【研究論文】 【令和3～4年度 県単独試験研究】

ガウス過程回帰を用いたガスセンサ濃度推定回帰モデルの構築
ー 機械学習を用いた時系列データ解析 ー

岩沢 正樹、阿部 宏之
機械電子情報技術部

半導体式ガスセンサにより測定した時系列データに対して、複数成分に対するガス濃度予測用回帰モデルの構築に取り組んだ。時系列データからの特徴量抽出には、変化点検知アルゴリズムとしてガス濃度を予測する時のリアルタイム性向上が期待できるChange Finderを採用した。回帰モデルの構築には、ガウス過程回帰を用いた。学習・推論した結果、検出するガスの単成分及び複数成分のガス濃度を予測する場合の回帰モデルについて、データの豊富な範囲での有効性を確認した。

キーワード: 時系列データ、回帰モデル、変化点検知、Change Finder、ガウス過程回帰、半導体式ガスセンサ

1 緒言

近年、深層学習に代表される機械学習分野の発展は著しく、その活用の幅が広がっている。特に、自然言語処理分野では、米国 OpenAI 社による ChatGPT の公開が社会に大きな影響を与えている [1]。また、日本では、経済産業省が、AI 導入ガイドブック [2]の作成や、より広く DX セレクション（中堅・中小企業等の DX 優良事例選定）[3]などを通じた中小企業への AI 導入を推進している。これらの動きの中、企業の製造現場においては、設備や装置の寿命予測や異常検知などの重要性が高い課題へ、機械学習の適用が期待され、今後、取り組む企業からの支援依頼が予想される。その中では、深層学習などの適用も考えられるが、機械学習のアルゴリズムは豊富であり、そのほかの様々なアルゴリズムを適材適所に活用することが重要である。そのためには、共通の事例に対して、様々なアルゴリズムを適用し、それぞれの長所と短所の理解を深めることが、当センターの支援活動において重要である。

そこで、当センターで開発中の半導体式ガスセンサ [4]で測定した電流値の時系列データを活用して、ガス濃度予測の回帰モデル構築の検討をおこなった。本報告では、先の報告 [5]で採用したニューラルネットワークによる回帰モデルとは異なるアプローチであるガウス過程回帰 [6]、[7]に注目し、回帰モデル構築の検討を実施した。

2 実験と解析方法

2.1 データの測定

開発中のガスセンサでは、検出対象ガスに対するセンサの出力電流の変化でガス濃度を検出する [5]。実際の測定では、2 チャンネルの半導体パラメータアナライザを用いたため、1 枚のガスセンサ基板上に形成されている 6 本のセンサ素子のうち 2 本からの出力電流値がそれぞれ同時に記録される。具体的には、ガスの導入・排出が可能な配管の付いた金属製の容器内にガスセンサを設置し、空気を流し、電流値が安定した時間から電流値の記録を開始した。一定の時間が経過した後に、空気から検出対象ガスに切り替えた。さらに一定の時間が経過した後に、検出対象ガスを流す前の出力電流値に戻すための回復ガスとして空気を流した。各ガスの流量を 50 ml/min とした。図1に 1 回の測定で得られた 2 本のセンサ素子の典型的な応答特性を示す。測定開始後、経過時間 100 秒、及び 300 秒で空気から検出対象ガスへ切り替え、経過時間 200 秒、及び 400 秒で検出対象ガスから空気に切り替え、経過時間 500 秒まで出力電流の記録を継続した。ここで、使用した検出対象ガスは、市販の 4 種混合ガス（一酸化炭素 CO：0.30％、酸素 O_2：20.30％、ヘリウム He：10.20％、窒素 N_2：69.20％）である。本稿には示していないが、濃度の異なる 7 種類の検出対象ガスに対して、2 本のセンサ素子で出力電流値の測定を行った。測定データは、時間と 2 本のセンサ素子の出力電流値がセットになった（ペアになった）時系列データで、2 回のガス濃度検出が 1 つのファイルに記録されている。

