資料５

独立行政法人
統計センター

National Statistics Center

総消費動向指数（CTIマクロ）作成における民間ビッグデータ等の利活用について

～ベイジアンモデル平均化法を用いた公的統計の予測～

令和８年３月27日

（独）統計センター　技術研究開発課

公的統計をつくる　活かす　支える

目次

1. 主な研究課題

2. 民間データを単独で用いたサービス産業動向調査の予測

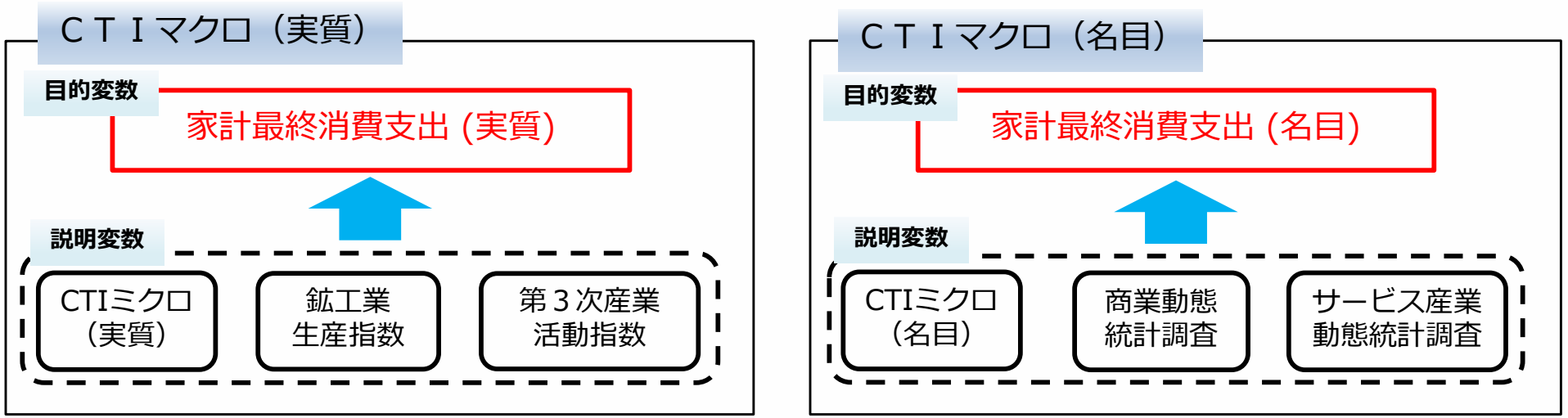
3. ベイジアンモデル平均化法（BMA）の概略

4. BMAによるサービス産業動向調査の予測と重要度が高い変数に特殊な変動があった場合のシミュレーション

5. まとめと課題

１．主な研究課題

現行のCTIマクロの時系列回帰モデルによる推定イメージ



- CTIマクロは、CTIミクロと４種類の公的統計を利用して推定している。
- 利用する公的統計のうち、経済産業省「第３次産業活動指数」と総務省「サービス産業動態統計調査」（旧：「サービス産業動向調査」）は公表時期の関係で、CTIマクロ推計時には最新月の公表結果が利用できない状況。現状では自身の値を状態空間モデルによって1カ月先を予測し、その予測値を推定時に利用している。そのため、経済に大きなショックが起きた時に、CTIマクロの推定にそのショックを充分取り込めない可能性がある。
- 民間データは入手時期が早く、CTIマクロ推定時には最新月のデータが揃っている状況。

入手時期が早い民間データ等を利用し、CTIマクロ推定時に利用する公的統計を予測できないか、研究を続けているところ。

研究課題の概要

CTIマクロ推定月のデータ状況イメージ図（数値は架空）

	サービス産業動態統計調査	商業動態統計調査	民間データ等
n－3月	340	500	672
n－2月	352	520	701
n－1月	382	541	720
n月	未公表	495	650

