Gaussian Process Regression and Monte Carlo Simulation to Determine VOC Biomarker Concentrations Via Chemiresistive Gas Nanosensors

Paula Angarita Rivera
Department of Mechanical and Energy Engineering, Indiana Univerisity-Purdue University of Indianapolis
Indianapolis IN, United States
pangarit@iu.edu

Mark Woollam
Department of Chemistry and Chemical Biology, Indiana Univerisity-Purdue University of Indianapolis
Indianapolis IN, United States
mwoollam@iu.edu

Amanda P. Siegel
Department of Chemistry and Chemical Biology, Indiana Univerisity-Purdue University of Indianapolis
Indianapolis IN, United States
apsiegel@iupui.edu

Mangilal Agarwal*
Department of Mechanical and Energy Engineering, Indiana Univerisity-Purdue University of Indianapolis
Indianapolis IN, United States
agarwal@iupui.edu

***Abstract*— Utilizing chemiresistive gas sensors for volatile organic compound (VOC) detection has been a growing area of investigation in the last decade. VOCs have been extensively studied as potential biomarkers for biomedical applications as they are byproducts of metabolic pathways which are dysregulated by disease. Therefore, sensor arrays have been fabricated in previous studies to detect VOC biomarkers. In the process of testing these sensors, it is highly advantageous to quantify the concentration of the VOC biomarkers with high accuracy to diagnose the disease with high sensitivity and specificity. To investigate, analyze, and understand the relation between the concentrations of the VOC to the sensor resistance response, Gaussian Process (GP) models were implemented to predict the behavior of the data with respect to the resistance when the sensor is exposed to a range of concentrations of VOCs. Additionally, the relation between the concentration and resistance of the sensor was studied to predict the concentration of the VOC when a resistance is obtained. Monte Carlo Simulation Sampling from the GP model was utilized to generate data to further understand the trend. The results demonstrated that the relation between the concentration and resistance is linear. The model was tested with sampling data and its accuracy was evaluated.**

***Keywords*— Sensors, Resistance, Gaussian Process Regression, Monte Carlo Simulation**

I. Introduction

Volatile organic compounds (VOCs) are airborne molecules emitted from manmade and organic structures. The study of VOCs first gained prominence when it was shown specific VOCs emitted by manmade structures demonstrated significant health risks, leading to the need to monitor VOC levels in buildings [1]. More recently, VOCs levels have been found to be able to quantitatively and non-invasively predict results in a wide variety of fields from food spoilage to soil status to even be able to be used as biomarkers for human health [2]. The growing value of detecting VOCs in many fields has led to a need for accurate gas sensors to monitor VOC concentrations [3]. One such type are chemiresistive gas sensor arrays. These devices have demonstrated numerous advantages such as high sensitivity, cross-selectivity, cost effectiveness, low degradation, and facile integration into complex systems [4]. For the field of medical diagnostic, VOCs are endogenous metabolites that are noninvasively expressed in alveolar air and different biofluids. Researchers have demonstrated that there is a relation between VOCs and diseases such as breast cancer [5], [6], prostate cancer [7], lung cancer [8], diabetes [9], and other diseases. VOCs that are exhaled from patients can contain biologically useful information for disease diagnostics [10]. Chemiresistive gas sensors can detect VOCs through a measurable change in resistance when the VOCs are exposed to the sensor. These sensors are a powerful technology that have the potential for many different biomedical diagnostic applications. Chemiresistive gas sensors are usually integrated into a sensor array to enhance their selectivity and sensitivity to different VOCs. The sensors in the array are coated with a combination of different conductive materials and polymers to create a nanocomposite that can sense the VOCs in human breath [11]. The fabrication and application of these sensors face many challenges, such as tolerance for humidity, life span of the nanocomposite, degradation of the electrodes, and ensuring appropriate limits of detection for the sensors [12].

Currently available devices such as the electronic nose or "e-nose" can detect VOCs, but have not demonstrated a high selectivity towards targeted VOC, therefore, they rely in machine learning and pattern recognition algorithms to differentiate VOCs. [13]. However, these sensors can differentially adsorb a wide range of VOCs. There is a need to develop sensors that can detect targeted VOCs with higher sensitivity and selectivity. To accomplish this, the resistance response of sensors must be collected at a range of concentrations. The purpose of this study is to understand the relationship between the concentration of the VOC with the measured resistance of the sensor. Data from sensor testing was collected by using a single VOC at three different concentrations (5ppm, 10ppm and 15ppm).