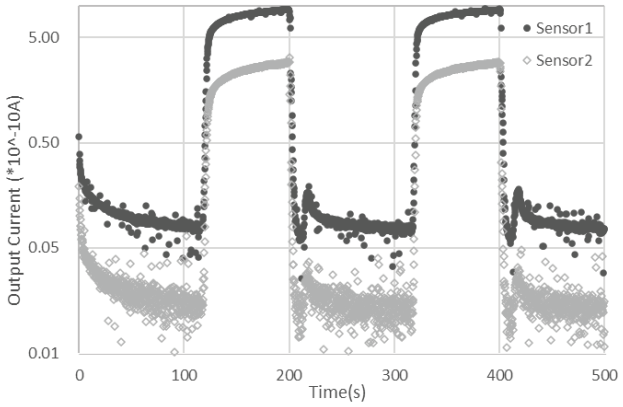

図1 2本のセンサ素子の典型的な応答特性

2.2 データセットの作成

本研究で使用したデータセットは、先の報告[5]と同様に、前述で得られたデータをセンサ素子1本ずつのデータに分離させた。このことにより、入力センサ数を変化させることが可能となった。次に、スパイクノイズを除去した後、先の報告と同様にChange Finder[8)-10)]で検出したフラット部(図2●点)の値を特徴量として採用した

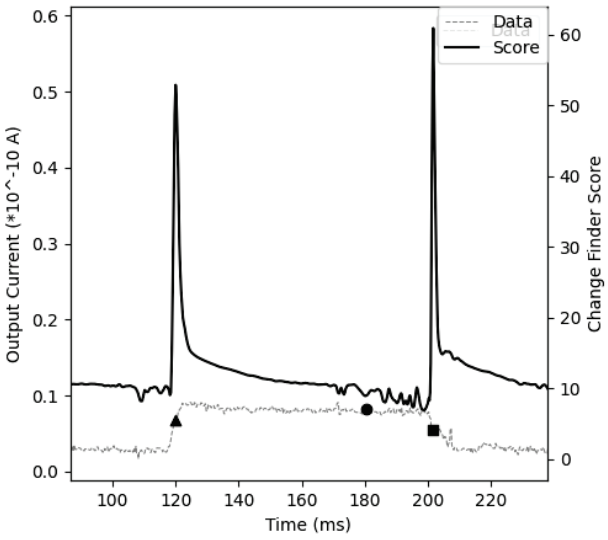

図 2 Change Finder による変化点検知

検出した特徴量から作成したデータセットの検出対象ガスについて7種類の CO ガス濃度、濃度毎のデータ数、及び回帰モデル構築に対する計算時の用途を表1に示す。測定データ中、CO ガス濃度が切り替わるタイミングのデータを除去している。CO ガス濃度7種類のうち、5 種類を学習に、残り 2 種類を予測値の推論に利用した。

表1 COガス濃度毎のデータの数とその用途

	CO濃度	センサ毎のデータ数	用途
1	0.30%	5	学習
2	0.10%	5	学習
3	0.05%	5	学習
4	0.04%	5	推論
5	0.03%	5	学習
6	0.02%	5	推論
7	0.01%	5	学習

2.3 ガウス過程回帰の概要

今回採用したガウス過程回帰[6)、7)]について概説する。ガウス過程回帰は、無限次元のガウス分布とも言われるガウス過程を回帰問題に適用した、ノンパラメトリックなモデルである。回帰モデルの構築以外にも、実験計画や最適化などへの応用が可能な手法である。

簡略化のため、入力1次元x・出力1次元yの場合を考える。任意のnについて、入力を$X = (x_1, x_2, \ \ldots, x_n)$とし、対応する出力を$Y = (y_1, y_2, \ \ldots, y_n)$とした時、どんな入力$X$についても出力$Y$の同時分布$p(y)$が多変量正規分布(この場合は、$n$次元)に従う時、$x$と$y$の関係は、関数の分布であるガウス過程$GP$に従うと定義される。ガウス過程は、平均関数$\mu(x)$とカーネル関数$k(x, x')$により特徴づけられ、

$$f \sim GP\left(\mu(x), \ k(x, x')\right)$$

と書き、ガウス過程GPから生成された関数fにより

$$y = f(x)$$

となる。

カーネル関数は、正規分布における分散(多変量正規分布の場合は、共分散)に相当するものである。様々なカーネル関数が提案されており、本報告では次式に示すRBFカーネルを利用した。

