An Evaluation of Smoothing Filters for Gas Sensor Signal Cleaning

Enobong Bassey, Jacqueline Whalley and Philip Sallis
Geoinformatics Research Centre, School of Computer and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand
e-mails: {ebassey@aut.ac.nz, jwhalley@aut.ac.nz, psallis@aut.ac.nz}

Abstract— **This paper explores signal response and methods for extracting the desired digital signal, from gas sensor arrays, while maintaining the shape and resolution of that signal. A comparative evaluation of Savitzky–Golay smoothing, moving average, local regression and robust local regression filters for cleaning signals obtained from gas sensor devices during the pre-processing phase is provided. It was found that the Savitzky–Golay smoothing filtering method provided the best approximation of the sensor response.**

Keywords-denoising; gas sensor arrays; signal processing.

I. INTRODUCTION

Typically, raw signals acquired from gas sensors are contaminated by noise and outliers and as a result the signal is occluded to a significant degree making accurate measurement of a sensor's response impossible. Noise in sensor systems has several possible sources and is introduced at various stages in the measurement process. Several forms of noise, including thermal and shot noise, are irreducible because they are inherent to the underlying physics of the sensors or electronic components. Other forms of noise which could be avoided originate from processes, and include 1/f noise, transmission, and quantization noise [1]. Noise introduced in the early measurement stages is considered to be the most harmful as it propagates and can be amplified through subsequent stages in the signal pathway [2].

While physical filters have been found to be successful in producing a cleaner signal they do not cover the full resolution and shape of the curve. In order to improve the interpretability, sensitivity and selectively of gas sensor array signals it is preferable to use the full resolution and coverage of the signal.

Several signal processing approaches have been investigated as an approach to reducing noise levels [3]. However, these approaches are typically static or steady state approaches and therefore do not encompass the full temporal signal [4].

In this paper we report on an evaluation of methods for feature extraction and denoising the digital signal from thin film zinc oxide and tin dioxide composite gas sensor devices. The aim was to find a method that not only cleaned the signal but that also maintained the shape, precision and resolution of the signal regardless of sensor composition. In Section II the stages of gas sensor array signal processing is outlined. Details of various approaches to signal pre-processing are then discussed in Section III. In Section IV the signal pre-processing methods evaluated are detailed. Section V gives the results of applying each of these signal denoising methods. Finally, Section VI provides a summary of our results and suggests possible avenues for future work.

II. GAS SENSOR ARRAY SIGNAL PROCESSING

The signal processing of gas sensor data can be divided into four steps [5]: (1) *pre-processing*, for further processing of the sensor signal (e.g., denoising, drift compensation, concentration normalization); (2) *dimensionality reduction* (of the input signal to avoid problems associated with high dimensionality data); (3) *prediction* (of the interesting properties of the sample, e.g., class membership, related odour samples); and (4) *validation*, where model and parameter settings are selected in order to optimize a criterion function (e.g., classification rate, mean-squared error). A useful summary of statistical and optimization methods that have been used to process gas sensor array signals is provided by Gutierrez-Galvez [6]. The work reported here focuses on improving existing pre-processing techniques used to eliminate noise, smooth and filter data, enhance sensor signals and ultimately improve measurement.

III. SIGNAL PRE-PROCESSING

Signal pre-processing facilitates noise elimination, data smoothing/filtering and signal enhancement, with the sole aim of increasing the signal-to-noise ratio without greatly distorting the signal. The choice of signal pre-processing method is known to have a significant impact on the results and performance of the pattern analysis system [1][7].When developing a pre-processing method three criteria should be considered [8]. The algorithm must: preserve the chemical selectivity differences between different profiles and limit run-to-run retention/migration time shift, be fast and less memory-demanding to deal with large numbers of data sets in a short period of time, and the resulting precision of retention/migration time estimation should be significantly improved in comparison with that initially provided by the instrumentation [1][7]. We propose that wavelet transform smoothing filters, for this purpose, should meet all three criteria.