主な研究課題

① 民間データ等から、公表前の公的統計を予測すること。
② 特殊な変動要因(※)が生じても、安定した予測結果を得られるようにすること。

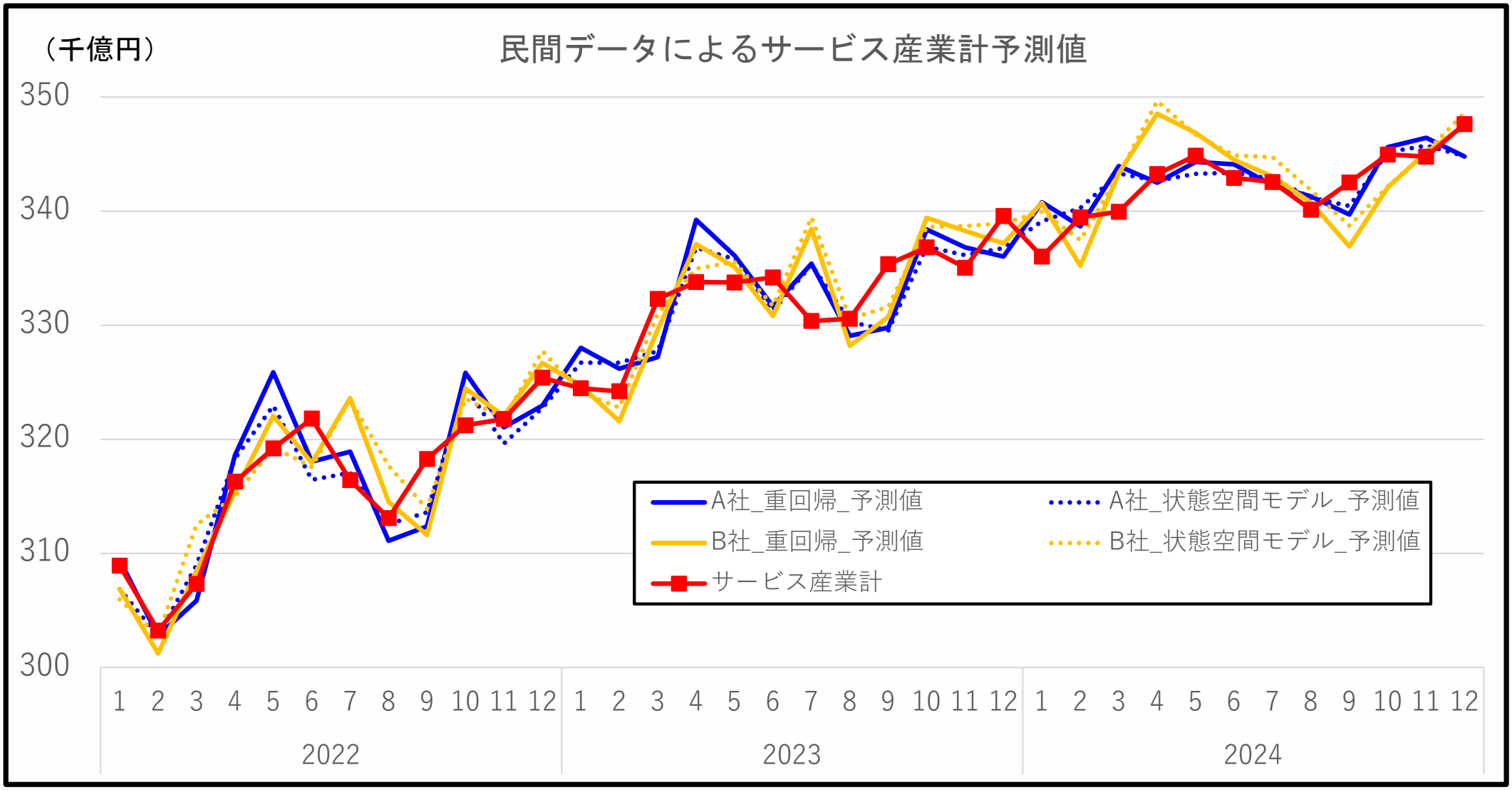
※民間データ等の分類の変更、販売促進のためのキャンペーンの実施等が考えられる。

→ ①・②に同時に資する方法として、ベイジアンモデル平均化法の利用を検討。

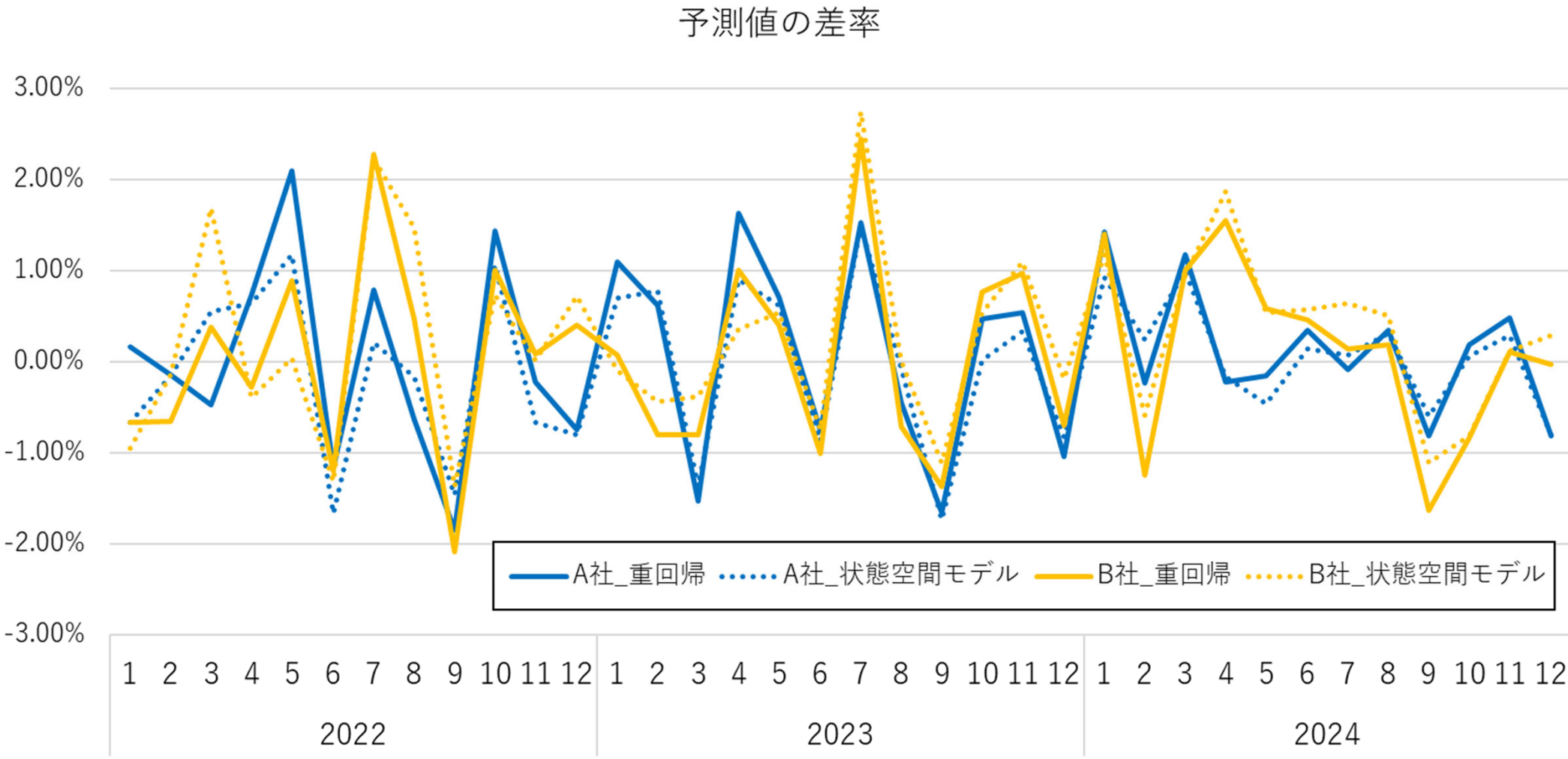
２．民間データを単独で用いたサービス産業動向調査の予測

単独の民間データを用いた取組み①

- 民間ビッグデータ（Ａ社・Ｂ社）を単独で使い、重回帰分析や状態空間モデル等を利用しサービス産業動向調査を予測。



単独の民間データを用いた取組み②



- それぞれのデータを単独で用いても、一定の精度での予測は可能。
- データに特殊な変動が起きた時の対応策の研究に着手中。

３．ベイジアンモデル平均化法（BMA）の概略

モデル平均化法の考え方

○ 例として世帯の1か月の消費額（C）を、収入（Y）、保有資産（S）、世帯人員（Z)の３つのデータから推定することを考える。

一般的な推定方法

① 有名な特定の関数を使って推定する方法。例えば、ケインズ型消費関数を使って推定する方法。

$$C = \beta_0 + \beta_1 Y$$

② 重回帰分析を用いてモデルを特定する方法。

$$C = \beta_0 + \beta_1 Y + \beta_2 S + \beta_3 Z$$

さらにAIC等を使ってモデル選択を行い、推定する方法。

$$C = \beta_0 + \beta_1 Y + \beta_2 S$$

モデル平均化法

- 複数のモデルを作成し、それらを何らかの方法で重みづけし、それを加重平均して推定結果とする推定方法。
  各モデルのウエイトをw_iとすると

$$C_1 = w_1 \cdot f(Y)$$
