経済時系列データにおける
外れ値検出指標のパフォーマンスに係る比較検討

尾﨑　雄太*
相田　政志†
田村　秀一‡

Comparative Analysis on Outlier Detection Indicators in Economic Time Series Data

OSAKI Yuta
AIDA Masashi
TAMURA Hidekazu

本稿は、統計調査実務への応用を念頭に、経済時系列データにおける外れ値を効率的に検出することを目的として、外れ値の存否の検出に用いる指標を検討する。シミュレーションと実データからデータセットを用意し、外れ値を検出する二項分類器に用いる検出指標のパフォーマンスを、ROC 曲線と PR 曲線を用いて評価した結果、自己相関を組み込んだモデルを活用した検出指標や、前月差から算出した検出指標が高いパフォーマンスを示した。また外れ値検出で用いるべきデータ期間の長さは検出指標により異なること、外れ値の乖離の大きさによって検出のパフォーマンスには差が見られることが示された。さらに検出指標別に、最も効率的に外れ値を検出できる閾値を得ることができた。本稿で得られた知見は、経済時系列データの調査実務の一環として実施されている外れ値検出に活用されることが期待される。

キーワード：経済時系列データ、外れ値検出、二項分類器、ROC 曲線、PR 曲線

This study, with an emphasis on application to statistical survey practice, aims to efficiently detect outliers in economic time series data by examining indicators used to predict the presence of outliers. Using datasets prepared from simulations and real-world data, we evaluated the performance of outlier detection indicators through ROC and PR curves. The results indicate that indicators based on models incorporating autocorrelation, as well as those derived from month-over-month differences, perform better than others. Furthermore, the optimal length of time series data required for effective outlier detection varies depending on the chosen indicator, and detection performance varies according to the size of deviation exhibited by outliers. Additionally, for each detection indicator, we identified threshold values that most efficiently detect outliers. The findings of this study are expected to contribute to improving outlier detection in the production process of economic time series data.

Keywords: Economic Time Series Data, Outlier Detection, Binary Classifier, ROC Curve, PR Curve

* 経済産業省大臣官房調査統計グループ総合調整室
† 経済産業省大臣官房調査統計グループ経済解析室
‡ 経済産業省大臣官房調査統計グループ鉱工業動態統計室

1　はじめに

1.1　経済時系列データにおける外れ値検出の必要性

経済時系列データにおいて最新の値に通常と異なる動きがみられるとき、その異常を適時適切に検出することは、統計調査実務における誤記入等の測定誤差の検知や、経済時系列データの動向の背景を把握する上で重要だ。

経済時系列データの作成においては、外れ値はしばしば調査対象の誤記入等の測定誤差により生じるが、そのような測定誤差は可能な限り迅速かつ的確に検出され、修正されることが望ましい。特にデータの真値性を客体に直接確認する場合は、調査対象に適時に接触することが求められ、限られた情報を基に迅速かつ効率的に外れ値を検出する必要がある。また経済時系列データの解釈において、経済活動における一時的なショックや景気局面等の実体経済の変化が経済時系列データにおける外れ値をもたらしていることが考えられる。このような外れ値をその統計の作成段階から検出することは、統計作成主体が経済時系列データ上の変動を適切に解釈し、基調判断等の合理的な意思決定を行う上でも有用である。

既往文献では、統計調査実務における外れ値の検出に関して、統計作成におけるデータ検証（data validation）の枠組みを定型化し、データ検証に用いるべき指標の特徴を論じるもの（Zio et al.（2018））や、横断企業調査の作成における編集や欠損値補定（imputation）に関して、推奨される誤差の検出やその処置の手法を含めて紹介するハンドブック（Orietta et al.（2007））がある。

統計研究彙報においても外れ値の検出に関した研究が掲載されており、統計実務で回答に誤りがないか確認する審査業務におけるレンジチェックに関して、外れ値の検出手法を比較するもの（野呂・和田（2015））や、多変量データにおける外れ値の検出手法について比較評価を行うもの（和田（2010））がある。また調査対象による記入漏れや記入誤りがあった際の対処の一つである欠損値補定について論じた研究として、諸外国での取組みを紹介するもの（小林（2009）、高橋（2012））や、補定手法を定量的に検討するもの（和田（2012）、高橋・伊藤（2013）、高橋・伊藤（2014）、和田・野呂（2019）等）がある。

このような研究の蓄積の一方で、特に月次等の経済時系列データに着目して、外れ値の検出に用いるべき指標を定量的に評価する試みは、これまでに十分なされているとは言えない。

1.2　本稿の目的と構成

このような背景から本稿は、統計調査実務への応用を念頭に、経済時系列データにおける外れ値を効率的に検出することを目的として、外れ値の存否の検出に用いる指標を検討する。本稿のリサーチ・クエスチョンは、（１）どのような検出指標を用いると最も効率的に経済時系列データの外れ値を検出できるのか、（２）検出指標を定義するに当たって、どの程度の期間だけ過去に遡ったデータを用いることが望ましいのか、（３）外れ値検出で検出するべき外れ値の乖離の大きさと、検出のパフォーマンスの間には、どのような関係が見られるのか、（４）各検出指標で最も効率的に外れ値の存否を判定できる閾値は何か、である。

本稿の構成は、第 2 節で検出指標を検討するための手法を説明する。続いて、第 3 節で候補となる検出指標を提示し、それらの特徴を整理する。第 4 節でシミュレーションを用いて、第 5 節で実データを用いて、検出指標別の外れ値検出のパフォーマンスを分析する。第 6 節で分析を通して得られた本稿のリサーチ・クエスチョンに対する答をまとめる。

なお本稿では、分析に当たって経済産業省生産動態統計調査の時系列データを用いた。当該統計は、1500 を超える数の広範な品目（末端品目数）について、生産量等の我が国事業所・企業の月次の生産活動に関する最も網羅的かつ詳細な時系列データを提供している。このことに鑑みれば、当該統計を用いた分析を行うことで、製造業が中心にはなるが、政府統計のみならず、業界統計等、他の生産活動に係る統計調査実務でも適用可能な汎用的な知見が得られることが期待される。

2　検討の手法

2.1　検討する外れ値検出の定義

時系列データに関する外れ値検出は、対象となるパラメーターに関して、過去の時系列データと直近の値しか観察できない状況下で、直近の値に関する外れ値の有無を検出することと言い換えることが出来る。数学的表現を用いて、本稿で検討する外れ値検出をより厳密に定義する。対象の時点（期末）をT、時点tにおけるパラメーターの値をY_t、時点Tにおいて過去の変動から推測される、異常な変動や測定誤差等を含まない「もっともらしい」パラメーターの値をY_T^*、期末のパラメーターの値を$Y_T = Y_T^* + error$とおく。このとき$error$はもっともらしい値からの乖離であり、この$error$が 0 であるか否を$Y_0, Y_1, \ldots, Y_T$から推測することが、外れ値検出の目的である（図 1）。

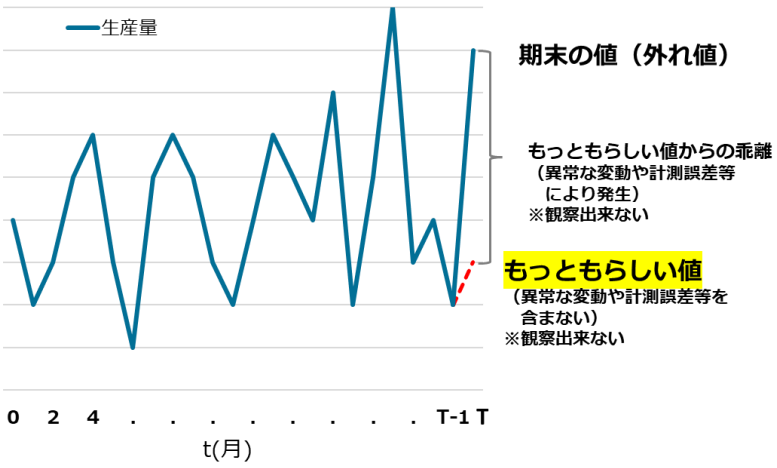

図 1　外れ値検出で使用するデータの模式図

外れ値はある確率分布に従って出現すると想定されるが、本稿では外れ値の大きさごとの結果を比較するため、一定の確率で一定の大きさの外れ値が生成される（正負で同じ確率とする）ものと仮定する。すなわち、大きさ$error\ size(>0)$のもっともらしい値からの乖離について、$0 \leq p \leq 1$に関して、

$$\begin{cases} \Pr(error = error\ size) = \Pr(error = -error\ size) = \dfrac{p}{2} \\ \Pr(error = 0) = 1 - p \end{cases} \tag{1}$$