Linear regression models are constructed for one independent variable x, where there is one dependent variable y [14]. This model could potentially work with the data proposed for sensor design and development. However, a constraint that it would face is that if the data do not fit in the regression model, then it would be considered to be an error. The aim for this study was to uncouple the data between the concentrations and understand the overall relations between concentration of the VOC and resistance of the sensor. In order to do this, two different methods were utilized: Gaussian Process (GP) regression and Monte Carlo Simulation (MCS) sampling from the GP application. GP regression is a supervised learning method that can efficiently solve regression and probabilistic classification problems. This method has a high accuracy response for simple and complex models. It also manages the kernels variable in the calculation that can be predefined, or it can be custom built depending on the data [15]. GP regression has been used in a wide range of applications including the design of composite material parts under dynamic loading [16], blast mitigation [17], crashworthiness design [18], design of lithium-ion batteries [19], multi-objective optimization [20], and multi-fidelity design optimization [21].

MCS sampling is also a probabilistic method for randomly sampling a probability distribution. This helps to generate data that could help optimize the model which increases its accuracy and effectiveness when the density is estimated [22]. MCS sampling from GP will be utilized to generate data points and understand how the model would behave when there is more data to adjust and update to make it more accurate. When the VOCs are exposed to the sensor, the data acquisition system records the change of resistance values continuously. However, when the data are being processed, each point is considered to be discrete to investigate critical sensor response. Therefore, MCS is a convenient and efficient methodology to generate a model for the missing data when processing.

II. Experimental Procedure

A. *Chemiresistive Gas Sensor Fabrication and Testing*

Sensors were fabricated in gold patterned over a silicon (Si)/silicon dioxide (SiO_2) substrate using a method similar to that described previously [4], with modifications that are not relevant for the modeling analysis presented here. The interdigitated electrodes (IDEs) are fabricated using photolithography. Polyetherimide (PEI) based nanoconductive material was drop casted and spin coated over the IDEs of the sensor and dried in a vacuum for 48hrs. The sensors will be used in diverse environments, therefore they need to be capable to detect VOCs of interest without interference of environmental gases [23]. The 14 sensors were tested in a simulated environment that mimics the ambient environment where the sensor would be utilized in the real world. Figure 1 represents an illustration of the sensor testing system which includes gases (air, water vapor and VOCs), flowmeters to regulate the flow of all gases exposed to the sensor, a gas mixer, a testing chamber where the sensors are located and a Keithley 2701 Digital Multimeter/Data Acquisition/ Data Logging system. The sensors are exposed to different concentrations of VOCs (5, 10 and 15 parts per million (ppm)) in order to investigate the relationship between the concentration of the VOC and the change in resistance measured by the sensor. Change in resistance divided by baseline resistance ($\Delta R/R_0$) were used to generate the 14 data readings used for the analysis.

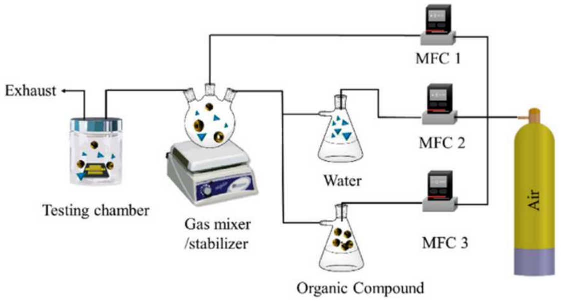

Figure 1 Schematic of experimental gas sensor testing setup [4].

B. *Gaussian Process Regression*

GP regression is a supervised learning method that employs GPs as probabilistic predictive models [15], [24], [25]. A GP is a collection of indexed random variables in which any finite set of them has a joint Gaussian distribution. A GP $f(\mathbf{x})$ is specified by two features: a mean function $m(\mathbf{x})$ and a covariance function $k(\mathbf{x},\mathbf{x}')$. A GP is denoted as

$$f(\mathbf{x}) \sim GP\big(m(\mathbf{x}), k(\mathbf{x},\mathbf{x}')\big). \qquad (1)$$

The mean and covariance functions need careful selection in order to capture any prior knowledge about the system to be modelled. In this work, two GP models were trained with different mean functions: a constant mean function and a linear mean function. The GP model with a constant mean function is a common practice in the machine learning community. This assumption is acceptable when there is no knowledge about the behavior of the data. In the second GP model, which uses a linear mean function, the aim is to exploit the prior knowledge about the behavior of the VOC concentration and the measured change in resistance. Linearity is expected as a result of a linear behavior in the operable range of the sensor.