$$C_2 = w_2 \cdot f(S)$$
$$C_3 = w_3 \cdot f(Z)$$
$$C_4 = w_4 \cdot f(Y, S)$$
$$C_5 = w_5 \cdot f(Y, Z)$$
$$C_6 = w_6 \cdot f(S, Z)$$
$$C_7 = w_7 \cdot f(Y, S, Z)$$
$$C_8 = w_8 \cdot f(\text{何も選ばない})$$

推定結果は加重平均なので、全てのモデルを使う場合、$\sum_{i=1}^{8} C_i$となる（ウエイトの高い上位n個のモデルのみを使う方法もある）。

モデル平均化法の実装上の利点

複数のモデルから予測値を算出可能

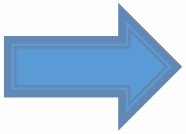
- A社データ・B社データを統合することなく、A社・B社両社のデータを使った予測値を算出可能。
- また、特定の１つのモデルを選ばないため、モデル選択の不確実性への対応も可能。

データに特殊な変動が生じた時の対応

- データに特殊な変動が生じた際、一旦そのデータを外すことが可能。

$$C_1 = w_1 \cdot f(Y)$$
$$C_2 = w_2 \cdot f(S)$$
$$C_3 = w_3 \cdot f(Z)$$
$$C_4 = w_4 \cdot f(Y,S)$$
$$C_5 = w_5 \cdot f(Y,Z)$$
$$C_6 = w_6 \cdot f(S,Z)$$
$$C_7 = w_7 \cdot f(Y,S,Z)$$
$$C_8 = w_8 \cdot f(\text{何も選ばない})$$

世帯人員(Z)に特殊な変動が生じた場合



データZを一旦外しても、予測結果を算出可能。

$$C_1 = w_1 \cdot f(Y)$$
$$C_2 = w_2 \cdot f(S)$$
$$C_3 = w_3 \cdot f(Y,S)$$
$$C_4 = w_4 \cdot f(\text{何も選ばない})$$