$$k(x, x') = \theta_1 \exp\left(-\frac{|x - x'|^2}{\theta_2}\right)$$

ここで、θ_1とθ_2はハイパーパラメータである。このカーネル関数の定義の仕方により、ガウス過程により生成される関数fが特徴づけられる。カーネル関数単体だけでなく、それらの和や積もカーネル関数として利用できる。このカーネル関数kを用いて、要素を

$$K_{i,j} = k(x_i, x_j)$$

とする行列Kが定義され、カーネル行列と呼ぶ。

既知の入力X、及び出力Yがガウス過程$GP\left(\mu(x)、\right.$ $\left.k(x、x')\right)$に従っているとき、任意の観測点x^*の平均と分散は、

$$\mu(x^*) = {k_*}^T K^{-1} Y$$
$$\sigma(x^*)^2 = k(x^*、x^*) - {k_*}^T K^{-1} k_*$$

と計算できる。ここで、

$$k_* = \left(k(x^*、x_1)、k(x^*、x_2)、\ \ldots\ 、k(x^*、x_n)\right)^T$$

である。この関係を用いることで、既知の値から未知の値を予測することが可能となり、回帰モデルを構築することができる。これらの式から、カーネル関数の選択方法により、フィッティングするために生成される関数が変化することがわかる。すなわち、回帰モデルを構築するときは、対象データの特徴を踏まえたカーネル関数の選択が重要となる。

2.4 ニューラルネットワークとガウス過程回帰の比較

今回のデータセットのケースにおいて、ガウス過程回帰、及び先の報告[5)]で活用したニューラルネットワークによる回帰モデル構築との比較をあえて試みる。それは、アルゴリズムの考え方が異なるため直接比較はできないが、アルゴリズムの特徴を明確にするためである。比較結果を、表2に示す。

表2 ニューラルネットワークとガウス過程回帰の概括的比較

	ニューラルネットワーク（NN）	ガウス過程回帰（GP）
モデルサイズ	△（大きい）	○（小さい）
計算時間	×（長い）	◎（早い）
設計スキル	△（NN設計）	△（カーネル設計）
予想結果から得られる情報	△（誤差のみ）	○（誤差＋分散）
多出力間の関係	○（NN内包）	△（設計事項）

ニューラルネットワークと比較した場合、ガウス過程回帰は、確率分布として推論できることから、推測値の分散を得ることができるところに特徴がある。これは、ガスセンサでのガス濃度推定のように推定値の確からしさを重視しなければならないケースにおいては、非常に重要な特徴となる。また、モデルサイズが小さくなるなどの特徴を有することに加えて、今回のようにデータ数が少ない場合は、計算時間が非常に短くなる（一般のガウス過程回帰の計算量は$O(N^3)$となることが知られている）。

出力yが多出力の場合、ニューラルネットワークでは、例えば、全結合層を中間層や出力層などに含むようにネットワーク構造を設計することで、各次元間の関係を暗黙的に含める形にすることができる。しかしながら、ガウス過程回帰の場合は、その関係を設計事項として考慮する必要があり、対象データに関するドメイン知識の活用が重要となる。

ニューラルネットワークとガウス過程回帰には、それぞれ長所と短所があり、互いに置き換えるものでは無く、それぞれの適材適所で活用することが求められる。

2.5 計算環境

学習並びに推論に用いた計算機環境を以下に示す。

CPU: Intel(R) Core(TM) i7-10750H CPU
GPU: NVIDIA GeForce RTX 2060、CUDA 11.0
OS: Windows 10 Pro バージョン 22H2
（OMEN by HP 15-dh1002TX上で構築）

また、今回の実装で利用したソフトウェア環境を以下に示す。回帰モデルの構築環境は、仮想環境を構築した。また、Change Finder の実装には、pipコマンドによりインストールされるChangefinderパッケージを利用した。ガウス過程に関するパッケージには幾つかあるが、扱いの容易さなどからGPyパッケージを採用した。