GPs are good models to use because the code can handle relatively noisy observations. The following additive model was employed

$$y = f(\mathbf{x}) + \epsilon, \qquad (2)$$

where $f(\mathbf{x})$ is a GP that captures the behavior of the data and $\epsilon \sim N(0,\sigma^2)$ is the noise in the data with variance σ^2.

Given a set of n training samples $\mathbf{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^n\}$ with observations $= \{\mathrm{y}^1, \dots, \mathrm{y}^n\}$, the joint prior distribution of and the predicted values $\mathbf{f}_*$ at test points $\mathbf{X}_*$ is

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X},\mathbf{X}) + \sigma_n^2 \boldsymbol{I} & K(\mathbf{X},\mathbf{X}_*) \\ K(\mathbf{X}_*,\mathbf{X}) & K(\mathbf{X}_*,\mathbf{X}_*) \end{bmatrix}\right), \qquad (3)$$

where the matrices $K(\mathbf{X},\mathbf{X})$, $K(\mathbf{X}_*,\mathbf{X})$, and $K(\mathbf{X}_*,\mathbf{X}_*)$ are generated by the evaluation of the covariance function $k(\mathbf{x},\mathbf{x}')$.

The expression above corresponds to the prior distribution of a GP with a zero mean function. The predictive equations for this regression model are

$$\mathbf{f}_*|\mathbf{X},\mathbf{y},\mathbf{X}_* \sim N(\bar{\mathbf{f}}_*, \operatorname{cov}(\mathbf{f}_*)), \tag{4}$$

where $\bar{\mathbf{f}}_*$ is the predictive mean

$$\bar{\mathbf{f}}_* = K(\mathbf{X}_*,\mathbf{X})\left[K(\mathbf{X},\mathbf{X}) + \sigma_n^2\boldsymbol{I}\right]^{-1}\mathbf{y} \tag{5}$$

and cov(**f**) is the covariance.

$$\operatorname{cov}(\mathbf{f}_*) = K(\mathbf{X}_*,\mathbf{X}_*) - K(\mathbf{X}_*,\mathbf{X})[K(\mathbf{X},\mathbf{X}) + \sigma_n^2\boldsymbol{I}]^{-1}K(\mathbf{X},\mathbf{X}_*). \tag{6}$$

In the case of GP models with non-zero mean functions, they can be modelled as

$$g(\mathbf{x}) = f(\mathbf{x}) + \mathbf{h}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta}, \tag{7}$$

where $\mathbf{h}(\mathbf{x})$ is a set of basis functions and $\boldsymbol{\beta}$ is a weighting vector with a prior $\boldsymbol{\beta} \sim N(\mathbf{b}, B)$. The resulting GP is

$$g(\mathbf{x}) \sim GP\Big(\mathbf{h}(\mathbf{x})^{\mathrm{T}}\mathbf{b}, k(\mathbf{x},\mathbf{x}') + \mathbf{h}(\mathbf{x})^{\mathrm{T}}B\mathbf{h}(\mathbf{x}')\Big). \tag{8}$$

The matrices H and H_* collect the values of $\mathbf{h}(\mathbf{x})$ for the training and test data. The predictive equations for the model then are

$$\bar{\mathbf{g}}_* = \bar{\mathbf{f}}_* + R^{\mathrm{T}}\bar{\beta} \tag{9}$$

and

$$\operatorname{cov}(\mathbf{g}_*) = \operatorname{cov}(\mathbf{f}_*) + R^{\mathrm{T}}\left(HK_y{}^{-1}H^{\mathrm{T}}\right)^{-1}R, \tag{10}$$

where $K_y = K(\mathbf{X},\mathbf{X}) + \sigma_n^2\boldsymbol{I}$, $K_* = K(\mathbf{X},\mathbf{X}_*)$, $R = H_* - HK_y{}^{-1}K_*$, and $\bar{\beta} = \left(B^{-1} + HK_y{}^{-1}H^{\mathrm{T}}\right)^{-1}(HK_y{}^{-1}\mathbf{y} + B^{-1}\mathbf{b})$.

In the expressions above, $\bar{\mathbf{g}}_*$ is the predictive mean and $\operatorname{cov}(\mathbf{g}_*)$ is the joint covariance of the GP model. In the special case of GP with a linear regression function, the linear regression function is incorporated into the matrix $\mathbf{h}(\mathbf{x})^{\mathrm{T}}$.

