is easier to filter out as they pose significantly different spectral property compared to the noise-free signal. However, structured noises occur for a brief interval of time while maintaining a contiguous shape; thus they appear as a part of the original signal. Again, their frequency spectrum is wide, often overlaps with the spectrum of the noise-free signal. Moreover, structured noise segments have variable length; their shape and statistical properties vary in between different data observation, which obsoletes the use of segment-based classification approach. Wavelet is a powerful spectral filtering tool which produces acceptable results in many a similar scenario. One main drawback of Wavelet analysis is that it requires the manual selection of basis function on which the filtering quality depends a lot. As our data has a high level of variation from one observation to another as well as one well to another, one selected fixed basis is not capable of working optimally for all the variations. A more robust method is required that can adapt to the variation of data. Moreover, our objective is not only to filter out but rather detect and localize structured noise segments in the noisy observation, which poses new challenges.

1.2 Related Works

Extensive studies have been carried out to perform noise removal on pressure data due to its usability in evaluating reservoir characteristics. Most related studies have performed noise removal on the original pressure data before calculating the derivative. But, removal of noise before calculating the derivative often time smooths out finer details. Also, noise is more pronounced in derivative data which requires more rigorous noise removal techniques. To the best of our knowledge, this is the *first attempt* to detect *structured noise in pressure derivative data*.

Noise reduction in the original pressure data is usually done using spectral methods and regression analysis. Wavelet transform has been widely used to reduce noise [23]. Some other common filtering methods are Butterworth, Locally Weighted Scatterplot Smooth (LOESS), Auto-regressive Moving Average (ARMA) [21]. In [19], they have shown a comparative evaluation of the aforementioned denoising methods for well pressure data. Among all these, wavelet performs better. But as mentioned earlier, one of the main challenges in applying wavelet is that we need to select the basis for wavelets manually and the performance of the noise detection varies depending on the choice of the basis. As a result of this basis dependency, it fails to adapt to data from different well types associated with a diverse range of structured noise.

1.3 Our Approach

In this paper, we propose a method where we use singular spectrum analysis (SSA) [3] to detect structured noise segments in real oil and gas well log-log pressure derivative data to automate the pressure transient analysis process. The SSA is a non-parametric time series analysis tool and does not make a prior assumption about the data. There are hyper-parameters involved to be adapted to the problem in hand. It has been used in a multitude of problem domains. SSA has been mainly used for time-series modeling [26], structural change detection [20], forecasting [10]. Recently it is being widely used in biomedical signal processing for noise and artifacts removal [7, 17, 18, 25].

SSA decomposes a signal into multiple additive components, which usually can be interpreted as the trend components, amplitude and phase modulated oscillatory components and the unstructured random noise components [11]. We use SSA to decompose our signal of interest into such components and we identify the relevant components that capture most of the information about the structured noise present in the data. As structured noise poses oscillatory behavior, it should be projected onto the decomposed oscillatory components and we use the eigenvalues associated with the decomposed components to identify these components. And, once we identify the relevant components, we localize the structured noise segments present in those components using a threshold value. To select the threshold value, we employ a *single-sample learning method* whenever we use our algorithm on new well data. After detecting the structured noise, we perform further boundary refinement for precise structured noise localization.

Experimental evaluation using real pressure data from oil and gas wells shows that our approach can identify a multitude of structured noise with 74.25% accuracy. Beside, the algorithm is invariant to the amount and types of noise. The algorithm is *generalizable* to structured noise detection in other kinds of data. By employing single sample learning to fine tune the threshold parameter each time we apply our method to new well data, we ensure that our method can adapt to a multitude of structured noise types which show large variation from well to well. Moreover, the algorithmic computation need not rerun for each threshold value change. In

Figure 1, we have shown an example of our noise detection result where the left figure shows accurate zero noise detection in the absence of noise and the right figure shows precise noise detection.

The rest of the paper is organized as follows. In section 2, we give a brief description of the Singular Spectrum Analysis method. In section 3, we describe our structured noise detection method using SSA. In section 4, we give a description of our method's practical experimentation and the obtained result. In section 5, we discuss various aspects of our algorithm and in section 6, we draw a conclusion.

2 SINGULAR SPECTRUM ANALYSIS

Singular Spectrum Analysis [3, 8] works with a one dimensional time series of finite length, and decomposes the time series into additive components. The algorithm consists of two main steps: *decomposition* and *reconstruction*. The decomposition stage is composed of **embedding** and **singular value decomposition**. The reconstruction stage is composed of **grouping** and **diagonally averaging**. Below we provide a brief description of each of these four steps.