とする。

ここで二項分類器を用いて、外れ値の存否を推測することを考える。本稿では、時系列データと直近の値の関数である特定の検出指標$Indicator(\geq 0)$が、事前に決めた閾値$threshold(\geq 0)$を超えた場合に、外れ値が存在すると推測する二項分類器を検討する。このとき、検出指標は

$$Indicator \coloneqq f(Y_0, Y_1, \dots, Y_{T-1}; Y_T) \tag{2}$$

と表され、二項分類器は外れ値の存否を、

$$Prediction = \begin{cases} 1\ if\ Indicator >\ threshold\ (i.e.\ Positive) \\ 0\ if\ Indicator \leq\ threshold\ (i.e.\ Negative) \end{cases} \tag{3}$$

から推測する。なお$Prediction = 1$は直近の値に外れ値が存在すると推測されることを表す。

2.2　検出指標の評価

実際の外れ値の存否と、二項分類器により推測される外れ値の存否の関係性は、混合行列により整理することができる（表 1）。外れ値がない時（$Y_T = Y_T^*$）に外れ値がないと二項分類器が推測することを真陰性（True Negative）と呼び、外れ値があると二項分類器が推測することを偽陽性（False Positive）と呼ぶ。逆に、外れ値がある時（$Y_T \neq Y_T^*$）に外れ値がないと二項分類器が推測することを偽陰性（False Negative）と呼び、外れ値があると二項分類器が推測することを偽陽性（True Positive）と呼ぶ。さらに、真陽性度ないし感度は$Sensitivity\ (or\ Recall) = \frac{TP}{TP+FN} (= 1 - \beta)$、真陰性度ないし特異度は$Specificity = \frac{TN}{TN+FP} (= 1 - \alpha)$と定義される。また適合率は$Precision = \frac{TP}{TP+FP}$、正解率は$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$と定義される。二項分類器の評価には、感度と適合率の調和平均である$F1\ score = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$が使われることもある。

表 1　外れ値検出に係る混合行列

		推測（外れ値検出）	
		外れ値なし（$Prediction = 0$）	外れ値あり（$Prediction = 1$）
実際	外れ値なし（$Y_T = Y_T^*$）	真陰性 TN （1-α）	偽陽性（第一種過誤）FP （α）
	外れ値あり（$Y_T \neq Y_T^*$）	偽陰性（第二種過誤）FN （β）	真陽性 TP （1-β）

二項分類器の使用に際しては、感度と特異度の両方が高くなるような閾値を設定することが望ましいが、一般にこれらの 2 つの値の間にはトレード・オフの関係があり、それを図示したものが受信者操作特性（Receiver Operating Characteristic, ROC）曲線である[4]。ROC 曲線は、ある二項分類器に関して、閾値を少しずつずらしながら(x, y) = (1 − Specificity, Sensitivity)をプロットすることで得られる曲線である。ROC 曲線がグラフ左上の角に近い位置にあるほど感度と特異度が優れていると考えることができる。ROC 曲線の曲線下面積（Area Under the Curve）（ROC-AUC）は、二項分類器のパフォーマンスの評価に用いられる。ROC-AUC は 0 以上 1 以下の値を取り、値が大きい二項分類器ほどパフォーマンスが高く、ROC-AUC= 1/2となる二項分類器のパフォーマンスはコイントス（当てずっぽう）と変わらない。

一方で陽性と陰性の発生確率に非均衡がある時には、ROC 曲線による分析は誤った解釈を招く可能性があり、陽性に占める真陽性の割合である適合率に着目した PR（Precision Recall）曲線とその線下面積（PR- AUC）を用いることが推奨されている（Saito and Rehmsmeier（2015））。PR 曲線は、ある二項分類器に関して、閾値を少しずつずらしながら(x, y) = (Sensitivity (or Recall) , Precision)をプロットすることで得られる曲線であり、PR-AUC は 0 以上 1 以下の値を取り、値が大きい二項分類器ほどパフォーマンスが高い。

ROC 曲線や PR 曲線を用いた先行研究は、特に医療や機械学習の分野で数多いが、社会科学領域における例を挙げると、ROC 曲線と ROC-AUC を用いて、経済活動の拡大局面と後退局面を評価するに適した指標を特定することを試みた研究（Berge and Jordá（2011））や、ローン申請者を「良い支払者」と「悪い支払者」に分類する際に効率的な手法を ROC-AUC を用いて検討した研究（Brown and Mues（2012））がある。

2.3　シミュレーションと実データを用いた検討

検出指標を評価するに当たって、直近の値に外れ値を含まないものとして扱えるデータセットを用意する必要がある。

そのため、第一に、シミュレーションを用いてデータセットを生成する。シミュレーションから生成されたデータに外れ値を導入し、それに検出指標を適用することで、検出指標のパフォーマンスを評価することができる。シミュレーションであれば、外れ値が存在しないクリーンなデータセットを用意できることに加え、任意のサイズのデータを生成することが可能だ。一方で、シミュレーションで用いるモデルに結果が依存してしまうため、実際の経済時系列データの変動の特徴を正確に再現できるようなシミュレーション手法を用いなければ、実用的な結果が得られないかも知れない。

このような欠点を補う観点で、第二に、過去の実データからなるデータセットを用いる。これにより、より現実的な設定で、外れ値検出の手法を比較検討できる。一方で、実データには外れ値が含まれる可能性があり、また利用可能なデータサイズには限りがある。

[4] ROC 曲線の概要や使用における留意点については、Fawcett（2007）等が詳しい。

2.4　パラメーターの設定

分析に当たって、(1),(2)式に登場する定数に代入する値を決める必要がある。まずもっともらしい値からの乖離の大きさError Sizeについては、通常では発生頻度が低い値を想定して、データ生成過程における撹乱項の標準偏差σに対して、$\text{Error Size} = 1.5\sigma, 2\sigma, 3\sigma$を設定した[5]。

検出指標の作成に当たって参照する時系列データの期間の長さTについては、過去 2 年または 5 年の時系列データを参照することを念頭に、$T = 26, 62$を設定した。

外れ値の発生確率pについては、本稿では$p = 0.5, 0.1, 0.05$を設定した[6]。

3　検出指標の候補

3.1　検出指標に求められる条件

たとえパフォーマンスが高くても、複雑で高度な検出指標が実用的であるとは限らない。検出指標に求められる条件として、検出のパフォーマンスが高いことに加えて、非専門家を含む指標の使用者が検出指標の特性を十分に理解できることが求められる。また、経済時系列データの特性を踏まえた検出指標であることが挙げられ、一般に経済時系列データでは直近の数値との相関関係や季節性がみられるため、それらを考慮できる検出指標の有用性が高いと思われる。

3.2　検出指標の改善案

先述の検出指標に求められる条件を踏まえて、本稿では 9 つの検出指標を候補に検討を行う。これらの検出指標はその特徴により大きく 4 つのグループに分けられる（表 2)。

まず、過去データの最大値最小値との比較による指標（MinMax)、過去データの四分位と四分位範囲による指標（Qntl)、そして標準偏差と平均値を用いて値を標準化した指標（SD）は、過去のデータをプールしてクロスセクションデータとして扱うものであり、時系列データの特性は考慮されていない。しかし、計算手法は比較的簡便なものである。

前年からの差分を標準化した指標（ChngPrvY）と前月からの差分を標準化した指標（ChngPrvM）は、過去データの 1 時点との比較を行うことで、季節性や、前期と今期の値の相関関係に対処するもので、計算手法は比較的簡便である[7]。

直近 12 か月分のデータを用いた二次の回帰曲線を延長し予測される値から観測値の差分を予測値の信頼区間の半分で割った指標（Reg12)、利用できる全期間分のデータ（長さ

---

[5] 実データを使用した分析においては、実際のデータ生成過程を知ることは出来ないため、実データからSARIMA モデルを用いて推定された撹乱項の標準偏差$\hat{\sigma}$を用いた。

[6] 外れ値の発生確率に関連して、先行事例においてアンケート調査等における記入誤差の発生確率には相当の幅があると明らかになっている。Schifeling et al.(2016)は、調査における誤答率を推定するモデルを構築し、最ももっともらしいモデルを使うと、米国での学歴に関する回答の誤答率が全体で 22%に上ると推定している。Celhay et al.(2024)は、ニューヨーク州のフードスタンプと公的扶助の受給有無に関する回答を、実際の行政記録と接合し、フードスタンプの受給者については 18~37%、公的扶助の受給者については 46~59%の誤答があった一方で、非受給者の誤答率はフードスタンプに関する設問で 1.1~1.3%、公的扶助に関する設問で 0.3~1.3%であったと報告している。