複数のモデルを使った予測が可能な手法。かつデータに特殊な変動が生じた時の対応としても有力。

BMAの計算方法の概略①

✓ 説明変数xの候補がk個あるとき、考えうるモデルを下記のイメージで列挙し、各モデルのパラメータと各モデルの事後確率を、マルコフ連鎖モンテカルロ法（MCMC)を用い推定する。

$$M_1 : \boldsymbol{Y} = \beta_0 + \beta_1 x_1 + \varepsilon$$
$$M_2 : \boldsymbol{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$
$$M_3 : \boldsymbol{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$
$$\vdots$$
$$M_{full} : \boldsymbol{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon$$

$M = \{M_1, M_2 \cdots, M_k\}$：候補となるモデル族
$\boldsymbol{Y} = \{y_1, y_2, \cdots, y_n\}$：目的変数

✓ ベイズ推定する際は、各モデルのパラメータβ_iと各モデルの事前確率が必要。

✓ 今回の報告では、β_iの事前分布にはZellnerのg事前分布を、各モデルの事前確率$P(M_i)$にはベータ二項分布を使用している。

BMAの計算方法の概略②

○ PMP（posterior model probability）

各モデルの良しあしを測るための指標。

モデルM_iのPMPは以下の式で定義される。

$$\mathrm{PMP}(M_i) = \frac{P(\boldsymbol{Y}|M_i)P(M_i)}{\sum_{j=1} P(\boldsymbol{Y}|M_j)P(M_j)} (= w_i)$$