2.1 Embedding

let's consider a real-valued non-zero time series $S = (s_1, s_2, ..., s_N)$ of length N where $N > 2$. In the embedding step, the time series S is converted into a trajectory matrix $\mathbf{X}$ by sliding an M-point window over the time series where M is called the *embedding dimension*. In this step the one-dimensional time series gets mapped into a sequence of $K = N - M + 1$ multi-dimensional lagged column vectors of length M that constitutes the column of the trajectory matrix $\mathbf{X}$.

$$\mathrm{X} = \begin{bmatrix} s_1 & s_2 & \cdots & s_K \\ s_2 & s_3 & \cdots & s_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ s_M & s_{M+1} & \cdots & s_N \end{bmatrix} \tag{1}$$

All of the anti-diagonal elements (when $i+j = constant$, where i and j are row and column indices) of the trajectory matrix (1) has same value (such matrices are also known as Hankel matrix). Embedding dimension M directly affects the decomposition quality. The optimal embedding dimension depends on the purpose of the analysis and the nature of the time series.

2.2 Singular Value Decomposition

Singular Value Decomposition (SVD) of a real non-zero matrix $\mathbf{X}$ of size $M \times K$ decomposes the matrix into a sum of rank one orthogonal elementary matrices (X_i).

$$X = \sum_{i=1}^{L} X_i = \sum_{i=1}^{L} \sqrt{\lambda_i} e_i v_i^T \tag{2}$$

Here, λ_i are the eigenvalues of the covariance matrix $\Sigma = \mathbf{X}\mathbf{X}^T$ in descending order, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_L \geq 0$, and $e_1, e_2, \cdots e_L$ are the corresponding eigenvectors. Here, $v_i = \frac{X^T e_i}{\sqrt{\lambda_i}}$ and L is the no of non-zero singular values of X, $L = \mathrm{argmax}_i \{\lambda_i > 0\}$. The collection $(\sqrt{\lambda_i} e_i v_i)$ is called the eigentriple of the SVD (2). The eigenvectors are also called empirical orthogonal functions (EOF) and can be considered as a set of data adaptive orthogonal basis functions for signal decomposition.

The eigenvalue λ_i represents the amount of partial variance in the direction of the corresponding eigenvectors or EOF v_i and the sum of the eigenvalues gives the total variance of the original time series. The energy contribution of the i-th eigenvector is given by the ratio $\frac{\lambda_i}{\sum_{i=1}^{L} \lambda_i}$ and is called the singular spectrum of the time series.

2.3 Grouping

In the grouping step, we partition the elementary matrices (X_1, $X_2, \cdots, X_L$) into m disjoint groups $I_1, I_2, \cdots I_m$ and then sum the matrices in each subset. For a group I_j whose member indices are $(i_{j_1}, \cdots i_{j_p})$ the resulting matrix would be $X_{I_j} = X_{i_{j_1}} + \cdots + X_{i_{j_p}}$ and we compute such matrix $(X_{I_1}, X_{I_2}, \cdots X_{I_m})$ for all the groups. Finally we obtain the expansion of the trajectory matrix X as a summation (3) of the grouped matrices:

$$X = \sum_{k=1}^{m} X_{I_k} = \sum_{k=1}^{m} \Big(\sum_{i \in I_k} X_i \Big) \tag{3}$$

There is no specific rule dictating how to perform the grouping and it depends on the purpose of the time series analysis and the type of signal and noise. Most often, the singular spectrum and the singular vectors are used to form groups of similar components.

2.4 Diagonal Averaging

In the last step of the reconstruction phase, each of the matrices formed after grouping is converted into a one-dimensional time series of length N by applying a linear transformation called diagonal averaging or *hankelization* [11], where the cross-diagonal elements are averaged to obtain a single value. If X_{I_k} is a reconstructed trajectory matrix with dimension $M \times K$, we average over the elements $i + j = q + 2$ to calculate the qth element of the converted time series, where i and j are row and column indices. If the elements of the matrix X_{I_k} are $Y_{i,j}$, then we can formalize the averaging procedure using following equations:

$$x_n^k = \begin{cases} \frac{1}{n} \sum_{m=1}^{n} Y_{m,n-m+1} & \text{for } 1 \leq n \leq M \\ \frac{1}{M} \sum_{m=1}^{M} Y_{m,n-m+1} & \text{for } M \leq n \leq K \\ \frac{1}{N-n+1} \sum_{m=n-K+1}^{M} Y_{m,n-m+1} & \text{for } K+1 \leq n \leq N \end{cases} \tag{4}$$

This digonalizing (4) of each of the reconstructed trajectory matrices $(X_{I_1}, X_{I_2}, \cdots, X_{I_m})$ produces m resultant time series, whom we refer as reconstructed components (RCs) from here after. These

reconstructed components are the final decomposition of the original time series and we can revert back to the original time series by summing them up (5).

$$s_n = \sum_{k=1}^{m} x_n^k \tag{5}$$

2.5 Why SSA?

SSA is a non-parametric, data-adaptive time series decomposition and analysis method introduced in [3]. SSA decomposes a time series into multiple interpretable additive components, which usually can be considered as the trend components, phase and amplitude modulated oscillatory components and the unstructured random noise components [11]. An important feature of SSA is that trends obtained in this way are not necessarily linear. As SSA is capable of separating out non-linear trends and oscillatory components it is an ideal method to decompose our pressure derivative data into a noise-free signal and structured noise components. Also, it does not make any underlying assumption (i.e.; stationarity, linearity, normality) about the time series of interest. The basis used in SSA to decompose the signal is computed independently for each time series in consideration [28]. Due to the data-adaptive property, SSA is highly suitable for our purpose as our data observation has a high amount of variation in their duration, amplitude offset, rate-of-change and noise structure.