[7] 前年や前月の値が 0 の時に、前年同月比や前月比は定義することが出来なくなるため、検出指標の候補には含めなかった。$Y_{T-12} \neq 0$のとき、前年同期比 $= {Y_T - Y_{T-12}}/{Y_{T-12}} = {Y_T - Y_{T-12}}/{\sqrt{Var(Y_t - Y_{t-12})}} \cdot {\sqrt{Var(Y_t - Y_{t-12})}}/{Y_{T-12}}$ = 前年同月差（$ChngPrevM$）$\cdot {\sqrt{Var(Y_t - Y_{t-12})}}/{Y_{T-12}}$の関係にある。

T）を用いた二次の回帰曲線を延長し予測される値を予測値の信頼区間の半分で割った指標（Reg）は、過去の推移を二次曲線で近似するものである[8]。時系列のトレンドを捉えることができるし、二次項を入れることで時系列データの周期性にも一定程度対応できると思われる。また散布図上に図示すれば視覚的にも分かり易い。

最後に、過去データからARIMA（自己回帰和分移動平均）モデルを推定し、予測される今期の値を観測値から引き、予測値の信頼区間の半分で割った指標（ARIMA）、過去データから季節 ARIMA モデルを推定し、予測される今期の値を観測値から引き、予測値の信頼区間の半分で割った指標（SARIMA）は、時系列データの自己相関を組み込んだモデルを活用したものである[9]。ARIMA、SARIMA モデルの推定に当たっては、最も当てはまりが良い次数(p,d,q)または$(p,d,q)(P,D,Q)_{12}$を選択して用いる[10][11]。ARIMA と SARIMA は、伝統的に時系列データによく用いられてきたモデルであり、季節性や直近の期との相関関係を良く考慮できる指標だと考えられる。一方で、その特性を理解するには専門的な知識が求められる。

[8] 区間推定において、95%信頼区間を構成した。
[9] 区間推定において、95%信頼区間を構成した。
[10] ARIMA、SARIMA モデルの推定に当たっては、R の forecast パッケージの関数 auto.arima()を用いて、ドリフト項を許容している。モデルの選択においては、AICC が最も小さくなるものを最も当てはまりのいいモデルとして選択している。
[11] SARIMA モデルの中には、$(P,D,Q)=(0,0,0)$となるものもある。すなわち、SARIMA の算出においても、ARIMA モデルが最良であれば、ARIMA モデルを採用している。

表 2　検出指標の候補

検出指標名	定義	特徴	理解しやすさ	時系列データの特徴を考慮
最大値最小値（MinMax）	$MinMax \coloneqq \begin{cases} {Y_T - \bar{Y}_t}/{\max(Y_t) - \bar{Y}_t} & (Y_T \geq \bar{Y}_t) \\ {Y_T - \bar{Y}_t}/{\min(Y_t) - \bar{Y}_t} & (Y_T < \bar{Y}_t) \end{cases}, t = 0, \ \ldots, T-1$	時系列データをクロスセクションデータとして扱う	○	×
四分位（Qntl）	$Qntl \coloneqq \begin{cases} {Y_T - median(Y_T)}/{IQR} & (Y_T \geq median(Y_T)) \\ {Y_T - median(Y_T)}/{IQR} & (Y_T < median(Y_T)) \end{cases}, t = 0, \ \ldots, T-1$			
標準偏差（SD）	$SD \coloneqq \left\lvert {Y_T - \bar{Y}_t}/{\sqrt{Var(Y_t)}} \right\rvert, t = 0, \ \ldots, T-1$			
前年からの差分（ChngPrvY）	$ChngPrev \coloneqq \left\lvert {Y_T - Y_{T-12}}/{\sqrt{Var(Y_t - Y_{t-12})}} \right\rvert, t = 12, \ \ldots, T-1$	過去の 1 時点と比較する	○	△
前月からの差分（ChngPrvM）	$ChngPrevM \coloneqq \left\lvert {Y_T - Y_{T-1}}/{\sqrt{Var(Y_t - Y_{t-1})}} \right\rvert, t = 1, \ \ldots, T-1$			
12 か月回帰（Reg12）	$Reg12 \coloneqq \left\lvert {Y_T - \widehat{Y_T}}/{\frac{CI_{\widehat{Y_t}}(T)}{2}} \right\rvert, \widehat{Y_t} = \hat{a} + \hat{b}t + \hat{c}t^2$ $(\hat{a}, \hat{b}, \hat{c}$は$t = T-13, \ \ldots, T-1$ のデータを元にOLS推定$)$	過去の推移を二次曲線で近似する	○	△
全期間回帰（Reg）	$Reg \coloneqq \left\lvert {Y_T - \widehat{Y_T}}/{\frac{CI_{\widehat{Y_t}}(T)}{2}} \right\rvert, \widehat{Y_t} = \hat{a} + \hat{b}t + \hat{c}t^2$ $(\hat{a}, \hat{b}, \hat{c}$は$t = 0 \ \ldots, T-1$ のデータを元にOLS推定$)$			
ARIMA モデル（ARIMA）	$ARIMA \coloneqq \left\lvert {Y_T - \widehat{Y_T}}/{\frac{CI_{\widehat{Y_t}}(T)}{2}} \right\rvert, \widehat{Y_t} = ARIMA(p,d,q)$ (モデルは$t = 0, \ \ldots, T-1$ のデータを元に推定)	時系列データの自己相関を組み込んだモデルを活用	△	○
SARIMA モデル（SARIMA）	$SARIMA \coloneqq \left\lvert {Y_T - \widehat{Y_T}}/{\frac{CI_{\widehat{Y_t}}(T)}{2}} \right\rvert, \widehat{Y_t} = ARIMA(p,d,q)(P,D,Q)_{12}$ (モデルは$t = 0, \ \ldots, T-1$ のデータを元に推定)			

4　シミュレーションによる分析

4.1　分析の流れ

以下の手順でシミュレーションによるデータを生成した。

経済産業省生産動態統計調査の公表されている調査結果のうち、2017 年 11 月～2024 年 12 月の品目別生産数量の時系列データを用いて、SARIMA モデルを推定した。その推定結果をランダムに抽出し、SARIMA モデルに基づく長さ T のシミュレーションデータを 10,000 個生成した[12]。さらにその期末の値について、一定の割合 p で外れ値の乖離の大きさに見立てた大きさ Error Size の値を加算した。

生成されたデータから前節で提示した検出指標をそれぞれ算出し、ROC 曲線と PR 曲線を描画するとともに、それぞれの曲線下面積（ROC-AUC と PR-AUC）を算出した。

[12] シミュレーションによるデータについては、初期値を 0 としたデータを 100+T 期分生成したうえで、101 期から 100+T 期までの値を使用した。

4.2　分析結果

ROC 曲線を用いた分析では、例えば T=62、外れ値の乖離の大きさを 2 に設定した図 2 中央を見ると、SARIMA を検出指標とした際の ROC-AUC が 0.814 と最も大きく、続いて ARIMA が 0.762、ChngPrvM が 0.725、Reg の ROC-AUC が 0.702 だった。

T=26 に設定した図 3 でも定性的には同様の結果が得られたが、SARIMA は外れ値検出に使う過去のデータのサイズ T が大きいほどパフォーマンスが高くなった一方で、MinMax、Qntl、SD については、T が小さいほどパフォーマンスが高くなった。例えば、外れ値の乖離の大きさを 2 とした場合、SARIMA の ROC-AUC は T=62 のとき 0.814 だったが、T=26 のとき 0.751 と小さくなっている。また MinMax の ROC-AUC は T=62 のとき 0.632 だったが、T=26 のとき 0.671 と大きくなっている。また T=26 の時の SARIMA と ARIMA の ROC-AUC の差は、T=62 の時より小さかった。

外れ値の乖離の大きさを変えてもパフォーマンスの順序は変わらなかったが、外れ値が小さいほど、同じ指標であっても ROC-AUC が小さくなる傾向が見られた。T=62 に設定した SARIMA について、Error Size=3 の時の ROC-AUC は 0.943 であった（図 2 左）のに対して、Error Size=1.5 の時は 0.722 と小さくなった（図 2 右）。