Python 3.8.10
Changefinder 0.03
GPy 1.10.0

3 モデル構築の結果と考察

3.1 ガウス過程回帰を用いたCOガス単成分回帰モデル

今回のデータセットに対して、ガウス過程回帰の適用を検討するに当たっては、開発中のガスセンサ全体に対して適用可能な汎用的なモデルを構築するのではなく、選択したセンサ個々に対する個別の回帰モデルを

構築することを目指した。それは、ニューラルネットワークでの回帰モデル構築の検討から、開発中のセンサにおいて、各センサ固有の構造などに由来すると考えられる特性のばらつきが見られるため、個別の回帰モデル構築が望ましいと判断したためである。そのため、データセットは学習用としたガス濃度における、選択したあるセンサの出力電流値で構成されるデータとなる。

最初に、学習データ全体に対するガウス過程回帰を適用した回帰モデル構築について述べる。カーネル関数には、

$$k(\boldsymbol{x}、\boldsymbol{x}') = \theta_1 \boldsymbol{x}^T \boldsymbol{x}' + \theta_2 \exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}'|^2}{\theta_3}\right)$$

を採用した。今回のデータセットでは、元々の測定データの電流に対する濃度の相関が非常に高く、その相関の高さを加味するための線形カーネル（第1項）に、非線形性を加えることを目指してRBFカーネル（第2項）を組み合わせた。GPy上では、複雑さを加味させるためにバイアスカーネルを加えた。

・GPyの式：GPy.kern.Linear(1) + GPy.kern.RBF(1) * GPy.kern.Bias(1)

このカーネル関数を用いたCOガス濃度のみを出力とした時の計算結果を図3に示す。ここで、点は学習に使用したデータ点、実線はフィッティングを行った結果の回帰モデル、薄く色が付いている範囲が各点での信頼区間である。

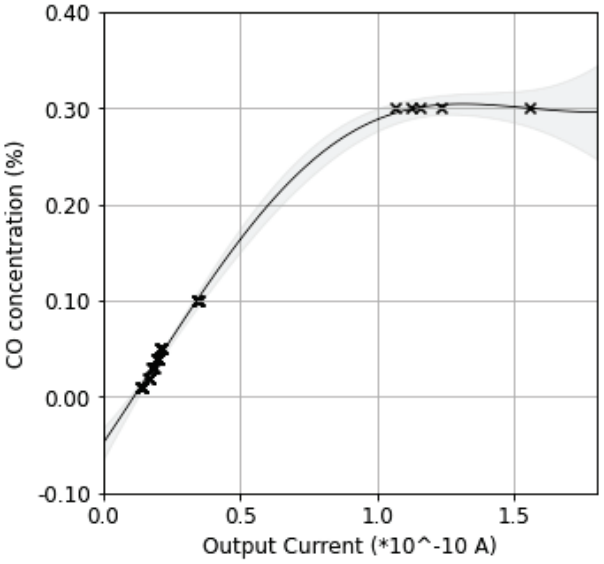

図 3　ガウス過程回帰による回帰モデル（全濃度）

今回の計算結果において、データセット中の高濃度（CO 0.30%）データのばらつきが低濃度（CO 0.10%以下）に比べて大きく、かつ、より高濃度（CO 0.30%以上）のデータが無いため、フラットな値に沿ってフィッティングしていることが分かる。これは、ガウス過程回帰は、データが無い領域では、フィッティングしようとする関数の分布が不確定となり、予測が不十分となるためである。

そこで、データセット中の最高濃度（CO 0.03%）のデータを除外し、データ間隔が狭く、データ数が多い低濃度領域（CO 0.10%以下）のデータで学習させた。ここで、カーネル関数には、

$$k(\boldsymbol{x}、\boldsymbol{x}') = \theta_1 \boldsymbol{x}^T \boldsymbol{x}' \times \theta_2 \exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}'|^2}{\theta_3}\right)$$

・GPyの式：GPy.kern.Linear(1) * GPy.kern.RBF(1) * GPy.kern.Bias(1)

を採用した。

低濃度領域での計算結果を、図4に示す。ここで、全濃度領域の結果と同様に、点は学習に使用したデータ点、実線はフィッティングを行った結果の回帰モデル、薄く色が付いている範囲が各点での信頼区間である。