3 STRUCTURED NOISE DETECTION

An ideal noise-free pressure data is smooth and slow varying. But, often they are affected by structured noise and unstructured random noise. Thus, real-life pressure derivative data are often a combination of noise-free signal, structured noise, random noise. The smooth, slowly varying noise-free signal can be regarded as the underlying trend of the observed noise corrupted pressure derivative data and the structured noises are randomly occurring deviations from the usual smooth signal for a brief interval of time and show oscillatory behavior. In Figure 2(left), we show the original pressure derivative data as a combination of noise-free signal and structured noises. In Figure 2(right), we subtract the underlying trend from the pressure derivative data and plot the residue which has non-zero values only in those segments that correspond to the structured noise in the original signal. From the spectral perspective, the noise-free signal lies on the lower range of the frequency spectrum, where else the structured noises are in the mid-range, and the random unstructured noises are in the higher range.

SSA is suitable for decomposing pressure derivative data into additive components that can be meaningfully categorized into trends representing the noise-free signal, phase and amplitude modulated oscillatory components capturing the structured noise components and the random noise components. However, three algorithmic questions remain unanswered.

(1) How do we decompose signals *efficiently* and *accurately* so the structured noise is separated from the trends? Poorly performed decomposition may combine structured noise and trend in the same component, hindering the noise removal process.

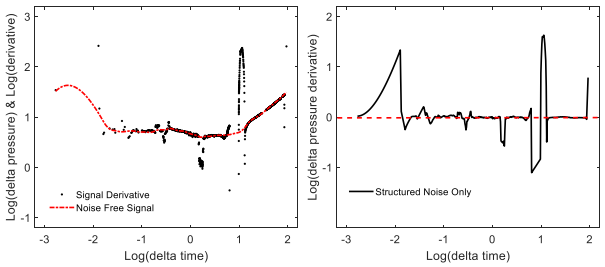

Figure 2: (left) Smooth trend over the noisy signal; (right) structured noise as residue after subtracting trend from the noisy signal (PBU-13-Well-A [4]).

(2) How do we identify components that contain only structured noise? There can be one or more components for structured noise.
(3) How do we temporally locate the structured noise segments in the structured noise components? The structured noise can be of variable length and appear at an arbitrary time.

We address these three issues in the subsequent subsections.

3.1 Signal Decomposition

The separation of components in SSA decomposition is a hard-pressed as well as a widely discussed topic in SSA literature [1, 9, 12]. There are several published techniques to measure the quality of the separation, for example, the weighted correlation of the decomposed components [14]. The quality of decomposition is limited by the nature of the signal as well as by the components to be extracted. Exact separability cannot be achieved for real-world signals as there are usually overlap in their frequency spectrum.

There are no fixed rules on how to select the embedding dimension; rather it depends on the purpose of the decomposition and the nature of the time series being decomposed. It is generally considered that the larger the embedding dimension, the better the decomposition quality and highest quality of decomposition can be achieved when the embedding dimension (M) is equal to the half of the signal length [11]. For periodic signal extraction there are some specific guidelines, for example, M should capture at least one full cycle of desired lowest frequency component [25].

For a small signal with a complex structure, a comparatively larger M may produce component mixing. Hence, it is suggested in [11], to select a small M for such a signal. On [27] it is demonstrated that the embedding dimension is related to the frequency bandwidth of each reconstructed component; the frequency bandwidth of each decomposed component is limited to f_s/M where f_s is the sampling rate of the signal. Hence, the selection of M can lead to a trade-off between component mixing vs. decomposition resolution. Also, a larger M corresponds to an increased amount of computation and time complexity.

Considering the above-mentioned fact, for a small signal with complex structure, in [27], the embedding dimension is selected according to the rule $M = f_s/f_b(6)$, where f_b is determined by the frequency bandwidth of desired signal structure to be extracted. As