Wavelets are a family of wavelet transforms that are considered to be a time-frequency representation for continuous-time (analog) signals [9]. They have a compact support (i.e., they differ from zero only in a limited time domain) and easily represent the different features of a signal, especially sharp signals and discontinuities. When applied to analytical signal processing wavelet transforms provide a simple procedure with short operation time, low

 ISBN: 978-1-61208-365-0

memory requirements, high precision, and good reproducibility [10]. Examples of wavelet transforms for denoising signals include the Savitzky–Golay Smoothing Filter (SGF) [11], Fast Fourier transform (FFT), Multivariate Wavelet Denoising (MWD), Discrete Wavelet Transform (DWT) [12], and Continuous Wavelet Transform (CWT) [2].

For this work, we have elected to evaluate the use of an SGF by comparing its performance with moving average filter and local regression methods. SGF is known to be superior to other adjacent averaging FIR filters because it tends to preserve the features of the data in the signal, such as peak height and width. Moreover, when using SGF it is possible to increase the smoothness of the result by changing the window size, or increasing the number of data points used, in each local regression. Although SGFs are considered to be less successful than standard averaging FIR filters at eliminating noise, they are more effective at preserving the pertinent high frequency components of the signal [4], and are optimal in minimising the least-squares error in fitting a polynomial to frames of noisy data. Moreover, SG filters can preserve more of the high-frequency content of a signal, but this is at the expense of reduced noise elimination.

Therefore, an SGF might prove to be a good choice for gas senor signal cleaning where it is important to preserve the height, width, amplitude and overall profile of the signal.

IV. METHODOLOGY

In order to evaluate the usefulness of wavelet transforms for preprocessing gas sensor arrays we elected to investigate the performance of an SGF, using tin dioxide (SnO_2) and zinc oxide (ZnO) sensor devices.

For a given experiment ***E***, a response matrix *R* is usually obtained in which each column represents a response matrix associated with the concentration of the target gas produced *C* at operating temperature *T* and the rows give the response matrices of each individual sensor (1).

$$E = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1T} \\ R_{21} & R_{22} & \cdots & R_{2T} \\ \vdots & \vdots & R_{pq} & \vdots \\ R_{C1} & R_{C2} & \cdots & R_{CT} \end{pmatrix} \quad (1)$$

For this work we opted, as a preliminary investigation, to evaluate the signal cleaning methods using two sensor devices rather than an array of sensors. Thus the experimental matrix ***E*** can be represented as vector $Rq = (R1q, R2q, Rpq... RCq)^T$. Each of these sensor devices were individually primed with 150ppm methanol at 250°C and operated at three different temperatures [150, 250 and 350 ºC] for the target gas, methanol, at three different concentrations [100, 150 and 200 ppm].

The data were visualised and smoothed using multiple fitting algorithms. The parameters of the curve after residual analysis were then analysed and fits generated. Then the curve was reconstructed to determine the accuracy of the models. Subsequently, an optimal model was selected for generating the best polynomial model.

The performance of the SGF was then compared with a moving average filter, lowess, loess, rlowess, and rloess methods. Details of these six methods and the results of the experiments are provided in the following sections.

A. Moving Average Filtering

A moving average filter, equivalent to low pass filtering, can be used to smooth data by replacing each data point with the average of the neighbouring data points within a specified span of data points. This process is described by the difference equation (2) [13], where $y_s(i)$ is the smoothed value for the i^{th} data point, N is the number of neighbouring data points on either side of $y_s(i)$, and 2N+1 is the span.

$$y_s(i) = \frac{1}{2N+1}\big(y(i+N-1) + \cdots + y(i-N)\big) \quad (2)$$

B. Local Regression Smoothing: rLowess & rLosess

The names lowess and loess are derived from the term "locally weighted scatter plot smooth," as both methods use locally weighted linear regression to smooth data.