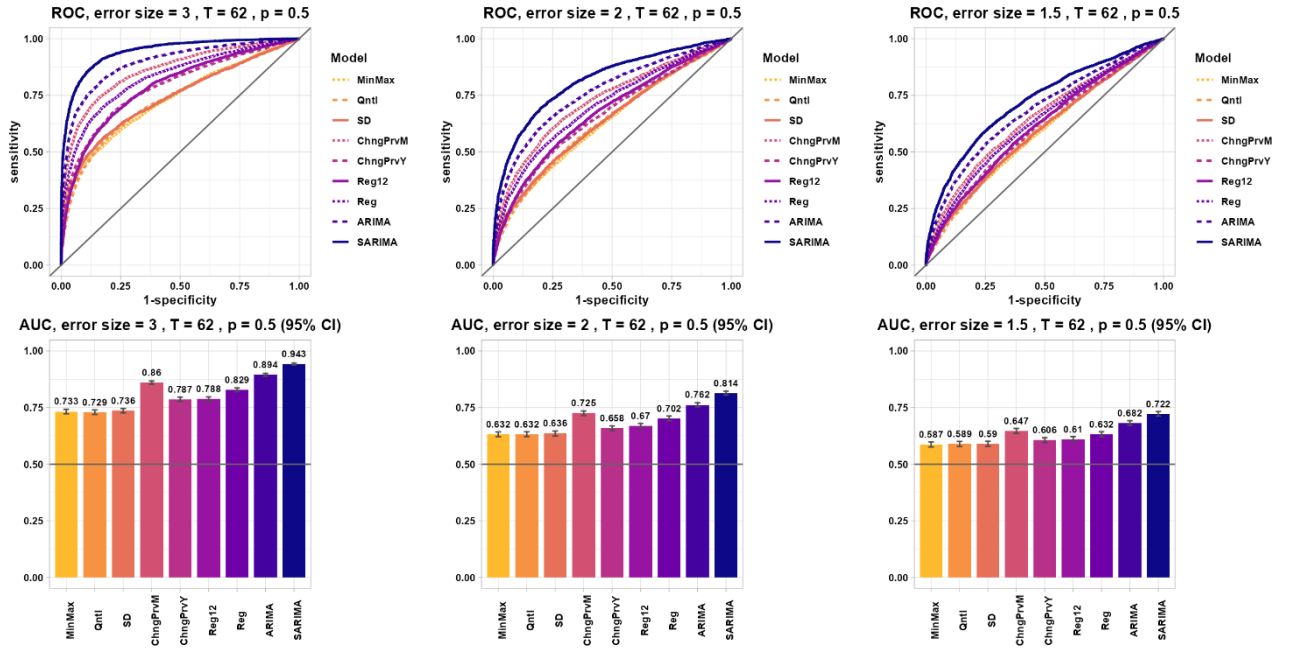

図 2　ROC 曲線と ROC-AUC（シミュレーション、T=62, p=0.5）

Note: 検出指標に関して、ROC 曲線（上段）とその線下面積（ROC-AUC）（下段）をまとめた。パラメーターは、検出指標の作成に当たって参照する時系列データの期間の長さ T=62、外れ値の発生確率 p = 0.5 に固定した。もっともらしい値からの乖離の大きさ Error Size は、左より 3,2,1.5 を設定した。ROC-AUC のエラーバーは、ブートストラップ法を用いて算出した 95%信頼区間。

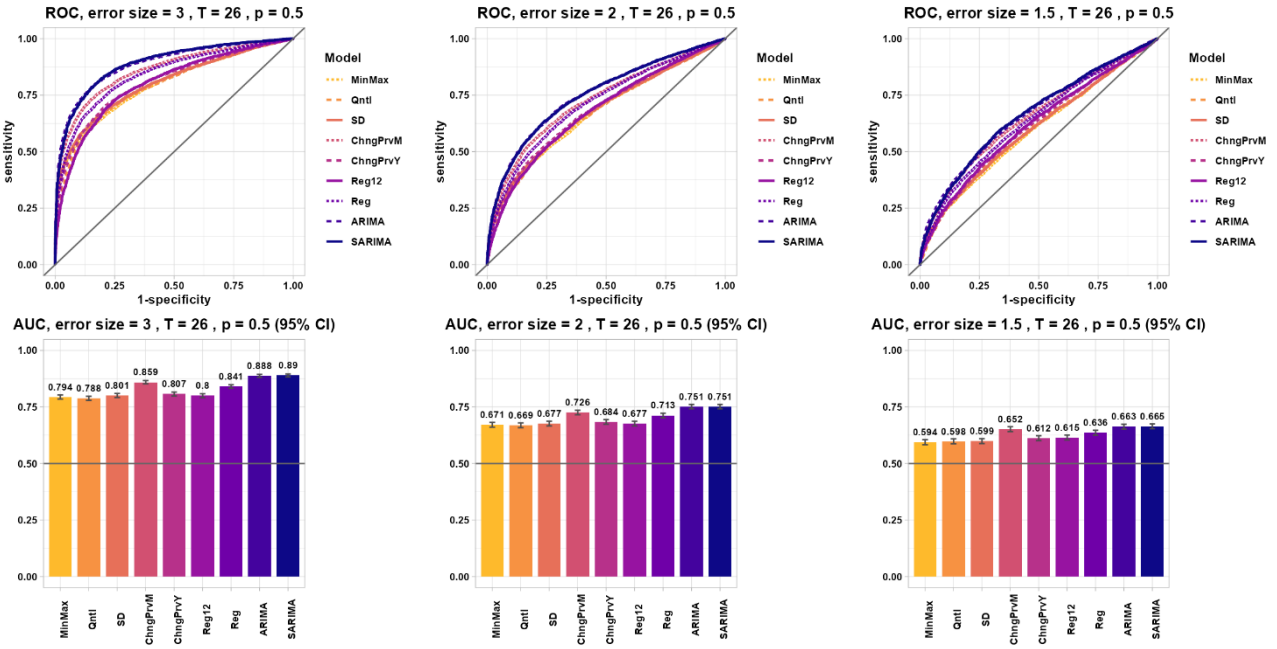

図 3　ROC 曲線と ROC-AUC（シミュレーション、T=26, p=0.5）

Note: 検出指標に関して、ROC 曲線（上段）とその線下面積（ROC-AUC）（下段）をまとめた。パラメーターは、検出指標の作成に当たって参照する時系列データの期間の長さ T=26、外れ値の発生確率 p = 0.5 に固定した。もっともらしい値からの乖離の大きさ Error Size は、左より 3,2,1.5 を設定した。ROC-AUC のエラーバーは、ブートストラップ法を用いて算出した 95%信頼区間。

続いて PR 曲線について、T=62、外れ値の乖離の大きさを 2 に設定した図 4 では、外れ値の発生率 p を 0.1 にした中図を見ると、SARIMA を検出指標とした際の PR-AUC が 0.458 と最も大きく、続いて ARIMA が 0.358、ChngPrvM が 0.307 だった。P の大きさを変えてもパフォーマンスの順序は変わらなかったが、p が小さいほど、同じ指標であっても PR-AUC が小さくなる傾向が見られた。また T=26 にした図 5 では、SARIMA と ARIMA の PR-AUC の差は T=62 の時より小さかった。