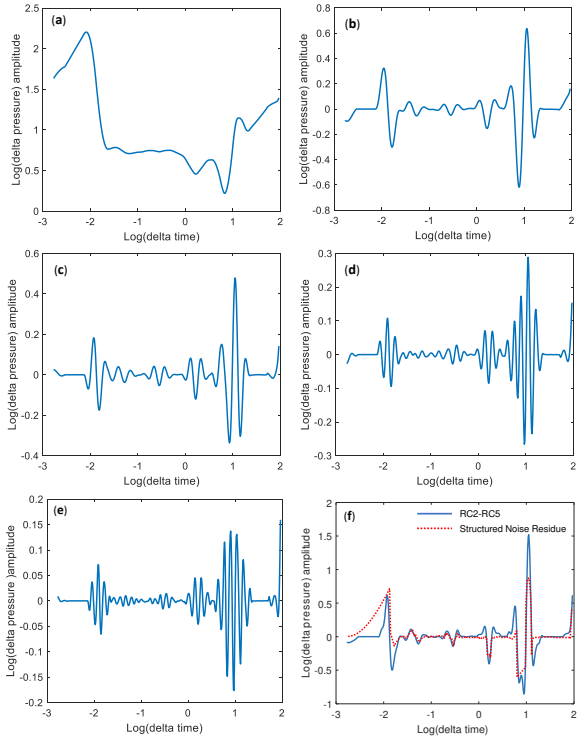

Figure 3: (a)-(e)Reconstructed components (RC1-RC5) after SSA decomposition (f) Sum of RC2 through RC5 overlapped with structured noise residue. (PBU-13-Well-A [4])

our main goal is to separate structured noise component, the selection of M needs to be based on noise property. In our data, the sampling frequency is 100Hz, which means a regular sampling of 100 data points per one log cycle of the time axis, and the aperiodic structured noise components observed in this experiment generally expands for 15-40 data points which leads to a frequency bandwidth $f_b \approx 4$Hz. Thus, based on, Eqtn (6), we use $M = fs/fb = 100/4 = 25$ as the embedding dimension. In Figure 3[(a)-(e)], we plot the first five reconstructed components (RC) and in Figure 3(f), we plot the summation of RC2 through RC5 overlapped with the structured noise residue for M=25 for PBU-13-Well-A data [4].

3.2 Structured Noise Component Identification

Once the time series is decomposed into additive components, the next objective is to identify the components that capture most of the structured noise information. This component of interest identification step is formally called the *grouping*. Similar to the embedding dimension selection step, there is no strict rule about how to perform this grouping, rather heuristics are used based on the type of the time series of interest and the purpose of the analysis [27].

An eigenvalue represents the amount of variance captured by the corresponding component[8]. In general, the largest eigenvalues are associated with the trend components like the smooth, slow varying noise-free signal, the intermediate ones are related with the mid-frequency components like the oscillatory structured noises,

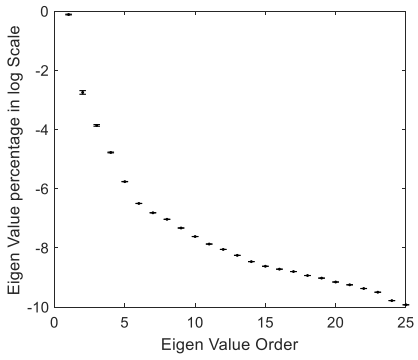

Figure 4: SSA eigenvalue spectra with 95% confidence limit for all 38 pressure derivative data of Well-A[4].

and lower values are associated with the high-frequency random noises. Our noise-free signal is supposed to be captured by the trend components which have high eigenvalues associated with them as they dictate the overall shape of the time series and hence accounts for most of the variance present in the signal. And the structured noise components should be projected onto the oscillating ones which has mid-range eigenvalues as they account for the short-duration oscillation around the trend.

Therefore, we use this discriminatory property of eigenvalues associated with the components to identify the trend components and the structured noise-capturing oscillatory components. In Figure 3, we have plotted the reconstructed components after performing SSA decomposition on PBU-13-Well-A [4]. In Figure 4, we plot the eigenvalue spectra of the associated decomposed components for the 38 pressure derivative data of Well-A. We observe that the first few eigenvalues account for most of the variance present in the signal, and after those, most of the eigenvalues are close to zero. These almost zero eigenvalues are usually associated with the noisy components. In [18], they have derived a rule to discard those noisy components. Based on their technique we reject all the components relating to the eigenvalues λ_i, if $i \geq \mathcal{L}$, where

$$\mathcal{L} = \underset{a}{\operatorname{argmin}} \left\{ \frac{\sum_{k=1}^{a} \lambda_k}{\sum_{j=1}^{l} \lambda_j} \geq 0.95 \right\}$$

The trend components usually have higher values in the similar range and the structured noise components have comparatively lower eigenvalues. In the Figure 3(a), we observe that the RC1 represents the trend and in Figure 3(f), the summation of RC2 through RC5 matches with the structured noise residue. Also, in the eigenvalue spectra, there is a significant gap between RC1 and the rest of the RCs. Hence, to generalize our procedure, we distinguish between the trend and the structured noise components by identifying the largest gap in between two consecutive eigenvalues and then based on that differentiating point we group the trend and structured noise components.

3.3 Structured Noise Localization

Ideally, after SSA decomposition, the trend component would represent the noise-free smooth signal, and the structured noise component should have non-zero values only in segments corresponding to structured noises in the original signal. But often time due to high amplitudes of the structured noises, component mixing occurs and they affect the trend components. Also, in the structured noise components, noise oscillations expand to neighboring regions, for example; where there are flow regimes in between two structured noises, or flow regimes close to high amplitude structured noises. Nevertheless, the amplitude of oscillatory components is still highly correlated to the likelihood of occurrence of structured noise. Thus, to localize structured noise segments from the structured noise components, we use a threshold value to compare against the absolute value of the summation of the structured noise components, and the sample data points exceeding the threshold value are identified as structured noise.