The covariance function for the model presented is

$$k(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_p - x_q)^2\right) + \sigma_n^2\delta_{pq}, \tag{2}$$

where the first term is the squared-exponential covariance function, which assumes a smooth behavior of the data, and the second term captures to the noise in the data. δ_{pq} is the Kronecker delta function. The parameters $\mathbf{b}$, B, σ_f^2, σ_n^2, and l are known as the hyperparameters of the GP model. A common approach to estimate the values of the hyperparameters is the maximum likelihood estimation [26]. For this study, we use the Statistics and Machine Learning Toolbox of MATLAB (which incorporates maximum likelihood estimation) to train the GP regression models. The inputs for the GP models are the concentrations of the VOC and the outputs correspond to the normalized change in the sensor's resistance.

C. *Monte Carlo Sampling from Gaussian Process Regression*

The GP regression model described in the previous section allows predictions of the normalized change in resistance for a given VOC concentration. Next, MCS and GP regression is used to solve the inverse problem, i.e., to predict the most probable concentration range of VOC given a measurement of the change in the resistance of the sensor. MCS is a numerical method to solve problems that have a probabilistic implementation. It relies on repeated random sampling and statistical analysis to estimate the probabilistic descriptors of the targets of study [22, 27]. These descriptors are derived by drawing samples of the inputs that follow a normal distribution and propagating their effect through the system. This part of the study employed normalized resistance values of R_0 = 1.5, R_0 = 2.8, and R_0 = 4.2, which corresponded to the prediction of the GP model at 5 ppm, 10 ppm and 15 ppm, respectively. These resistances values were selected to determine if samples from the GP model follow the distribution of the experimental results. After sampling, histograms of the concentration were generated using a window that encloses the samples for each tested resistance [28]. The window had a size of 2 *h* and it was located at ± *h* from R_0. The value of *h* is 0.01. From the histogram, relevant statistical information was obtained such as most probable value, range, and 95% CI.

The solution of the inverse problem followed a three-step process: (1) Use the GP regression model to determine normal distribution, i.e., the predicted mean and variance, of the change in resistance for different VOC concentrations. (2) Draw samples from each normal distribution. (3) Estimate the mean concentration, given a defined resistance change and determine a 95% confidence interval by capturing the samples that fall within a prescribed threshold.

III. Results and Discussion

A. *Gaussian Process Regression Model*

The experimental data used to create the Gaussian model include for the Training Samples a set of 3 concentrations (***X*** = 5, 10, and 15 ppm***)*** and 14 data points ***y*** per concentration. Range of ***y*** varies from 0% to roughly 5% $\Delta R/R_0$. The GP regression model has two components: the first component contains the regression function (mean function), and the second component has the correlation function. The mean function provides information about the general behavior of the data. The correlation function, on the other hand, explains the local behavior of the data. In the figures below, the green line represents the model, and the red line is the 95% confidence interval (CI) (±2σ).

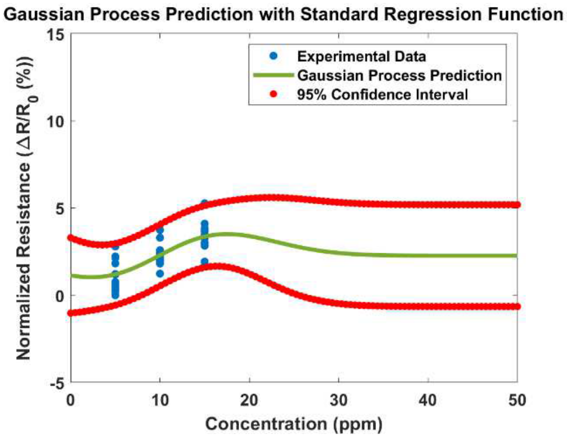

Figure 2 Gaussian process prediction for the standard model.

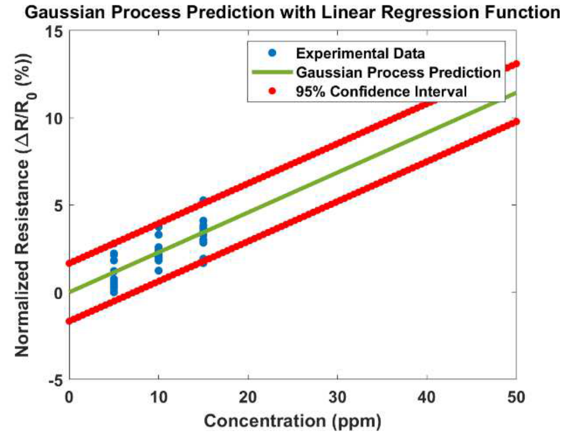

Figure 3 Gaussian process prediction for the linear model with the mean and standard deviation from the experimental data.