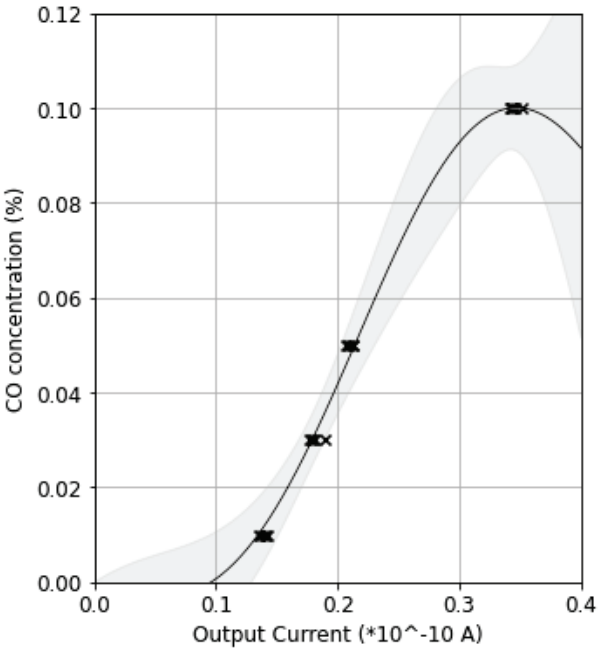

図 4　ガウス過程回帰による回帰モデル（低濃度）

データが存在する領域では、適切なフィッティングが行われている。一方で、より低濃度領域（出力電流値で約0.13×10^{-10} A以下）では、データが存在しない領域のため、フィッティングが不正確である。

フィッティングした曲線は、理論上は任意の入力に対して出力を得ることができるが、そもそもデータの存在しない領域での出力値には意味がない。そのため、データ間を補間するような回帰モデルとしての利用が適切であることがわかる。この低濃度領域での回帰モデルに対

して、COガス濃度0.04％と0.02％に対する予測結果を図5と表3に示す。

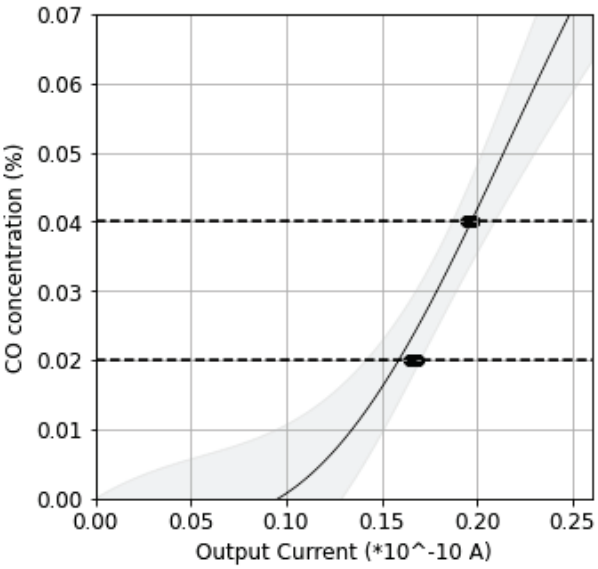

図 5 ガウス過程回帰による予測（低濃度）

表3 ガウス過程回帰による予測値

電流 （×10^{-10} A）	CO濃度 予測値	標準偏差	CO濃度 真値
0.164	0.023％	0.003％	0.02％
0.169	0.025％	0.003％	
0.166	0.024％	0.003％	
0.165	0.023％	0.003％	
0.168	0.024％	0.003％	
0.194	0.038％	0.003％	0.04％
0.197	0.040％	0.003％	
0.198	0.041％	0.003％	
0.198	0.041％	0.003％	
0.198	0.040％	0.003％	

各濃度に対して、データ数が5点あるため、各電流値に対するCOガス濃度の予測値は、広がりを持っている。この比較結果から、予測値の推定においては、COガス濃度 0.04％の方が、目的値を含み、より適切に予測していると考えられる。これは、図5のグラフ範囲外に存在するデータ（COガス濃度 0.10％）の影響を受けたために、より適切に推測していると考えられる。一方、COガス濃度 0.02％では、より低濃度側（COガス濃度 0.01％以下）のデータが無い影響を受け、目的値からずれて予測されていると考えられる。以上の結果から、今回のケースにおいては、回帰モデルの有効範囲としては、およそ0.03％以上、0.05％以下のCOガス濃度の予測ができると考えられる。