3.3.1 Single Sample Threshold Selection. In our pressure derivative data, in some observation the structured noises have a high amplitude, where else in some observation they have low deviation from the actual trend. Though the structured noise properties maintain somewhat similarities in a single well, the variation becomes prominent in between observation of two different well. So, it is quite impossible to obtain a globally optimal threshold. In summary, we need to have a mechanism to modify our threshold from well to well. To address this issue, we employ a semi-supervised learning approach where we select the threshold by manual evaluation of noise detection performance for a range of threshold values using one PBU data per well. Then, we select the threshold which results in best performance to be used as a fixed threshold for the rest of the observation of that well. Though this process might seem laborious, in practical scenarios, we can automate the noise detection procedure for a few dozens of other pressure derivative observation by performing a manual evaluation on only a single data observation.

3.3.2 Boundary Refinement. The selected structured noise components have high values in the center of the noise segment, and the region where noise transcends into signal has comparatively lower values. In Figure 4(left), we have plotted the absolute value of the summation of the oscillatory components, which portrays the aforementioned scenario. To tackle this problem, we calculate a windowed average of the absolute value of the summation of the oscillatory components which is plotted in Figure 4(right). In this case, the boundary region of structured noise segments has increased amplitude than before which helps in the precise localization of whole noise segment.

4 EXPERIMENT AND RESULTS

We use real log-log pressure data and the semi-log derivative of log-log pressure data from two real gas well and one oil well for quantitative accuracy calculation. Moreover, we use three separate oil well pressure data for qualitative evaluation by the domain expert.

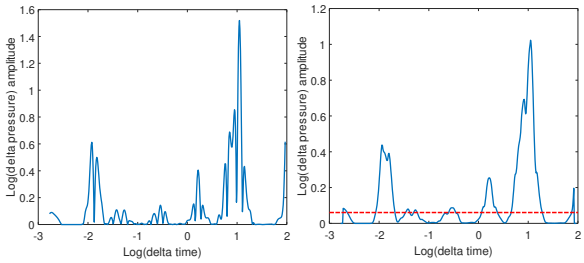

Figure 5: (left) Sum of the structured noise component values as an indicator of structured noise presence; (right) windowed averaging performed on the sum of the structured noise components. (red dashed line indicates threshold limit, PBU-13 Well-A [4])

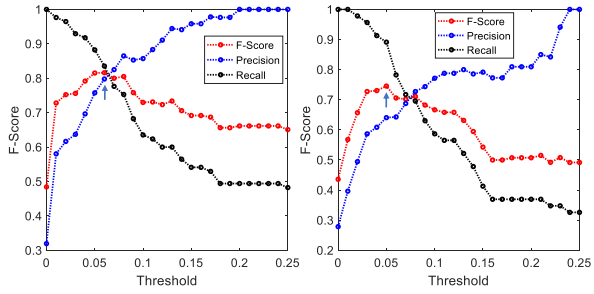

Figure 6: (left) PBU-13-Well-A, (right) PBU-24-Well-A [4]; Two demonstrations of using F-score as performance evaluation measure for semi-supervised threshold selection. (maximum F-score values are shown by arrow)

4.1 Data Collection and Preprocessing

The pressure data is extracted from the shut-in period of the continuous stream of pressure data recorded by the permanent downhole pressure gauge installed in wells drilled in oil and gas fields. The extracted shut-in period pressure values were converted into pressure change data with reference to the pressure value at the beginning of the shut-in period. Both the pressure change data and time axis was converted into log scale which in turn produced the log-log pressure change data. Finally, the semi-log derivative of this pressure change was calculated using the method of Bourdet [2] to produce the log-log pressure derivative plot. From this plot, we remove the data points with an undefined value, which occurs due to a negative change in pressure value around that point in time. Also. we perform a linear interpolation to create evenly spaced data points from irregularly sampled and log-scale transformed data points.

4.2 Data Labeling and Accuracy Calculation

We use *F-score* measure to calculate the accuracy of our proposed method. We use log-log pressure derivative data where both structured noise and flow regime segments have been manually labeled by domain experts. We label the flow regimes in addition to the structured noise segments as there are some segments present in