$P(\boldsymbol{Y}|M_i)$：モデルM_iの尤度　$P(M_i)$：モデルM_iの事前確率

分母が各モデルの事後確率の合計、分子がPMPを計算したいモデルの事後確率である。

このPMPが各モデルのウエイトw_iになる。

○ PIP（posterior inclusion probabilites）

説明変数の重要度を表す指標。

説明変数x_iのPIPは次式で定義される。

$$\mathrm{PIP}(x_i) = \sum_{j=1} \delta(x_i) \times \mathrm{PMP}(M_j)$$

$\delta(x_i)$：説明変数x_iがモデルM_jに含まれれば1、そうでなければ0を取る関数

○ BMAの推定値

各モデルM_iの推定値を$\hat{y}_i$とすると、BMAによる推定値$\hat{y}$は各モデルの加重平均であるから、次式で定義される。

$$\hat{y} = \sum_{i=1} \hat{y}_i \times w_i$$

モデル平均化法を使った先行研究事例

○ＧＤＰの予測

- Fernández, Ley and Steel (2001)

→説明変数が全て揃う72か国のデータを用い、41の説明変数を使って成長回帰分析をBMAを使って行った。最も事後確率が高いモデルでもその確率は1.24%で、多くのモデルに事後確率が分散されていた。

- Ley and Steel(2009)

→上記Fernández, et al.(2001)等のデータを用い、BMAで推定する際のモデルの事前確率を変え同じく成長回帰分析を行っている。重要とされる説明変数が使う事前分布によって変わり得ることが示された。

- Bencivelli, Marcellino and Moretti (2017)

→BMAブリッジモデルを用いてユーロ圏、ドイツ、フランス、イタリアのGDPを予測。ドイツ、フランス、イタリアの予測において、BMAブリッジモデルに基づく予測は、標準的なブリッジモデルよりも予測誤差が小さいとしている。

○インフレ率の予測

- Koop and Korobilis (2012)

→Dynamic Model Averaging (DMA)という手法を使い、米国のインフレ率を予測。TVPモデル等よりも精度の高い予測結果を得られたと報告。

４．　BMAによるサービス産業動向調査の予測と重要度が高い変数に特殊な変動があった場合のシミュレーション

予測方法の説明①

- 前述したベイジアンモデル平均化法を利用し、サービス産業動向調査の「サービス産業計」の予測を行った。説明変数を変えパターン①・②で予測を行っている。
- 比較のため、パターン②の説明変数をAICを使って変数選択し、重回帰分析を使った予測値も試算している。

パターン①（５変数）

- ✓ 日経平均月末終値（1期ラグ）
- ✓ 新車登録台数(※1)（普通車＋小型車）
- ✓ サービス産業計の1期ラグ
- ✓ A社総額データ
- ✓ B社総額データ

パターン②（11変数）

- ✓ 日経平均月末終値（1期ラグ）
- ✓ 新車登録台数(※1)（普通車＋小型車）
- ✓ サービス産業計の1期ラグ
- ✓ A社分類別データ（7分類）(※2)
- ✓ B社総額データ

※1：データ出所：国土交通省「自動車保有車両数」
※2：過去の研究から、説明力のありそうな「宿泊業」、「飲食業」、「旅行・レジャー業」、「交通」、「医療・介護」、「通信」、「その他」の7つを選定。

予測方法の説明②

データの前処理

- ✓ 全ての変数を1か月ごとにDecomp法を用い、時系列変動成分のうち季節成分とノイズ成分を除去。
- ✓ 見せかけの回帰への対応のため、全ての変数について対数差分を取る。

推定するモデル族

$$M_1: \Delta \ln(y) = w_1 \cdot \{\beta_0 + \beta_1 \Delta \ln(x_1) + \varepsilon\}$$
$$M_2: \Delta \ln(y) = w_2 \cdot \{\beta_0 + \beta_1 \Delta \ln(x_1) + \beta_2 \Delta \ln(x_2) + \varepsilon\}$$
$$\vdots$$
$$M_{full}: \Delta \ln(y) = w_{ful} \cdot \{\beta_0 + \beta_1 \Delta \ln(x_1) + \beta_2 \Delta \ln(x_2) + \cdots \beta_k \Delta \ln(x_k) + \varepsilon\}$$

Δ：階差　y：サービス計　x_i：説明変数
w_i：モデルiのウエイト　β_i：パラメータ

推定に利用する事前分布等

- ✓ パラメータβ_iの事前分布：Zellnerのg事前分布
- ✓ 各モデルの事前確率$P(M_i)$：ベータ二項分布
- ✓ MCMCは4000回実行（最初1000回はburn-in分）。