The smoothing process is considered local because each smoothed value is determined by neighbouring data points defined within the span. The process is weighted because a regression weight function is defined for the data points contained within the span. The local regression first and second degree polynomial models are lowess and loess, respectively. The robust local regression method (rlowess and rloess) assigns a lower weight to outliers in the regression and assigns a zero weight to data outside six mean absolute deviations.

Like the moving average method, the lowess and loess smoothed value is determined by neighbouring data points defined within a span [14]. However, in this case the process is weighted by a regression weight function that is defined for all the data points contained within the specified span. In addition to the regression weight function, a robust weight function can be used; this makes the process resistant to outliers.

Lowess and loess are differentiated by the model used in the regression: lowess uses a linear polynomial, while loess uses a quadratic polynomial.

C. Savitzky–Golay Smoothing Filter (SGF)

The SGF is based on local least-squares polynomial approximation [11]. It is a generalization of the finite-impulse response (FIR) or moving average filter with filter coefficients determined by an unweighted linear least-squares regression and a polynomial model of specified degree. The process, equivalent to discrete convolution with a fixed impulse response, involves fitting a polynomial to a set of input data and evaluating the resulting polynomial at a single point within the approximation interval. The SG smoothing procedure consists of replacing the central point of a window (containing an odd number of points, 2p + 1) with the value obtained from the polynomial fit. The window is moved one data point at a time until the whole signal is scanned; thus, creating a new, smoothed value for each data point. The smoothed signal g(t) is then calculated by convolving the signal f(t) with a smoothing (or convolution)

function h(t) [9] for all observed data points p where f(m) is the curve function at point m and h(m - t) ≠ 0 (3). The convolution function h(t) is defined for each combination of degree of the polynomial and window size.

$$g(t) = f(t) * h(t) = \frac{\Sigma f(m)h(m-t)}{\Sigma h(m)} \quad (3)$$

A typical digital filter can be applied to a series of equally spaced data values $f_i \equiv f(t_i)$, where $t_i \equiv t_0 + i\Delta$ for some constant sample spacing Δ and i = …-2, -1, 0, 1, 2, …. Therefore, the SG smoothing operations consist of the replacement of each data point f_i with a linear combination of g_i and a number of nearby neighbours *n* [10] where *nL* is the number of neighbouring points prior to the data point i and *nR* is the number of neighbours after data point *i* (4)

$$g_i = \sum_{n=-nL}^{nR} c_n f_{i+n} \quad (4)$$

and where the coefficients c_n are the weights of the linear combination and a causal filter would have an *nR* of zero [2][15][16]. For the simplest possible averaging smoothing filter (similar to the moving average window), the smoothed point is the average of an odd number of neighbouring data points. This moving window average (5) is computed as g_i, i.e., as the average of the data points from f_{i-nL} to f_{i+nR}, for some fixed nL = nR = M; and the weights $c_n = 1/(nL + nR + 1)$ [12]:

$$g_i = \sum_{n=-M}^{M} \frac{f_{i+n}}{2M+1} \quad (5)$$

The weights c_n are chosen in such a way that the smoothed data point gi is the value of a polynomial fitted by least-squares to all (nL + nR + 1) points in the moving window. That is, for the group of 2M+1 data centered at n = 0, and the coefficient of the polynomial is obtained as (6) [15][16].

$$c_n = p(n) = \sum_{k=0}^{N} a_k n^k \quad (6)$$

This minimises the mean-squared approximation error (7) for the group of input samples centred on n = 0:

$$\varepsilon_n = \sum_{n=-M}^{M} \left(p(n) - x[n]\right)^2 = \sum_{n=-M}^{M} \left(\sum_{k=0}^{k} a_k n^k - x[n]\right)^2 \quad (7)$$

Therefore, the smoothed data point g_i by the Savitzky–Golay algorithm [10] is given by (8):

$$g_i = \frac{\sum_{n=-nL}^{nR} c_n f_{i+n}}{\sum_{n=-nL}^{nR} c_n} \quad (8)$$

V. RESULTS

A. Smoothing

Local regression smoothing (lowess and loess) and robust local regression (rlowess and rloess) were carried out using a span of 10% of the data points. Moving average and SGFs were used to smooth the data using a span of 5 and 55. These values were chosen because they gave comparable results.