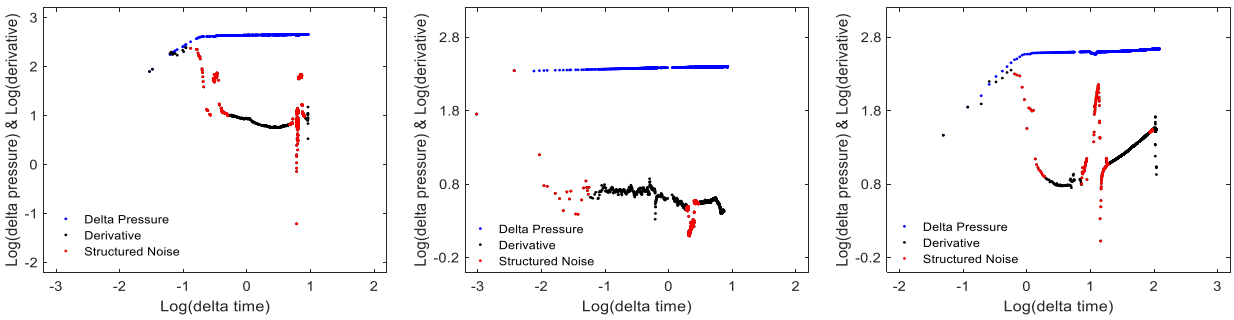

Figure 7: Automatic noise detection result on real pressure build up data from *Well-A* [4]; (left) structured noise due to unknown reason; (middle) structured noise due to unknown reason; (right) structured noise due to pressure leak at flow line.

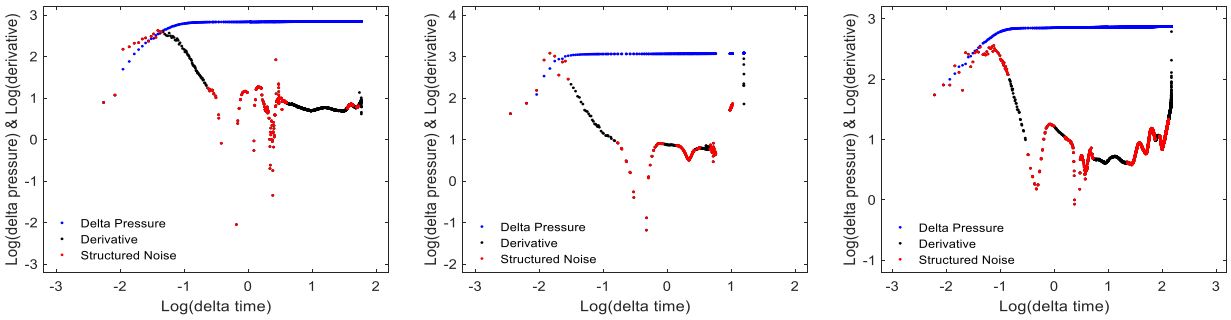

Figure 8: Automatic noise detection result on real pressure build up data from *Well-C* [4]; (left) structured noise due to fluid segregation; (middle) structured noise due to fluid segregation; (right) structured noise due fluid segregation.

Table 1: Accuracy

Well	No Lower Bound	Minimum Threshold > 0	Minimum Threshold > 0.02
A	69.73%	74.07%	75.02%
B	63.22%	66.22%	67.14%
C	80.36%	80.36%	80.59%
Average	71.1%	73.55%	74.25%

the pressure derivative data which are neither noise nor flow. While calculating the accuracy of our algorithm we consider only the labeled flow regimes and structured noise segments.

We compare our detected structured noise segments with the labeled ground truth data and calculate true positive (*TP*) , false positive (*FP*), and false negative (*FN*). Then, we use the (*TP*), (*FP*) and (*FN*) to calculate precision,recall, and finally the F-score.

$$precision = \frac{TP}{TP + FP}, recall = \frac{TP}{TP + FN}$$

$$F - score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

4.3 Threshold Selection

The threshold parameter is selected in a semi-supervised fashion where one pressure derivative data is used to select the threshold parameter for rest of the data of the same well. First, the user manually labels one selected pressure derivative data. Then noise detection is performed on that data for a range of threshold values, and F-score is calculated for each of the threshold values to select the threshold corresponding to the highest F-score. This best performing threshold value is the default for the rest of the data of that well. In the Figure 6(left) for PBU-13-Well-A, we observe that the maximum F-score is achieved for the threshold value 0.06. So if we select this PBU observation for threshold selection step, our automatic threshold value for the rest of the observation from Well-A would be 0.06. Similarly, in Figure 6(right), we see that for PBU-24-Well-A, the optimal threshold is 0.05.

4.4 Results

The final noise detection performance of our method depends mainly on the selection of threshold; which can be selected using any PBU data. To get an overall performance evaluation, for each well, we use each PBU once for manual threshold selection. And then we use that threshold as the fixed threshold for the rest to perform noise detection and accuracy calculation. Finally, we calculate the average accuracy for all the threshold. In table 1 we show the average accuracy for three different well, where we notice that, without