この結果を踏まえて、今後、より広範囲のガス濃度に対応した回帰モデルへ拡張を試みる場合、今回の測定データ中で欠損している領域（例えば、COガス濃度 0.10％以上0.3％以下）での測定データが必要である。その時には、測定回数などを軽減するために、ベイズ最適化の適用なども期待できる。

3.2 ガウス過程回帰を用いた多成分回帰モデル

出力を一酸化炭素、酸素、ヘリウム、窒素の全成分を対象とした、低濃度領域での多成分ガウス過程回帰の計算結果を図6に示す。多成分のモデル構築には、GPyに実装されているアルゴリズム[11]を利用した。また、単成分の結果を踏まえて、低濃度領域と同じカーネル関数と同じものを利用した。

多成分でのガウス過程回帰においても、各成分のフィッティングが実施できることを確認できた。データの少ない領域の影響を受け、データの端の部分では、不正確なフィッティングとなっている。そのため、回帰モデルの活用時には、事前のデータ測定において、適切な測定範囲を設定したうえで、回帰モデルの適用範囲を定める必要がある。なお、今回のデータセットでは、既存の測定データがガス成分空間中での線上データであるため、その適否判断は限定的だと言える。それは、あるガス成分の濃度をある値に固定させて、他の成分の濃度を変化させて電流値を測定したデータではないためである。今後、あるガス成分の濃度を固定させた、ガス成分空間中での面や領域中でのデータが取得でき次第、検討が必要と考えている。

4 結言

開発中のガスセンサで測定した時系列データを用いた、ガス濃度予測用回帰モデル構築の検討を行った。ガウス過程回帰による回帰モデル構築を行った結果、ガウス過程の特徴である分散が計算できることによる予測値の確からしさの推定ができることが分かった。また、回帰モデルの適用時には、適切な適用範囲を定める必要があることがわかった。今後、ガス成分空間中での面上でのデータの測定などが進む中で、更なる検討を実施する予定である。