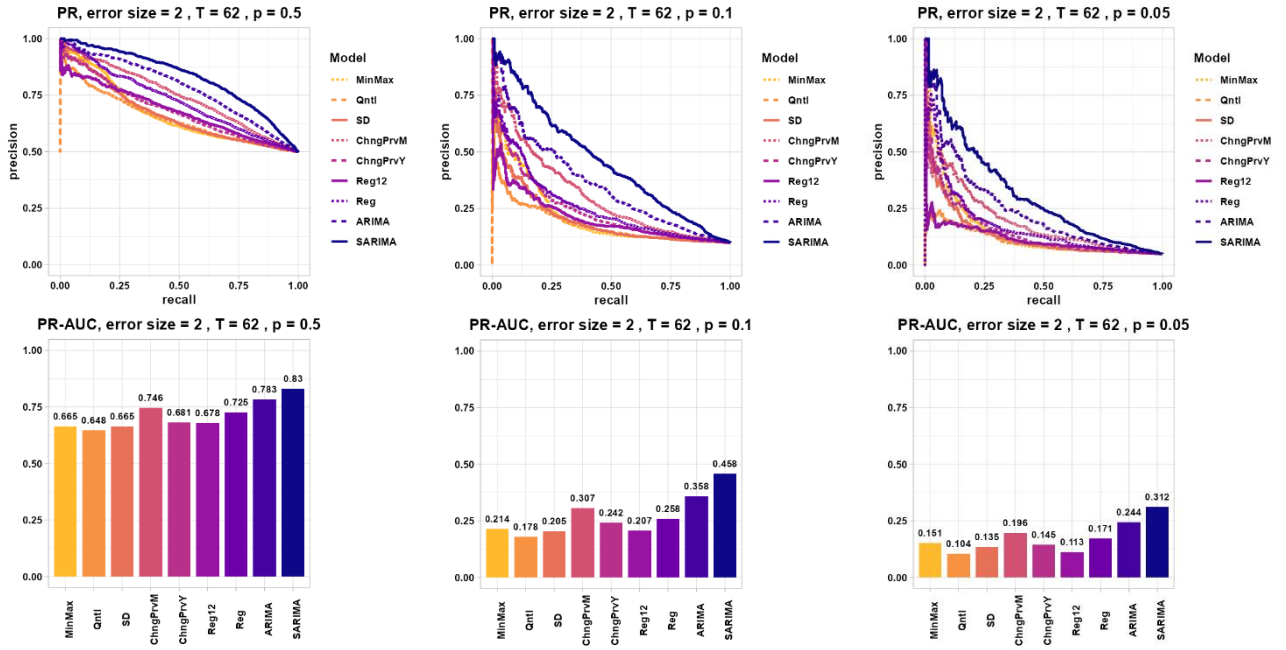

図 4　PR 曲線と PR-AUC（シミュレーション、T=62, Error Size=2）

Note: 検出指標に関して、PR 曲線（上段）とその線下面積（PR-AUC）（下段）をまとめた。パラメーターは、検出指標の作成に当たって参照する時系列データの期間の長さ T=62、もっともらしい値からの乖離の大きさ Error Size=2 に固定した。外れ値の発生確率pについては、左より 0.5,0.1,0.05 を設定した。

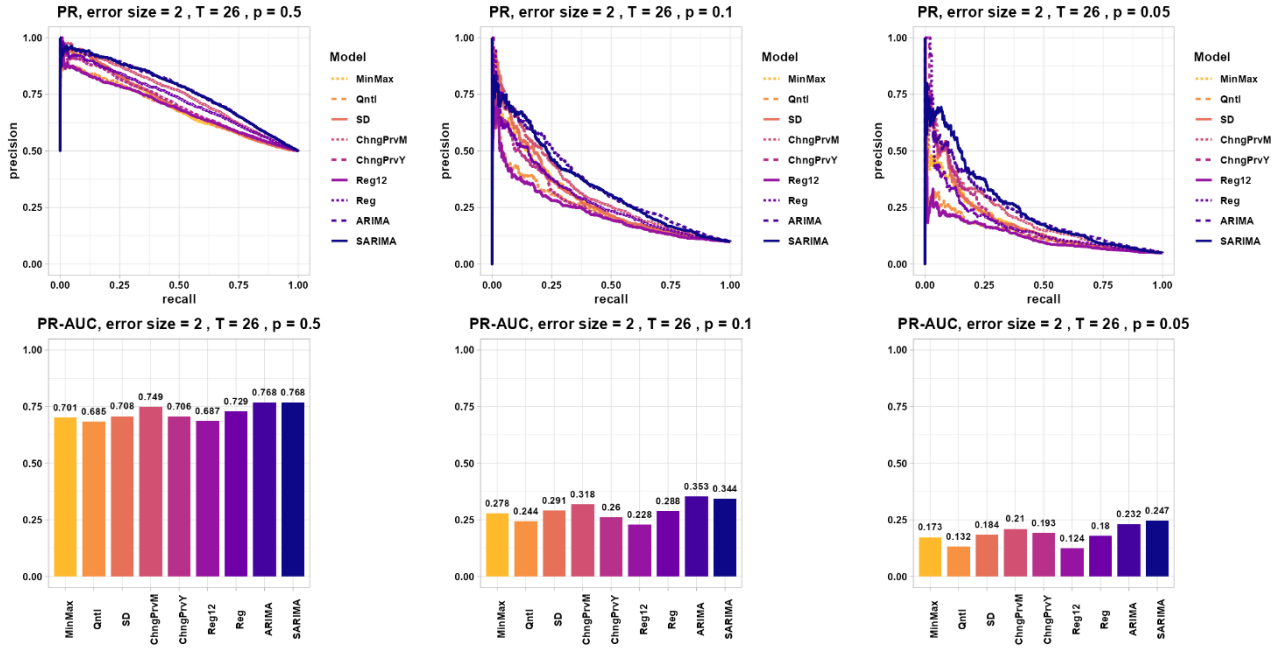

図 5　PR 曲線と PR-AUC（シミュレーション、T=26, Error Size=2）

Note: 検出指標に関して、PR 曲線（上段）とその線下面積（PR-AUC）（下段）をまとめた。パラメーターは、検出指標の作成に当たって参照する時系列データの期間の長さ T=26、もっともらしい値からの乖離の大きさ Error Size=2 に固定した。外れ値の発生確率pについては、左より 0.5,0.1,0.05 を設定した。

最適な閾値を設定するために、T=62, Error Size=2 の時に、F1 スコアを最大化する閾値を確認した（表 3）。Error Size=2, T=62, p=0.1 の時には、SARIMA では 0.984 を閾値に設定した場合の F1 値が最も大きく 0.463 で、感度が 0.499、特異度が 0.927、適合率が 0.432、正解率が 0.884 であった。またそれぞれの検出指標で F1 値を最大化する閾値を取った時に、SARIMA の F1 値、適合度、正解率はいずれも、他の検出指標より大きかった。

表 3　各検出指標のパフォーマンス（シミュレーション、T=62, Error Size=2）

						Sensitivity/			
				Threshold	F1 Score	Recall	Specificity	Precision	Accuracy
p	Indicator	ROC-AUC	PR-AUC	(under the maximum F1 score)					
0.5	MinMax	0.632	0.665	-	0.667	1	0	0.5	0.5
	Qntl	0.632	0.648	0.004	0.667	0.999	0.005	0.501	0.502
	SD	0.636	0.665	0.001	0.667	1	0	0.5	0.5
	ChngPrvM	0.725	0.746	0.748	0.685	0.764	0.535	0.621	0.649
	ChngPrvY	0.658	0.681	0	0.667	1	0	0.5	0.5
	Reg12	0.67	0.678	0.116	0.668	0.934	0.138	0.52	0.536
	Reg	0.702	0.725	0.631	0.675	0.857	0.32	0.557	0.588
	ARIMA	0.762	0.783	0.393	0.709	0.807	0.53	0.632	0.669
	SARIMA	0.814	0.83	0.486	0.75	0.815	0.64	0.693	0.728
0.1	MinMax	0.614	0.214	1.059	0.257	0.236	0.934	0.283	0.864
	Qntl	0.616	0.178	1.477	0.241	0.29	0.876	0.207	0.818
	SD	0.618	0.205	1.977	0.248	0.302	0.874	0.211	0.817
	ChngPrvM	0.712	0.307	1.841	0.353	0.363	0.923	0.344	0.867
	ChngPrvY	0.664	0.242	1.816	0.289	0.34	0.888	0.251	0.833
	Reg12	0.658	0.207	1.18	0.263	0.303	0.889	0.232	0.83
	Reg	0.687	0.258	2.375	0.302	0.389	0.869	0.247	0.821
	ARIMA	0.761	0.358	0.923	0.398	0.468	0.902	0.346	0.858
	SARIMA	0.816	0.458	0.984	0.463	0.499	0.927	0.432	0.884
0.05	MinMax	0.638	0.151	1.088	0.206	0.226	0.949	0.189	0.913
	Qntl	0.638	0.104	1.709	0.185	0.246	0.925	0.148	0.891
	SD	0.641	0.135	2.314	0.192	0.23	0.939	0.165	0.903
	ChngPrvM	0.731	0.196	2.129	0.261	0.27	0.958	0.253	0.924
	ChngPrvY	0.659	0.145	2.252	0.218	0.23	0.954	0.207	0.918
	Reg12	0.675	0.113	1.299	0.198	0.288	0.915	0.151	0.883
	Reg	0.705	0.171	2.889	0.223	0.284	0.933	0.183	0.901
	ARIMA	0.76	0.244	1.185	0.3	0.294	0.965	0.307	0.932
	SARIMA	0.807	0.312	1.208	0.355	0.334	0.971	0.379	0.939