The standard GP regression model is represented in Figure 2. This model starts with an initial assumption which is a constant value. However, as it gets closer to the data, the model is optimized to increase prediction accuracy. Data were obtained for three different concentrations: 5ppm, 10ppm and 15ppm. As the model reaches the area of the data, it starts making more meaningful predictions. In the range from 5ppm to 15ppm, the slope of the model is directly proportional to the change of resistance of the sensor. However, when it reaches 20 ppm the model does not have any more data that helps update the initial assumption. Therefore, the model goes back to the initial assumption, which was a constant resistance. From previous knowledge of the data behavior, the standard GP regression model was updated to enhance the model. GP has an advantage that any prior knowledge of the model can be used to create the most accurate model possible for the type of data given.

In Figure 3 the model was updated with the prior knowledge obtained from the standard GP prediction that it was considered to be a linear regression model. As it can be seen, the model is able to estimate the behavior of the data in a range of concentrations where experimental data are missing. This demonstrates that when the model is known to be linear, it gets more accurate with respect to the data points that are unknown. Figure 3 shows a linear relationship between the concentration of the VOC and the resistance. As the concentration of the VOC increases, the change in resistance increases as well. The uncertainty of this model is illustrated by creating a 95% CI. This means that there is a 95% confidence that the samples for those concentrations will fall within that region. There are many ways to enhance the prediction accuracy of the linear model. However, the addition of incremental data points would help the most with the performance of the model. The more data points that the model has to work with, the higher the accuracy. The mean value for each of the concentration's sets (using GP regression) are used to analyze and understand the relation of the concentration of the VOC with respect to a given resistance of the sensor. For 5 ppm the R_0=1.5, 10 ppm the R_0=2.8, and 15ppm the R_0=4.2.

The GP regression model function becomes more accurate in predicting behavior in the data as more samples are added to the function. As it can be seen in these results, there is insufficient data for describing the complete behavior, especially regarding high VOC concentrations. For example, the model predicts that the data model would remain linear, even for higher concentrations. This is a potential concern, because at some point the sensor could be exposed to a highly concentrated VOC and this could cause sensor saturation, requiring modification of the model. Therefore, it is important to determine at what concentration the sensor reaches a plateau.

B. *Monte Carlo Simulation Sampling from Regression Model*

MCS sampling was used to generate data for VOC concentrations ranging from 0 ppm to 50 ppm with a window size *h* of 0.01. Each of the sets of data had a known mean and standard deviation which were used to sample and produce 3000 data points for each concentration. The normal distribution has a support that contains all real numbers ($\mathbb{R}$). Therefore, some samples might present negatives concentrations and negative resistances. A future work includes the use of a different stochastic process (regression model) with finite support (positive values). However, the mode of the mass of the GP is positive. Figure 4 shows the linear relationship between the measured change in resistance and concentration for the normally distributed Monte Carlo simulated data. This aspect of the study addresses the second question: if the resistance is given, what is the most probable concentration of the VOC. Therefore, in this case R_0 is considered to be a constant.

For the purpose of simplification of this study, only three constant R_0 were analyzed (values found in the Gaussian Process Regression model for the three experimental concentrations). The most probable VOC concentration was calculated based on a given change in resistance. In this case, the change in resistance is considered to be constant and the normal distributions of all the concentrations are considered to be cut-off. The constant then has an upper and lower concentration range that captures the number of data points that it has for that specific change in resistance. Using this information, a histogram was produced to observe the number of repetitions that each of the concentrations have for the given resistance. The average concentration was calculated along with the 95% CI for constant resistance values of R_0 = 1.5, R_0 = 2.8 and R_0 = 4.2, respectively. In this way, a concentration range can be adequately predicted based on a measured change in resistance via the chemiresistive gas sensor.

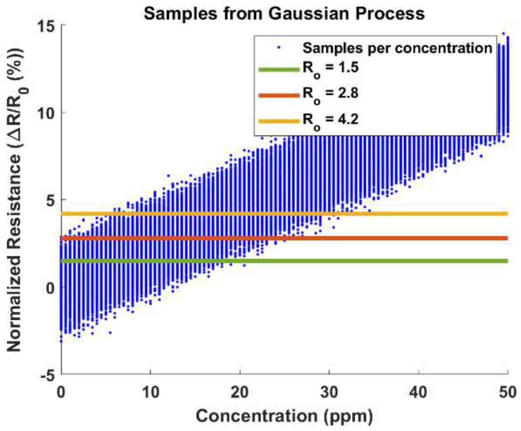

Figure 4 Monte Carlo Sampling with three constant resistances.