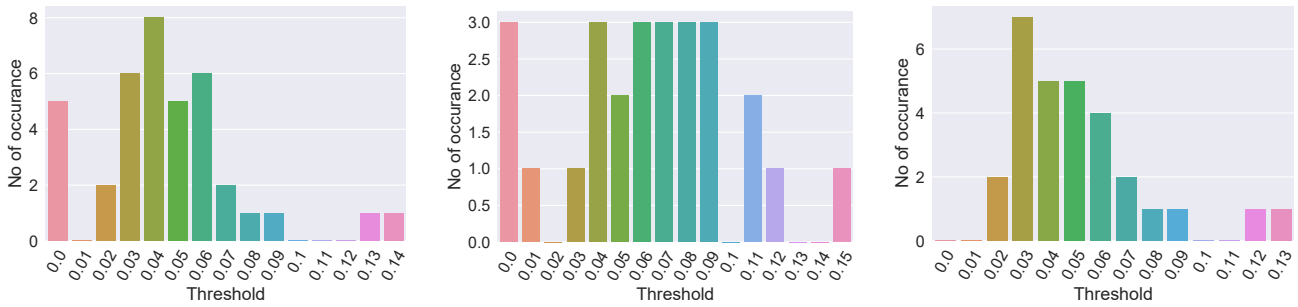

Figure 9: Histogram plot of the number of times a threshold value is selected using single-sample learning; (left) Well-A, (middle) Well-B, (right) Well-C [4].

using any constraint on the threshold value, the average accuracy among three well is 71.1%. However, if we put some lower-bound on the threshold value, it increases up to 74.25%. For blind performance analysis, we run our method on three more oil well [4] data, which was evaluated by domain experts to assert the usability of our algorithm in practical application. In Figure 7 and 8, we give a few graphical demonstration of our structured noise detection performance.

Runtime. The run time of our method is always less than 0.05 seconds for a single PBU averaged over several runs on different well data (using a core i5 2.70 GHz desktop computer).

5 DISCUSSION

In this section, we discuss various aspects of our proposed method.

5.1 Parameter Sensitivity

In Figure 9, for three different well, we plot histogram which shows how many times a particular threshold value is selected from a well. In Figure 10, we plot the overall accuracy for a range of threshold values for three different well. In Figure 9 for Well-A, we observe that the median of the selected threshold value is 0.06 and the noise detection accuracy is also highest around that threshold value. This same scenario is noted for the other wells too, which implies that if we select an observation which bears some structured noises, it would, in turn, produce a good enough threshold for the rest of the observation of that well.

5.2 Interpretability

The solitary parameter of our method, the threshold value has an *intuitive interpretation* which the user can readily control to tune the outcome according to specific well characteristics. Here, a higher threshold value implies that the method will be more selective in labeling data segments as noise, where else a lower threshold value implies that the algorithm will be more inclined towards detecting a particular data segment as structured noise. Due to this interpretable user control mechanism, this method can be effectively deployed in practical field usage where the users would be able to fine-tune the method accordingly without requiring any knowledge about the underlying algorithm.

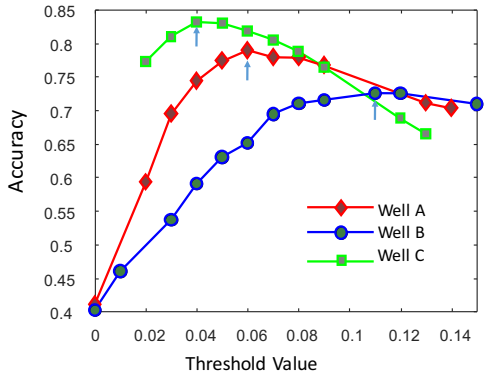

Figure 10: Accuracy of structured noise detection performance for various threshold values on Well-(A,B,C) [4]. The maximum accuracy values (shown by arrows) vary across wells.

6 CONCLUSION

In this paper, we show a practical example of using signal processing techniques to improve the usability of data science methodologies on process monitoring dataset. Our developed method is adaptive to the variation of structured noises, and the run-time is not dependent on the amount of noise present in the data. The proposed method is fast, computationally inexpensive and requires minimal manual intervention, which makes it perfectly suitable for practical day-to-day deployment. The solitary parameter is readily interpretable and intuitive to fine-tune. Experimental evaluation, as well as assessment by domain experts have validated the accuracy and effectiveness of the method. In future, we will use this method to increase the accuracy of classification and clustering tasks on additional datasets, which will demonstrate the generalizability of this method.

7 ACKNOWLEDGEMENTS

The authors thank ExxonMobil Production Company, Anadarko, Petrobras and Eni for the permission to publish data. The author also thanks ExxonMobil Upstream Research Company for the permission to publish this paper.

REFERENCES

[1] Theodore Alexandrov. 2008. A method of trend extraction using singular spectrum analysis. *arXiv preprint arXiv:0804.3367* (2008).

[2] Dominique Bourdet, JA Ayoub, YM Pirard, et al. 1989. Use of pressure derivative in well test interpretation. *SPE Formation Evaluation* 4, 02 (1989), 293–302.

[3] David S Broomhead and Gregory P King. 1986. Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena* 20, 2-3 (1986), 217–236.

[4] Well Data Provided by Upstream Research Company. 2018. Well-[A,B,C,D,E,F].

[5] Robert C Earlougher. 1977. *Advances in well test analysis.* Henry L. Doherty Memorial Fund of AIME New York.