Note: 検出指標のパフォーマンスをまとめた。パラメーターは、検出指標の作成に当たって参照する時系列データの期間の長さ T=62、もっともらしい値からの乖離の大きさ Error Size=2 に固定した。外れ値の発生確率pは、0.5, 0.1, 0.05 を設定した。Threshold より右の列については、各指標の F1 score を最大化する値に閾値を固定したときの値を報告している。

5　実データによる分析

5.1　分析の流れ

シミュレーションによる分析から検出指標毎のパフォーマンスを比較することが出来たが、シミュレーションで高いパフォーマンスを見せた検出指標について、検出指標として適切であるがゆえに ROC-AUC や PR-AUC が大きくなったのか、検出指標の算出に用いた計算式とシミュレーションのモデルが似ているため ROC-AUC や PR-AUC が大きくなったのか、明らかではない。特に SARIMA と ARIMA については、単にシミュレーションデータの生成過程と検出指標の算出法が類似しているためパフォーマンスが高かった可能性があるが、実際のデータが SARIMA によるシミュレーションと同じような過程で生成されているという確証はない。実際、現実の鉱工業の生産量は一定の上限・下限の範囲内で変動すると思われる[13]が、シミュレーションデータではそのような範囲は設定していない。そこで、実際のデータで検出指標を活用したときにどのような結果が得られるか確認する。

経済産業省生産動態統計調査における 2017 年 11 月～2024 年 12 月の時系列データのうち T か月分をランダムに抽出し、T か月目の値に対してランダムに外れ値を加える。また外れ値を加えた後の T か月目の値が負になる場合は、0 で置換する。その後、シミュレーションによる分析と同様の過程で検出指標を算出し、パフォーマンスを比較した。

5.2　分析結果

実データによる検証の結果、ROC 曲線に着目すると、大きい順に SARIMA、ARIMA、ChgPrevM の ROC-AUC が他の検出指標より大きく、T=62, Error Size=2 の場合（図 6 中）は外れ値の乖離の大きさに関わらずパフォーマンスが高い検出指標の大小関係はシミュレーションによる分析結果と概ね同じだった。しかし、T=26, Error Size=2 の場合（図 7 中）は、SARIMA、ARIMA、ChgPrevM の ROC-AUC は小さくなり、互いの差も縮小している。また外れ値の乖離が小さいほど、同じ指標であっても ROC-AUC が小さくなる傾向が見られた。

[13] 工場の生産活動を考えたときに、需要が大幅に増加したとしても工場の生産能力を超えた生産を行うことは出来ない。逆に需要が激減したとしても生産量を 0 未満にすることは出来ないし、工場の設備規模や労働者数が非弾性的で固定費用が大きい場合は、赤字になったとしても価格が操業停止点を下回らない限り生産し続けることが考えられるから、0 より大きい生産下限を想定することもできる。

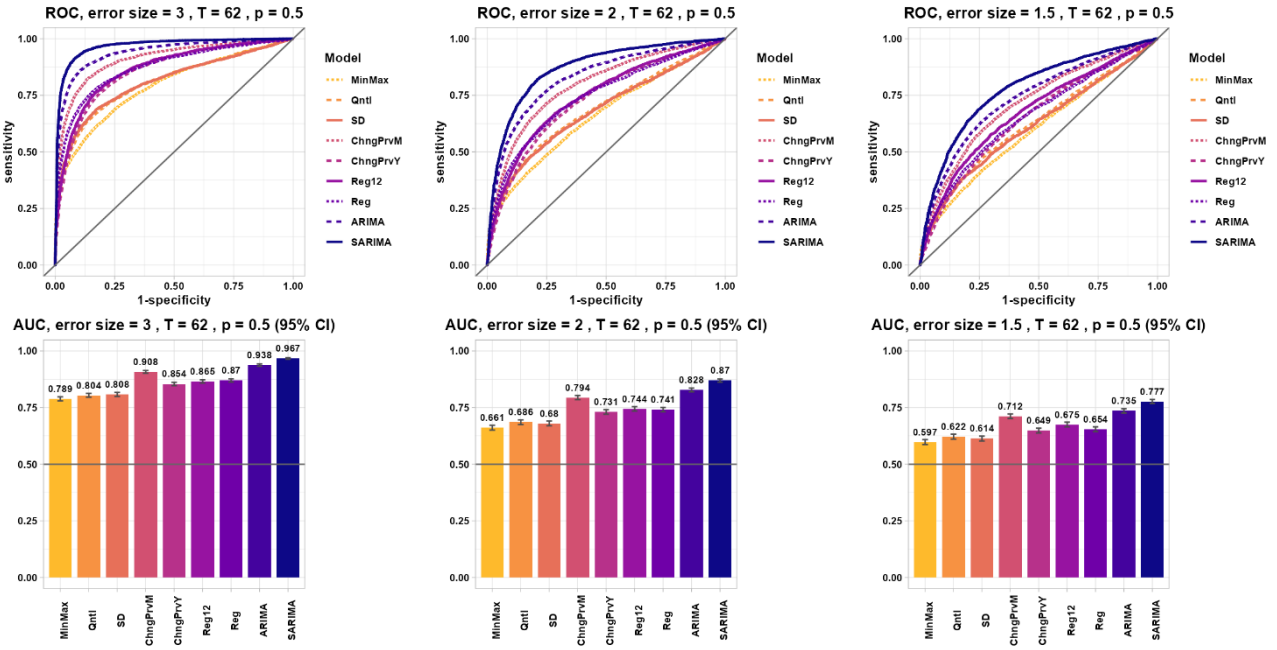

図 6　ROC 曲線と ROC-AUC（実データ、T=62, p=0.5）

Note: 検出指標に関して、ROC 曲線（上段）とその線下面積（ROC-AUC）（下段）をまとめた。パラメーターは、検出指標の作成に当たって参照する時系列データの期間の長さ T=62、外れ値の発生確率p = 0.5 に固定した。もっともらしい値からの乖離の大きさ Error Size は、左より 3,2,1.5 を設定した。ROC-AUC のエラーバーは、ブートストラップ法を用いて算出した 95%信頼区間。

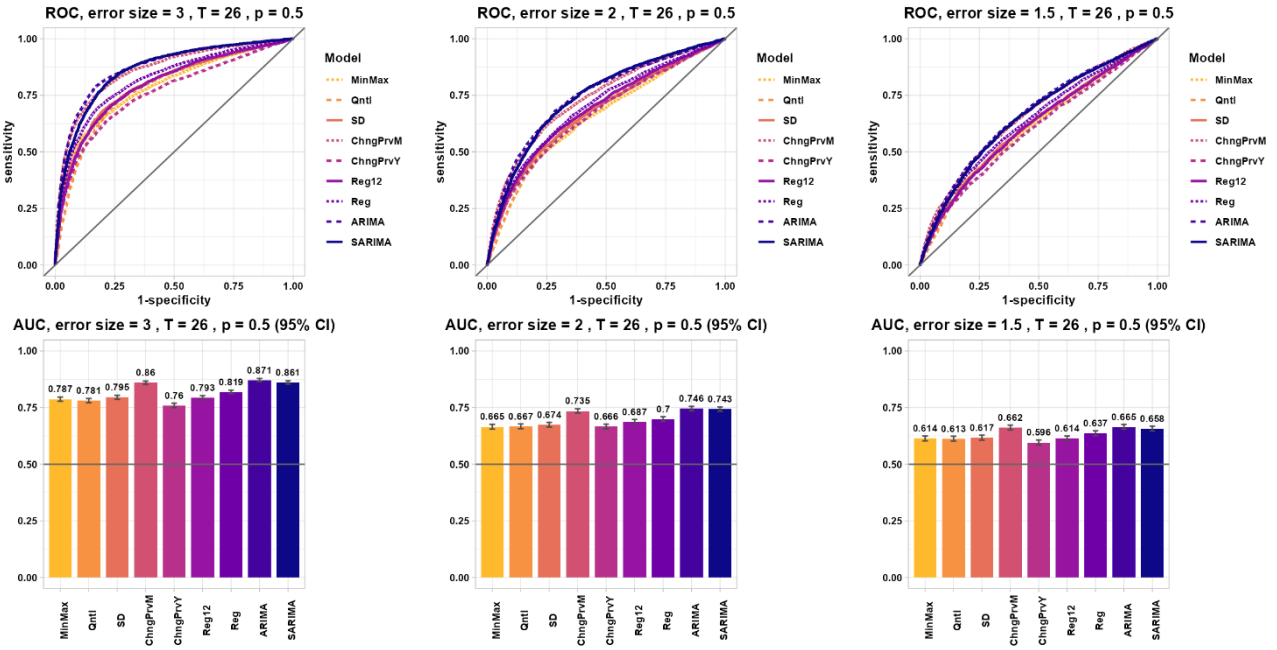

図 7　ROC 曲線と ROC-AUC（実データ、T=26, p=0.5）

Note: 検出指標に関して、ROC 曲線（上段）とその線下面積（ROC-AUC）（下段）をまとめた。パラメーターは、検出指標の作成に当たって参照する時系列データの期間の長さ T=26、外れ値の発生確率p = 0.5 に固定した。もっともらしい値からの乖離の大きさ Error Size は、左より 3,2,1.5 を設定した。ROC-AUC のエラーバーは、ブートストラップ法を用いて算出した 95%信頼区間。