Figure 4 shows the constant R_0 for the mean concentration of 5 ppm, 10 ppm and 15 ppm. Visually, the range of the most possible concentration can be hard to determine because the data show a range from 0 ppm to 25 ppm as possible concentrations for all constant R_0. This is not ideal because this makes the model less accurate due that the range of possible concentrations is too wide. The window was created and then the most probable concentration can be validated in this case because it is already known from previous results.

A way to reduce the range of possible concentrations could be by adding some limitations of what is the minimum number of repetitions that can be in the window in order to be considered to be part of the optimal concentration. Therefore, the 95% confidence interval would be helpful to find a more concise range that it is still statistically significant but is not as wide as the one that can be visually seen from the Figure 5. Limiting the window size with the 95% CI is also useful because it can reduce overlap between two discrete resistance values. This could be a challenge for the process if the 95% CI range is so wide that there are multiple concentrations that can relate to a given resistance.

In order to evaluate the accuracy and precision of the MCS Sampling with the GP regression model, the number of repetitions within the window size h = 0.01 of the constant resistance were counted. Figure 5 shows the distribution of the concentrations with respect to each of the resistance values studied. From this graph, it can be concluded that the median concentration for the given resistance R_0=1.5 in a system capable of registering concentrations from 0 to 50 ppm is 6.5 ppm (95% CI [0 ppm – 14 ppm]). From prior knowledge it is known that the actual concentration is 5 ppm. As previously mentioned, the distribution has a support $\mathbb{R}$ which in order to find the most probable concentration for the given resistance the distribution takes in consideration negative concentrations. For a given resistance of R_0=2.8, the median concentration is 12.5 ppm (95% CI [6 ppm – 19.5 ppm]), when the actual concentration is 10 ppm. Lastly, for the highest resistance which it was R_0 = 4.2, the median concentration is 18.5 ppm (95% CI [11.5 ppm – 26 ppm]). The actual concentration is considered to be 15 ppm. The median concentration values are closer to the experimental results, but the 95% CI range is very consistent across the three resistance values sampled. Moreover, the MCS values are significantly different (p-value < 0.001, Student's T-test) between studied resistance values. Sensors that show higher reproducibility in the experimental data will have much tighter confidence intervals, and therefore will generate more meaningful results.

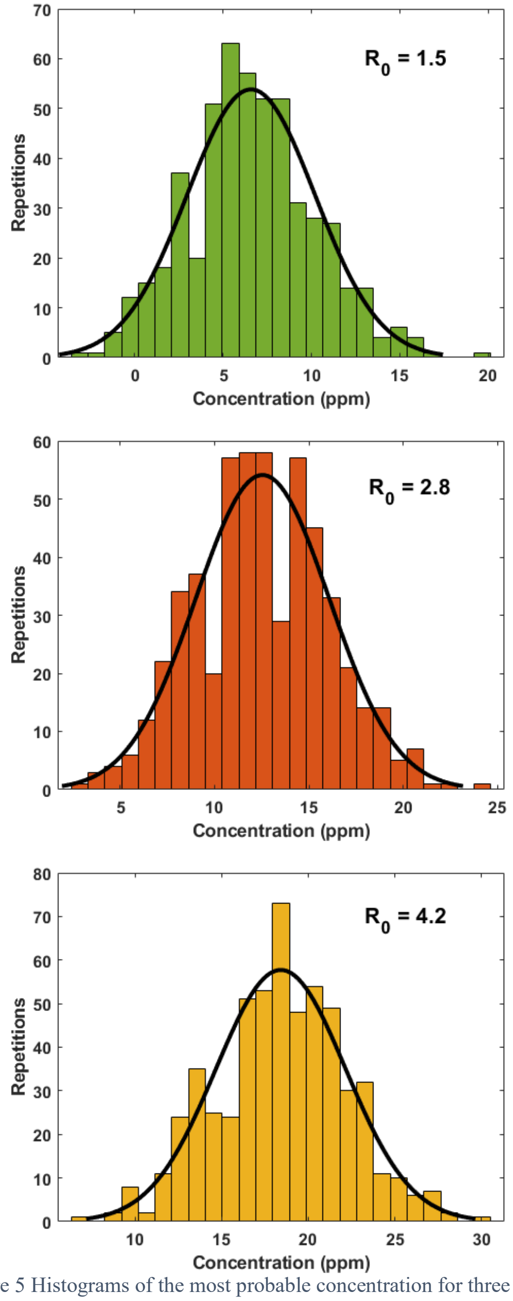

Figure 5 Histograms of the most probable concentration for three constant resistances (note discussion in text for why 5 ppm includes negative values).