The results from smoothing the raw data using local regression smoothing and robust local regression were found to give essentially the same shape resolution (Fig. 1). With the moving average and SGFs, using a span of 55 gave better smoothing/shape resolution than using a span of 5, as shown in Fig. 2. For the ZnO device rloess, rlowess, and losses gave improved smoothing over lowess (Fig. 3). Moving average and SGF with a span of 55 gave better smoothing than with a span of 5 as shown in Fig. 4.

However in all cases, although smoothing was improved, the approximation of the curve was poorer because less raw data points were fitted.

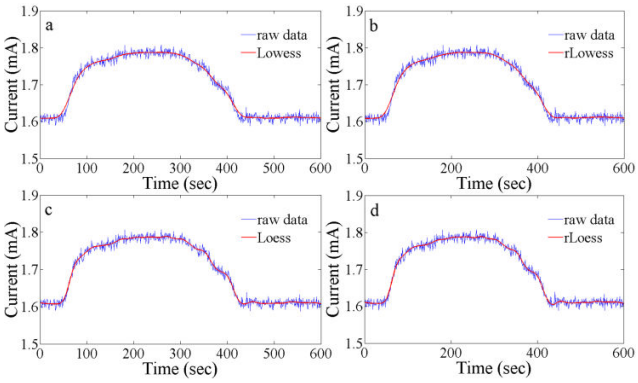

Figure 1. SnO_2 Sensor Local Regression Smoothing (a) Lowess, (b) rLowess and Robust Local Regression Smoothing; (c) Loess (d) rLoess.

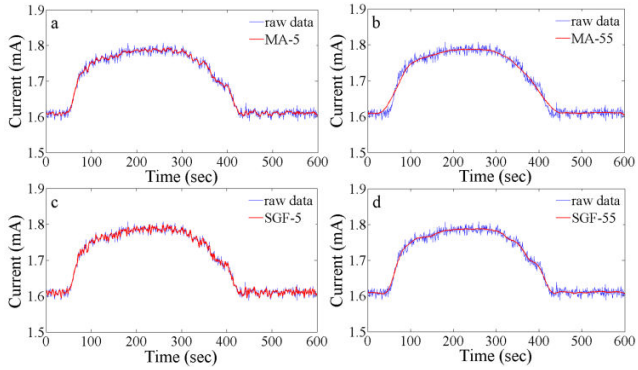

Figure 2. SnO_2 Sensor (a) MA, span = 5, (b) MA span = 55, (c) SGF, span = 5, and (d) SGF, span = 55.

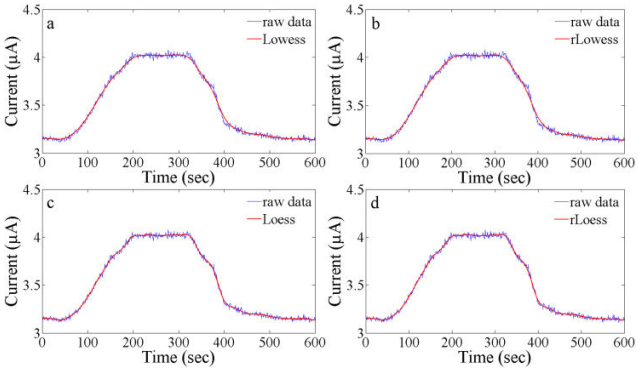

Figure 3. ZnO Sensor Local Regression Smoothing (a) Lowess, (b) rLowess and Robust Local Regression Smoothing; (c) Loess (d) rLoess.