続いて PR 曲線を用いて、T=62、外れ値の乖離の大きさを 2 に設定した図 8 では、外れ値の発生率 p を 0.1 にした中図を見ると、SARIMA を検出指標とした際の PR-AUC が 0.452 と最も大きく、続いて ARIMA が 0.401、ChngPrvM が 0.335 だった。P を変えてもパフォーマンスの優劣は大きく変わらなかったが、p が小さいほど、同じ指標であっても PR-AUC が小さくなる傾向が見られた。また T=26 にした図 9 では、SARIMA の PR-AUC は ARIMA や ChgPrevM より小さくなった。

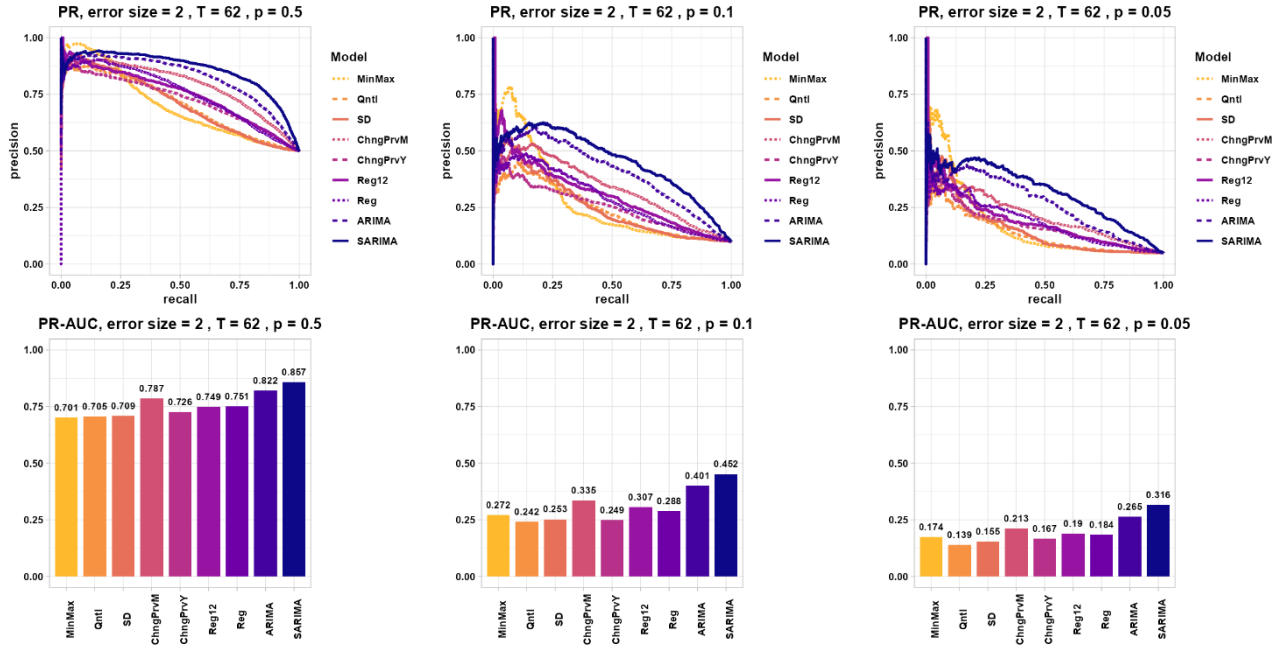

図 8　PR 曲線と PR-AUC（実データ、T=62, Error Size=2）

Note: 検出指標に関して、PR 曲線（上段）とその線下面積（PR-AUC）（下段）をまとめた。パラメーターは、検出指標の作成に当たって参照する時系列データの期間の長さ T=62、もっともらしい値からの乖離の大きさ Error Size=2 に固定した。外れ値の発生確率pについては、左より 0.5,0.1,0.05 を設定した。

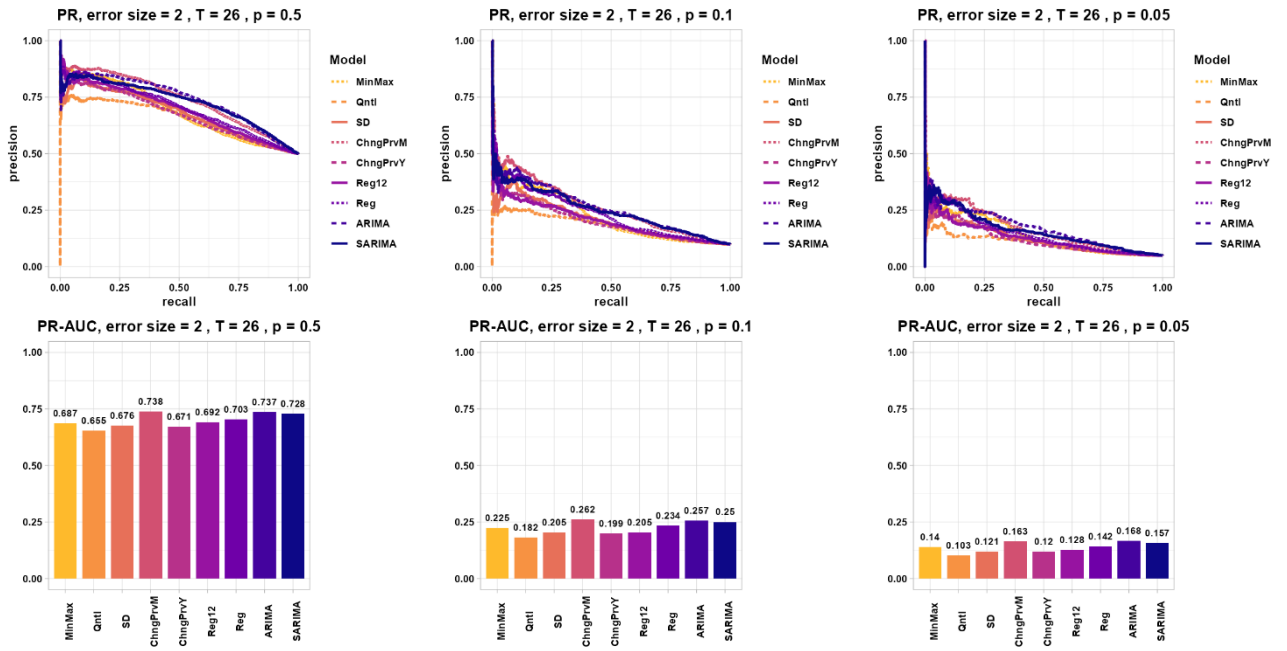

図 9　PR 曲線と PR-AUC（実データ、T=26, Error Size=2）

Note: 検出指標に関して、PR 曲線（上段）とその線下面積（PR-AUC）（下段）をまとめた。パラメーターは、検出指標の作成に当たって参照する時系列データの期間の長さ T=26、もっともらしい値からの乖離の大きさ Error Size=2 に固定した。外れ値の発生確率pについては、左より 0.5,0.1,0.05 を設定した。

各検出指標について最適な閾値を設定するために、F1 スコアを最大化する閾値を確認した（表 4）。Error Size=2, T=62, p=0.1 の時には、SARIMA では 0.893 を閾値に設定した時の F1 値が最も大きく 0.517 で、感度が 0.601、特異度が 0.920、適合率が 0.454、正解率が 0.888 であった。また閾値を変化させたときに取り得る F1 値の最大値は、SARIMA が最も大きかった。