IV. Conclusion

The purpose was to investigate the relationship between the measured change in resistance of a chemiresistive nanosensor and the concentration of a VOC exposed to the sensor surface. Two different methodologies were utilized to understand this relationship. From the standard GP regression model, the results showed that the data is linear in the range of the data that it was provided. However, when there is no data for a particular concentration, the model returns to its initial assumption. From the linear GP regression model, prior knowledge was applied to the model by previously demonstrating linearity with a positive correlation. This approach had much higher accuracy and had the ability to extrapolate data. One of the challenges with model was that utilizes a stationary covariance to model the experimental noise. The stationary covariance assumed that the amplitude of the noise is constant through the whole input domain. Therefore, future work includes the use of more sophisticated GP models with covariance functions capable of capturing non-stationarities in the experimental noise, i.e., noise with varying amplitude. The MCS sampling generated a new set of data that correlates with the experimental data analyzed by linear GP regression. The simulation also showed the capability to predict the concentrations of VOCs exposed to the sensor based on an constant change in resistance. This caused the overestimation of concentration because the simulated range contained data that was greater than the experimental range. These models can be applied for any sensor data considered to be linear. The MATLAB source can be updated with the corresponding data. The model can also be modified to study the sensor response at higher and lower ranges of concentrations with smaller or higher step sizes.

Acknowledgment

The authors would like to thank NSF (#1502310). A special thanks to Homero Valladares for the assistance with MATLAB and supplementary instruction. Thanks to Dr. Ali Daneshkhah and Sanskar Thakur for training of sensor fabrication and testing. Thank to Dr. Xiaoping Du for the support in the class.

References

[1] J Sundell, "On the history of indoor air quality and health," (in eng), *Indoor Air,* vol 14 Suppl 7, pp 51-8, 2004, doi: 10 1111/j 1600-0668 2004 00273 x

[2] J D Fenske and S E Paulson, "Human breath emissions of VOCs," (in eng), *J Air Waste Manag Assoc,* vol 49, no 5, pp 594-8, May 1999, doi: 10 1080/10473289 1999 10463831

[3] A Daneshkhah, S Shrestha, M Agarwal, and K Varahramyan, "Poly(vinylidene fluoride-hexafluoropropylene) composite sensors for volatile organic compounds detection in breath," *Sensors and Actuators B: Chemical,* vol 221, pp 635-643, 2015/12/31/ 2015, doi: https://doi org/10 1016/j snb 2015 06 145

[4] A Daneshkhah, S Vij, A P Siegel, and M Agarwal, "Polyetherimide/carbon black composite sensors demonstrate selective detection of medium-chain aldehydes including nonanal," *Chemical Engineering Journal,* vol 383, p 123104, 2020/03/01/ 2020, doi: https://doi org/10 1016/j cej 2019 123104

[5] M Woollam *et al.*, "Urinary Volatile Terpenes Analyzed by Gas Chromatography–Mass Spectrometry to Monitor Breast Cancer Treatment Efficacy in Mice," *Journal of Proteome Research,* vol 19, no 5, pp 1913-1922, 2020/05/01 2020, doi: 10 1021/acs jproteome 9b00722

[6] K Taunk *et al.*, "A non-invasive approach to explore the discriminatory potential of the urinary volatilome of invasive ductal carcinoma of the breast," *RSC Advances,* vol 8, no 44, pp 25040-25050, 2018, doi: 10 1039/c8ra02083c

[7] T Khalid *et al.*, "Urinary Volatile Organic Compounds for the Detection of Prostate Cancer," (in eng), *PLoS One,* vol 10, no 11, p e0143283, 2015, doi: 10 1371/journal pone 0143283

[8] Y Saalberg and M Wolff, "VOC breath biomarkers in lung cancer," *Clinica Chimica Acta,* vol 459, pp 5-9, 2016/08/01/ 2016, doi: https://doi org/10 1016/j cca 2016 05 013

[9] A P Siegel, A Daneshkhah, D S Hardin, S Shrestha, K Varahramyan, and M Agarwal, "Analyzing breath samples of hypoglycemic events in type 1 diabetes patients: towards developing an alternative to diabetes alert dogs," (in eng), *J Breath Res,* vol 11, no 2, p 026007, Jun 1 2017, doi: 10 1088/1752-7163/aa6ac6