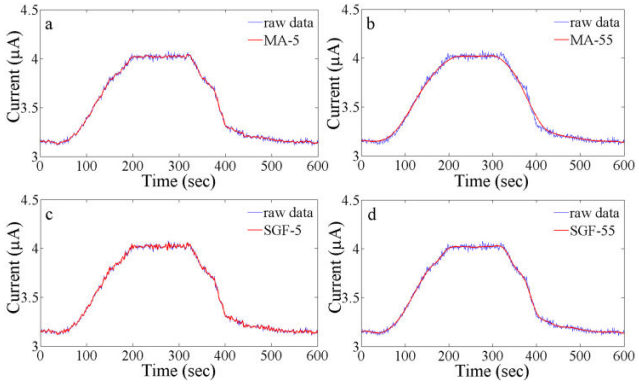

Figure 4. ZnO Sensor (a) Moving Average (Span = 5), (b) Moving Average (Span = 55), (c) SGF, span = 5, and (d) SGF, span = 55.

It was also found that using SGF, the height and width of narrow peaks is accurately captured by higher degree polynomials, but wider peaks are poorly smoothed. For optimality, a polynomial degree of three was applied for the implementation of the SG filtering.

B. *Curve Fitting Accuracy*

Curve fitting was undertaken for each of the smoothing processes and the coefficient of determination (R-squared (R^2)) was calculated using a polynomial of three (9).

R-squared indicates how well data points fit a statistical model and provides a measure of how well observed outcomes are replicated by the model, as the proportion of total variation of outcomes explained by the model [17]. In other words, R^2 is proportional to the variability of the response signal in the polynomial model.

In our case, R^2 indicates the proportionate amount of variation in the response signal explained by the independent variables *t* in the polynomial model where SSE is the sum of squared error, SSR is the sum of squared regression, and SST is the sum of squared total.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (9)$$

In all cases, the R^2 value for the SnO_2 device was greater than that observed for the ZnO device indicating that a better fit of the model is obtained for the SnO_2 device (Fig. 5)

The best value of R^2 was obtained using the moving average smoothing method with a span of 55, followed closely by the lowess local regression for both the SnO_2 and ZnO devices.

However, visual observation of the smoothing results indicates that the SGF and the loess methods gave better shape resolution. The norms of residuals were found to be the same regardless of the smoothing method for each device (5.25E-04 and 3.62E-06 for SnO_2 and ZnO, respectively).

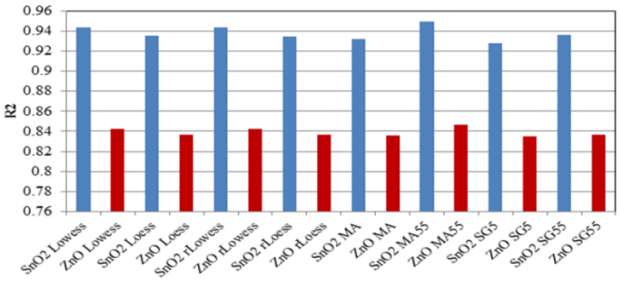

Figure 5. Coefficient of Determination vs. the Curve Fitting Process.

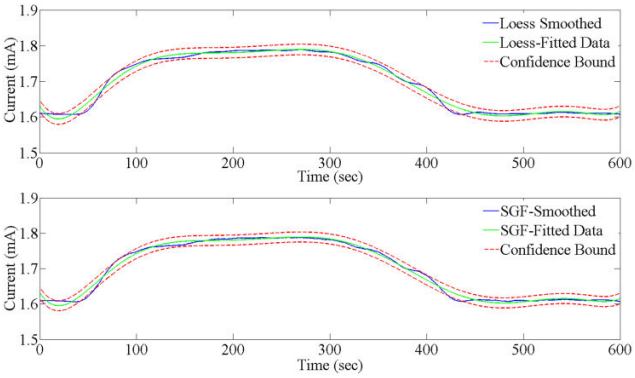

Figure 6. Polynomial Fit with Confidence Bounds and Smoothed SnO_2 Device Data: (top) Loess, (bottom) SGF-55.