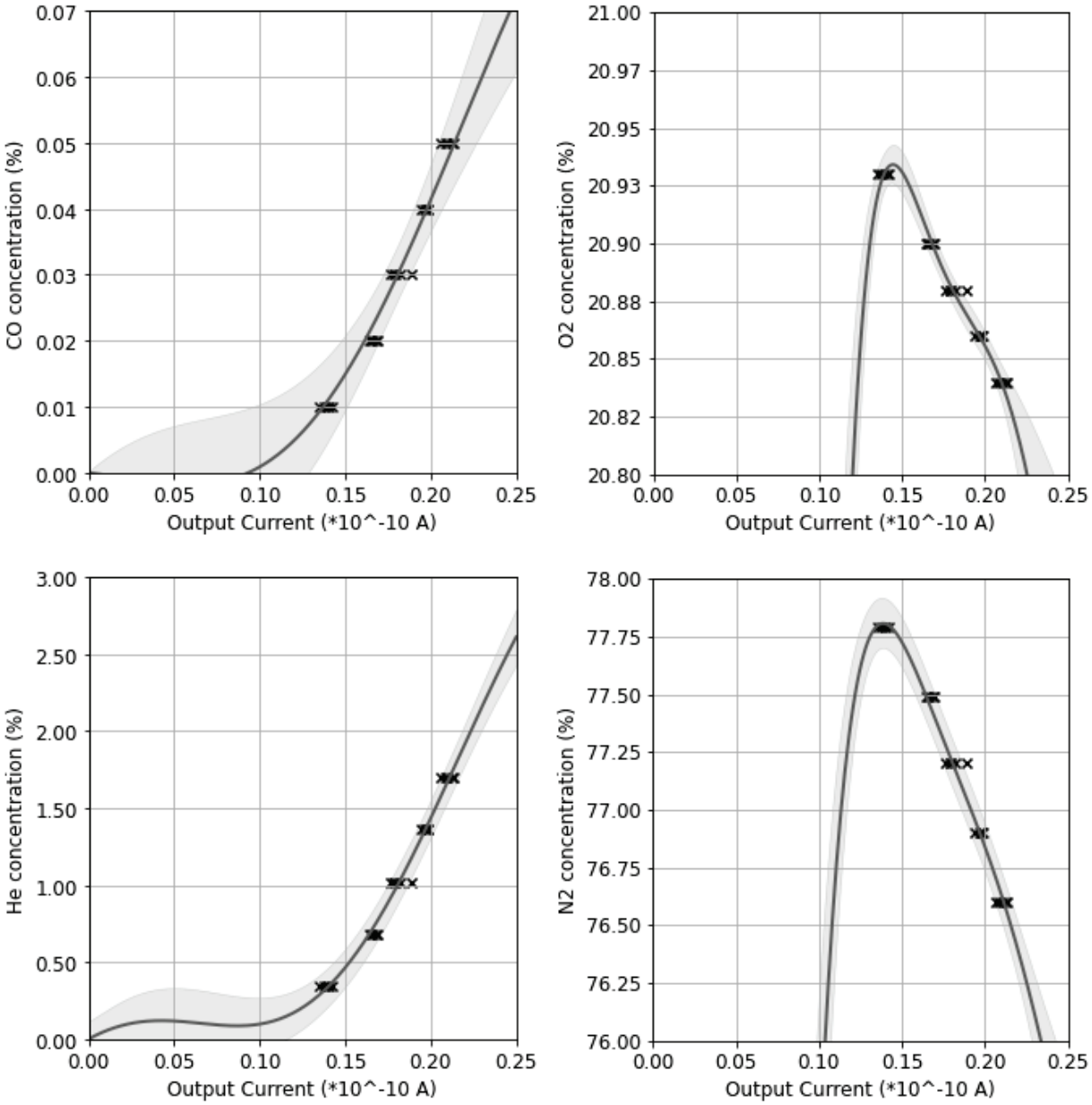

図 6 ガウス過程回帰による多成分回帰モデル（低濃度）

謝辞

本研究を進めるに当たり、有益な助言を頂いた東北大学 庭野道夫名誉教授、東北福祉大学 岩田一樹准教授に謝意を表します。

参考文献

1） OpenAI、 ChatGPT、 https://openai.com/blog/chatgpt （参照 2023-05-17）

2） 経済産業省、“中小企業向けAI導入ガイドブック”、2022、 https://www.meti.go.jp/press/2022/04/20220408001/20220408001.html （参照 2023-05-17）

3） 経済産業省、“DXセレクション（中堅・中小企業等のDX優良事例選定）”、 https://www.meti.go.jp/policy/it_policy/investment/dx-selection/dx-selection.html （参照 2023-05-17）

4) 阿部 宏之、岩田 一樹、馬 騰、但木 大介、平野 愛弓、木村 康男、庭野 道夫、“集積化ガスセンサへの機械学習の適用”、 センサ・マイクロマシンと応用システム」シンポジウム論文集 電気学会センサ・マイクロマシン部門［編］、 2020.

5) 岩沢正樹、 阿部 宏之、“機械学習を用いた時系列データ解析”、 宮城県産業技術総合センター研究報告、 No19、 2022

6) Carl Edward Rasmussen, Christopher K. I. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2006

7) 持橋 大地、大羽 成征、 ガウス過程と機械学習、講談社、 2019

8) J. Takeuchi and K. Yamanishi, “A Unifying Framework for Detecting Outliers and Change Points from Time Series”, IEEE Transaction on Knowledge and Data Engineering, 2006, 18(4), p.482-492

9) 山西 健司、 データマイニングによる異常検知、共立出版、 2009.

10) 山西 健司、 情報論的学習とデータマイニング(数理工学ライブラリー 3)、 朝倉書店、 2014

11) Zhenwen Dai, Mauricio A. Álvarez, Neil D. Lawrence, “Efficient Modeling of Latent Information in Supervised Learning using Gaussian Processes.”, NIPS, 2017.