[6] Chris Fair, Ricardo Flores, Bilal Hakim, Soumitra Nande, et al. 2014. Using the Results From Automated Petroleum Engineering Calculations to Accelerate Decision Workflows. In *SPE Asia Pacific Oil & Gas Conference and Exhibition*. Society of Petroleum Engineers.

[7] Foad Ghaderi, Hamid R Mohseni, and Saeid Sanei. 2011. Localizing heart sounds in respiratory signals using singular spectrum analysis. *IEEE Transactions on Biomedical Engineering* 58, 12 (2011), 3360–3367.

[8] Michael Ghil, MR Allen, MD Dettinger, K Ide, D Kondrashov, ME Mann, Andrew W Robertson, A Saunders, Y Tian, F Varadi, et al. 2002. Advanced spectral methods for climatic time series. *Reviews of geophysics* 40, 1 (2002), 3–1.

[9] Nina Golyandina. 2010. On the choice of parameters in singular spectrum analysis and related subspace-based methods. *arXiv preprint arXiv:1005.4374* (2010).

[10] Nina Golyandina and Anton Korobeynikov. 2014. Basic singular spectrum analysis and forecasting with R. *Computational Statistics & Data Analysis* 71 (2014), 934–954.

[11] Nina Golyandina, Vladimir Nekrutkin, and Anatoly A Zhigljavsky. 2001. *Analysis of time series structure: SSA and related techniques.* Chapman and Hall/CRC.

[12] Nina Golyandina and Alex Shlemov. 2013. Variations of singular spectrum analysis for separability improvement: non-orthogonal decompositions of time series. *arXiv preprint arXiv:1308.4022* (2013).

[13] Valery Guralnik and Jaideep Srivastava. 1999. Event detection from time series data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 33–42.

[14] Hossein Hassani. 2007. Singular spectrum analysis: methodology and comparison. (2007).

[15] Dilhan Ilk, David M Anderson, Garth WJ Stotts, Louis Mattar, Thomas Blasingame, et al. 2010. Production data analysis–Challenges, pitfalls, diagnostics. *SPE Reservoir Evaluation & Engineering* 13, 03 (2010), 538–552.

[16] S Joe Qin. 2003. Statistical process monitoring: basics and beyond. *Journal of Chemometrics: A Journal of the Chemometrics Society* 17, 8-9 (2003), 480–502.

[17] Ajay Kumar Maddirala and Rafi Ahamed Shaik. 2016. Removal of EOG artifacts from single channel EEG signals using combined singular spectrum analysis and adaptive noise canceler. *IEEE Sensors Journal* 16, 23 (2016), 8279–8287.

[18] Jonathan Mamou and Ernest J Feleppa. 2007. Singular spectrum analysis applied to ultrasonic detection and imaging of brachytherapy seeds. *The Journal of the Acoustical Society of America* 121, 3 (2007), 1790–1801.

[19] Seyedeh Robab Moosavi, Jafar Qajar, and Masoud Riazi. 2018. A comparison of methods for denoising of well test pressure data. *Journal of Petroleum Exploration and Production Technology* (2018), 1–16.

[20] Valentina Moskvina and Anatoly Zhigljavsky. 2003. An algorithm based on singular spectrum analysis for change-point detection. *Communications in Statistics-Simulation and Computation* 32, 2 (2003), 319–352.

[21] Alan V Oppenheim. 1999. *Discrete-time signal processing.* Pearson Education India.

[22] Hugh Richard Rees, John Foot, Richard Heddle, et al. 2011. Automated Pressure Transient Analysis with Smart Technology. In *SPE Digital Energy Conference and Exhibition*. Society of Petroleum Engineers.

[23] MY Soliman, J Ansah, S Stephenson, B Mandal, et al. 2001. Application of wavelet transform to analysis of pressure transient data. In *SPE annual technical conference and exhibition*. Society of Petroleum Engineers.

[24] Satomi Suzuki et al. 2018. Using Similarity-Based Pattern Detection to Automate Pressure Transient Analysis. In *Abu Dhabi International Petroleum Exhibition & Conference*. Society of Petroleum Engineers.

[25] AR Teixeira, AM Tomé, EW Lang, P Gruber, and A Martins da Silva. 2005. On the use of clustering and local singular spectrum analysis to remove ocular artifacts from electroencephalograms. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, Vol. 4. IEEE, 2514–2519.

[26] Robert Vautard and Michael Ghil. 1989. Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Physica D: Nonlinear Phenomena* 35, 3 (1989), 395–424.

[27] Shanzhi Xu, Hai Hu, Linhong Ji, and Peng Wang. 2018. Embedding Dimension Selection for Adaptive Singular Spectrum Analysis of EEG Signal. *Sensors* 18, 3 (2018), 697.

[28] Pascal Yiou, Didier Sornette, and Michael Ghil. 2000. Data-adaptive wavelets and multi-scale singular-spectrum analysis. *Physica D: Nonlinear Phenomena* 142, 3-4 (2000), 254–290.