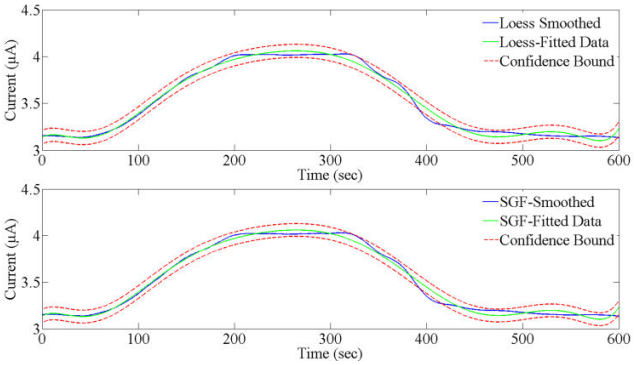

Figure 7. Polynomial Fit with Confidence Bounds and Smoothed ZnO Device Data: (top) Loess, (bottom) SGF-55.

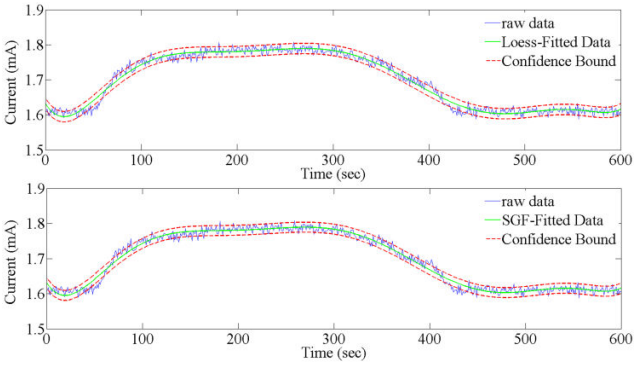

Figure 8. Fitting the Confidence Bounds over the SnO_2 Device Raw Data: (top) Loess, (bottom) SGF.

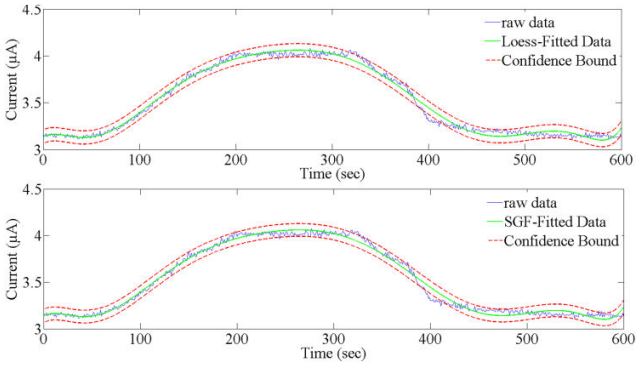

Figure 9. Fitting the Confidence Bounds over the ZnO Device raw data: (top) Loess , (bottom) SGF.

C. Model Calibration

To calibrate the model, different polynomial models were tested on the raw data to determine the best fit within the confidence bounds. An attempt to fit the data using a higher degree polynomial is presented using the SGF (span = 55) and the loess (span = 10%) methods. The best fit was found to occur using a polynomial degree of nine for both the loess and the SGF methods. For the SnO_2 device, the norms of residuals were 1.8034E-04 for loess and 1.714E-04 for the SGF respectively (Fig. 6). For the ZnO device, the norms of residuals were 8.47E-07 and 8.27E-07 for the loess and SGF methods, respectively (Fig. 7). Further analysis of the raw data and the fitted data showed that the two smoothing methods (loess and SGF) present very good confidence bounds when fitted over the raw data for both devices (see Fig. 8 and Fig. 9).

VI. Conclusion and Future Work

This study has explored signal response, and methods for extracting the desired digital signal while maintaining the shape and resolution of that signal. A simple procedure to test different polynomial models, with confidence bounds, on the raw data was developed for easy application to quantised gas sensor response data. Curve fitting approaches were used to validate the results of three possible methods.