[10] M Mansurova, B E Ebert, L M Blank, and A J Ibáñez, "A breath of information: the volatilome," *Current Genetics,* vol 64, no 4, pp 959-964, 2018/08/01 2018, doi: 10 1007/s00294-017-0800-x

[11] G Konvalina and H Haick, "Sensors for Breath Testing: From Nanomaterials to Comprehensive Disease Detection," *Accounts of Chemical Research,* vol 47, no 1, pp 66-76, 2014/01/21 2014, doi: 10 1021/ar400070m

[12] A H Jalal, F Alam, S Roychoudhury, Y Umasankar, N Pala, and S Bhansali, "Prospects and Challenges of Volatile Organic Compound Sensors in Human Healthcare," *ACS Sensors,* vol 3, no 7, pp 1246-1263, 2018/07/27 2018, doi: 10 1021/acssensors 8b00400

[13] W Cuypers and P A Lieberzeit, "Combining Two Selection Principles: Sensor Arrays Based on Both Biomimetic Recognition and Chemometrics," (in English), *Frontiers in Chemistry,* Mini Review vol 6, no 268, 2018-August-02 2018, doi: 10 3389/fchem 2018 00268

[14] N Altman and M Krzywinski, "Simple linear regression," *Nature Methods,* vol 12, no 11, pp 999-1000, 2015/11/01 2015, doi: 10 1038/nmeth 3627

[15] E Schulz, M Speekenbrink, and A Krause, "A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions," *Journal of Mathematical Psychology,* vol 85, pp 1-16, 2018/08/01/ 2018, doi: https://doi org/10 1016/j jmp 2018 03 001

[16] H Valladares, A Jones, and A Tovar, "Surrogate-Based Global Optimization of Composite Material Parts under Dynamic Loading," 2018 [Online] Available: https://doi org/10 4271/2018-01-1023

[17] H Valladares and A Tovar, "Multilevel Design of Sandwich Composite Armors for Blast Mitigation using Bayesian Optimization and Non-Uniform Rational B-Splines," presented at the SAE WCX Digital Summit, 2021

[18] K Liu, T Wu, D Detwiler, J Panchal, and A Tovar, "Design for Crashworthiness of Categorical Multimaterial Structures Using Cluster Analysis and Bayesian Optimization," *Journal of Mechanical Design,* vol 141, pp 1-44, 09/12 2019, doi: 10 1115/1 4044838

[19] H Valladares *et al.*, "Bayesian Optimization of Active Materials for Lithium-ion Batteries," presented at the SAE WCX Digital Summit, 2021

[20] H Valladares and A Tovar, "A Simple and Effective Methodology to Perform Multi-Objective Bayesian Optimization: An Application in the Design of Sandwich Composite Armors for Blast Mitigation," in *ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2020, vol Volume 11A: 46th Design Automation Conference (DAC), V11AT11A056, doi: 10 1115/detc2020-22564 [Online] Available: https://doi org/10 1115/DETC2020-22564

[21] H Valladares and A Tovar, "Design Optimization of Sandwich Composite Armors for Blast Mitigation Using Bayesian Optimization with Single and Multi-Fidelity Data," 2020 [Online] Available: https://doi org/10 4271/2020-01-0170

[22] S Raychaudhuri, "Introduction to monte carlo simulation," presented at the 2008 Winter simulation conference, 2008

[23] Z Geng, F Yang, X Chen, and N Wu, "Gaussian process based modeling and experimental design for sensor calibration in drifting environments," *Sensors and Actuators B: Chemical,* vol 216, pp 321-331, 2015/09/01/ 2015, doi: https://doi org/10 1016/j snb 2015 03 071

[24] C E R a C K Williams, "Gaussian Process for Machine Learning," ed: Adaptive Computation and Machine Learning: MIT Press Cambridge, 2006

[25] T D I Couckuyt, and P Demeester, "ooDACE toolbox: a flexible object-oriented Kriging implementation," *The Journal of Machine Learning Research,* pp 3183–3186, 2013

[26] P I Frazier A tutorial on bayesian optimization

[27] G A Bird, "Monte Carlo Simulation in an Engineering Context," presented at the AIAA, New York, 1981, 1, 1981

[28] Z Fan *et al.*, "Monte Carlo Optimization for Sliding Window Size in Dixon Quality Control of Environmental Monitoring Time Series Data," *Applied Sciences,* vol 10, no 5, p 1876, 2020 [Online] Available: https://www mdpi com/2076-3417/10/5/1876