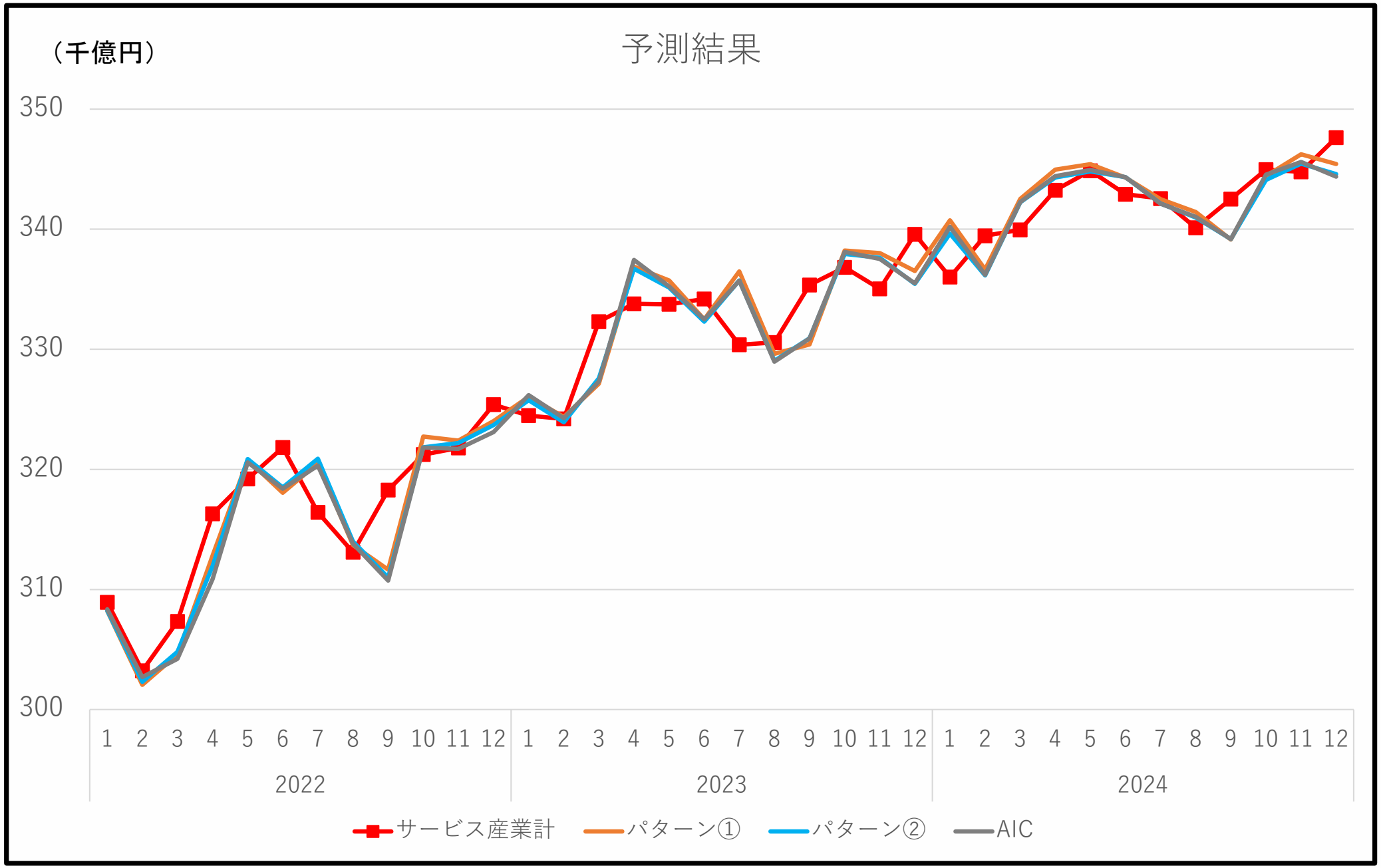
予測値の算出

- ✓ 推定に使用するデータは、2017年4月～2024年12月。
  予測期間は2022年１月～2024年12月。
  予測したい前の月(n-1月)までのデータを学習データとし、予測月(n月)のサービス産業を予測する。
- ✓ サービス産業計の対数差分を推定したことになるため、変化率の近似値を推定したとみなし、以下の算式で水準の予測値を算出する。