Of the six methods investigated, and as expected, it was found that the SGFs give the best resolution and best maintain the shape of the signal and therefore provided the best approximation of the sensor response. After testing SGF with various polynomial models, the ninth degree of the polynomial model was observed to provide the best fit to the raw data for both the SnO2 and ZnO sensor devices. For both sensor devices, the raw data were observed to fall within the confidence bounds simulated from the polynomial models. This is a promising result and future work will involve testing the SGF signal pre-processing method on signals produced from an array of sensors.

References

[1] I. García-Pérez, M. Vallejo, A. García, C. Legido-Quigley, and C. Barbas, "Metabolic fingerprinting with capillary electrophoresis," J. Chromatogr. A, vol. 1204, issue 2, 2008, pp. 130-139.

[2] R. Gutierrez-Osuna, H. T. Nagle, B. Kermani, and S. S. Schiffman, "Signal Conditioning and Preprocessing," Handbook of Machine Olfaction: Electronic Nose Technology. Wiley-VCH Verlag GmbH & Co. KGaA, pp. 105-132, 2004.

[3] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Numerical Recipes: The Art of Scientific Computing," Cambridge, New York Cambridge University Press, 2007.

[4] S. J. Orfanidis, "Introduction to Signal Processing," New Jersey, USA: Prentice Hall, Inc, 1996.

[5] R. Gutierrez-Osuna, "Pattern analysis for machine olfaction: a review," Sensors Journal, IEEE, vol. 2, issue 3, 2002, pp. 189-202.

[6] A. Gutierrez-Galvez, "Coding and learning of chemosensor array patterns in a neurodynamic model of the olfactory system," Ph.D., Texas A&M University, United States – Texas, 2006.

[7] E. Szymańska et al., "Increasing conclusiveness of metabonomic studies by cheminformatic preprocessing of capillary electrophoretic data on urinary nucleoside profiles," Journal of Pharmaceutical and Biomedical Analysis, vol. 43, issue 2, 2007, pp. 413-420.

[8] K. J. Johnson, B. W. Wright, K. H. Jarman, and R E.. Synovec, "High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis," Journal of Chromatography A, vol. 996, issues (1–2), 2003, pp. 141-155.

[9] C. Perrin, B. Walczak, and D. L. Massart, "The Use of Wavelets for Signal Denoising in Capillary Electrophoresis," Analytical Chemistry, vol. 73, issue 20, 2001, pp. 4903-4917.

[10] L. Bao, J. Mo, and Z. Tang, "The Application in Processing Analytical Chemistry Signals of a Cardinal Spline Approach to Wavelets," Analytical Chemistry, vol. 69, issue 15, 1997, pp. 3053-3057.

[11] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," Analytical Chemistry, vol. 36, issue 8, 1964, pp. 1627-1639.

[12] M. Zuppa, C. Distante, P. Siciliano, and K. C. Persaud, "Drift counteraction with multiple self-organising maps for an electronic nose," Sensors and Actuators B: Chemical, vol. 98, issues 2–3, 2004, pp. 305-317.

[13] MathWorks(R) MATLAB: Curve Fitting Toolbox, Filtering and Smoothing Data. Available from: http://www.mathworks.com/help/curvefit/smoothing-data.html [retrieved: April 2014].

[14] MathWorks(R) MATLAB: Wavelet Toolbox, Multivariate Wavelet Denoising. Available from: http://www.mathworks.com/help/wavelet/examples/multivariate-wavelet-denoising.html [retrieved: April 2014].

[15] R. Schafer, "What Is a Savitzky-Golay Filter?," Signal Processing Magazine, IEEE, vol. 28, issue 4, 2011, pp. 111-117.

[16] D. J. Thornley, "Novel Anisotropic Multidimensional Convolutional Filters for Derivative Estimation and Reconstruction," IEEE International Conference on Signal Processing and Communications ICSPC 2007, 2007, pp. 253 - 256.

[17] R. G .D. Steel and J. H. Torrie, "Principles and Procedures of Statistics with Special Reference to the Biological Sciences," McGraw Hill, pp. 187-287, 1960.

 ISBN: 978-1-61208-365-0