表 4　各検出指標のパフォーマンス（実データ、T=62, Error Size=2）

						Sensitivity/			
				Threshold	F1 Score	Recall	Specificity	Precision	Accuracy
p	Indicator	ROC-AUC	PR-AUC	(under the maximum F1 score)					
0.5	MinMax	0.661	0.701	0.011	0.667	0.992	0.017	0.502	0.504
	Qntl	0.686	0.705	0.039	0.669	0.986	0.038	0.506	0.512
	SD	0.68	0.709	0.03	0.667	0.99	0.02	0.503	0.505
	ChngPrvM	0.794	0.787	0.745	0.738	0.808	0.617	0.678	0.712
	ChngPrvY	0.731	0.726	0.489	0.7	0.832	0.454	0.604	0.643
	Reg12	0.744	0.749	0.424	0.702	0.813	0.497	0.618	0.655
	Reg	0.741	0.751	0.97	0.697	0.783	0.537	0.629	0.66
	ARIMA	0.828	0.822	0.491	0.768	0.808	0.703	0.731	0.756
	SARIMA	0.87	0.857	0.586	0.809	0.83	0.777	0.789	0.804
0.1	MinMax	0.66	0.272	1.043	0.303	0.233	0.966	0.435	0.893
	Qntl	0.685	0.242	1.356	0.331	0.369	0.904	0.3	0.851
	SD	0.678	0.253	1.889	0.326	0.324	0.926	0.327	0.866
	ChngPrvM	0.795	0.335	1.384	0.411	0.546	0.877	0.33	0.844
	ChngPrvY	0.728	0.249	1.255	0.34	0.481	0.85	0.263	0.813
	Reg12	0.756	0.307	0.985	0.377	0.524	0.861	0.295	0.827
	Reg	0.737	0.288	2.323	0.376	0.405	0.916	0.35	0.865
	ARIMA	0.834	0.401	0.809	0.476	0.571	0.908	0.408	0.874
	SARIMA	0.872	0.452	0.893	0.517	0.601	0.92	0.454	0.888
0.05	MinMax	0.636	0.174	1.045	0.237	0.218	0.967	0.261	0.93
	Qntl	0.656	0.139	1.494	0.224	0.294	0.93	0.181	0.898
	SD	0.648	0.155	1.902	0.238	0.314	0.93	0.191	0.899
	ChngPrvM	0.802	0.213	1.703	0.303	0.396	0.936	0.246	0.909
	ChngPrvY	0.751	0.167	1.585	0.25	0.362	0.919	0.191	0.892
	Reg12	0.761	0.19	1.302	0.259	0.338	0.933	0.21	0.903
	Reg	0.747	0.184	2.383	0.284	0.402	0.925	0.219	0.899
	ARIMA	0.823	0.265	1.003	0.373	0.41	0.959	0.342	0.931
	SARIMA	0.875	0.316	1.004	0.415	0.518	0.949	0.347	0.927

Note: 検出指標のパフォーマンスをまとめた。パラメーターは、検出指標の作成に当たって参照する時系列データの期間の長さ T=62、もっともらしい値からの乖離の大きさ Error Size=2 に固定した。外れ値の発生確率pは、0.5, 0.1, 0.05 を設定した。Threshold より右の列については、各指標の F1 score を最大化する値に閾値を固定したときの値を報告している。

6　まとめ

本稿では、統計調査実務への応用を念頭に、外れ値の存否の検出に用いる指標を検討した。その結果、SARIMA、ARIMA については、シミュレーションと実データによる検証のそれぞれで、パフォーマンスが他の手法と比べて高かった。ChngPrvM については、シミュレーションと実データのそれぞれで、比較的パフォーマンスが高かった。回帰分析を必要としない比較的簡便な手法の中で、検出指標の次善策として候補になり得る。

SARIMA は、外れ値検出に使う過去のデータのサイズ T が大きいほどパフォーマンスが高くなった一方で、MinMax、Qntl、SD については、T が小さいほどパフォーマンスが高くなる場合もあり、使用するデータ期間が長いほど良い結果が得られるとは限らなかった。特に時系列データの特性を考慮しない検出指標では、昔と直近の値を同じ重みづけで扱うため、T を大きくすると最新の値との相関が弱い古いデータがノイズとなり、パフォーマンスが低下したと考えられる。

外れ値の乖離の大きさと検出のパフォーマンスについて、実データにおいて撹乱項の標準偏差の 3 倍の外れ値であれば最大で 0.97 程度、撹乱項の標準偏差の 2 倍の外れ値であれば最大で 0.87 程度の ROC-AUC が得られ、統計調査実務で用いるに十分なパフォーマンスを有する検出指標が存在すると分かった。一方で撹乱項の標準偏差の 1.5 倍の外れ値に対して ROC-AUC は最大でも 0.78 程度と、外れ値の乖離が小さいと検出のパフォーマンスは低下した。

また検出指標別に、F1 値を最大化する閾値を求めることができた。

本稿で得られた知見は、経済時系列データの作成と解釈における外れ値検出に応用されることが期待される。本稿の限界として、経済産業省生産動態統計を用いた分析に留まることから、サービス業等の製造業以外の分野における時系列データに関して、どのように外れ値を検出することが望ましいか把握するには更なる分析を要する。また金融危機等に起因するマクロ経済上の変動における外れ値検出について、取り立てて本稿では扱わなかったが、今後の課題として位置づけたい。

謝辞

本稿について有益なコメントをいただいた２名の匿名査読者に対して、深く感謝を申し上げたい。

参考文献

[1] 小林良行 (2009), 「ヨーロッパにおけるデータエディティング及び補定に関する調査報告 ～EDIMBUS プロジェクトを中心に～」, 『統計研究彙報』, 66, 101-129.

[2] 野呂竜夫・和田かず美 (2015), 「統計実務におけるレンジチェックのための外れ値検出方法」, 『統計研究彙報』, 72(72), 41-54.

[3] 高橋将宜 (2012), 「諸外国のデータエディティング及び混淆正規分布モデルによる多変量外れ値検出法についての研究」, 『統計センター 製表技術参考資料』, 17.

[4] 高橋将宜・伊藤孝之 (2013), 「経済調査における売上高の欠測値補定方法について ～多重代入法による精度の評価～」, 『統計研究彙報』, 70, 19-86.

[5] 高橋将宜・伊藤孝之 (2014), 「様々な多重代入法アルゴリズムの比較 ～大規模経済系データを用いた分析～」, 『統計研究彙報』, 71, 39-82.

[6] 和田かず美 (2010), 「多変量外れ値の検出 ～MSD 法とその改良手法について～」, 『統計研究彙報』, 67, 89-157.

[7] 和田かず美 (2012), 「多変量外れ値の検出 ～繰返し加重最小二乗（IRLS）法による欠測値の補定方法～」, 『統計研究彙報』, 69, 23-52.

[8] 和田かず美・野呂竜夫 (2019), 「ロバスト回帰推定へのウェイト関数や残差尺度の影響について」, 『統計研究彙報』, 76, 101-113.

[9] Berge, T. J., & Jordá, Ó. (2011), *Evaluating the classification of economic activity into recessions and expansions*, American Economic Journal: Macroeconomics, 3(2), 246–277.

[10] Brown, I., & Mues, C. (2012), *An experimental comparison of classification algorithms for imbalanced credit scoring data sets*, Expert Systems with Applications, 39(3), 3446–3453.

[11] Celhay, P., Meyer, B. D., & Mittag, N. (2024), *What leads to measurement errors: Evidence*

from reports of program participation in three surveys, Journal of Econometrics, 238(2), 105581.

[12] Fawcett, T. (2007), *ROC graphs: Notes and practical considerations for researchers*.

[13] Zio, D. M., Fursova, T., Gelsema, T., Gießing, S., Guarnera, U., Petrauskienė, J., Quensel von Kalben, L., Scanu, M., ten Bosch, K. O., van der Loo, M., & Walsdorfer Nadežda, K. (2018), *Methodology for data validation 1.1*, Essnet Validat Foundation.

[14] Orietta, L., Waal, T., Hulliger, B., Di Zio, M., Pannekoek, J., Kilchmann, D., Guarnera, U., Hoogland, J., Manzari, A., & Tempelman, C. (2007), *Recommended practices for editing and imputation in cross-sectional business surveys*.

[15] Saito, T., & Rehmsmeier, M. (2015), *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets*, PLoS One, 10(3), e0118432.

[16] Schifeling, T., Reiter, J. P., & DeYoreo, M. (2016), *Data fusion for correcting measurement errors*, arXiv.