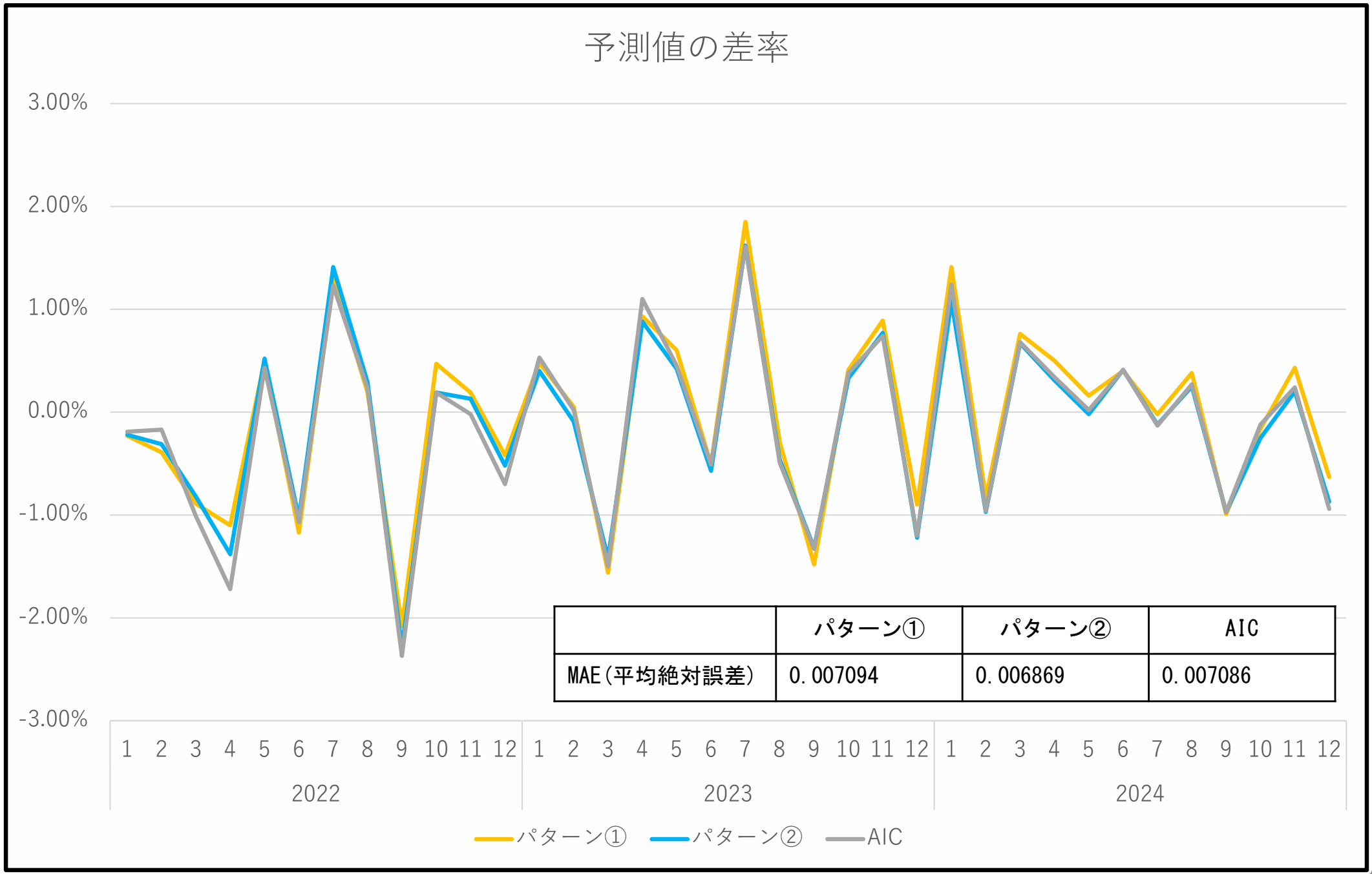
$$\hat{y}_t = y_{(t-1)} \times (1 + \Delta\widehat{\ln(y_{(t)})})$$

$\hat{y}_t$：t月のサービス産業計の予測値　$y_{(t-1)}$：t-1月のサービス産業計実測値
$\Delta\widehat{\ln(y_{(t)})}$：t月の民間データから推定したt月のサービス産業計予測変化率

予測結果


（千億円）
予測結果
350
340
330
320
310
300
1 2 3 4 5 6 7 8 9 10 11 12 1 2 3 4 5 6 7 8 9 10 11 12 1 2 3 4 5 6 7 8 9 10 11 12
2022
2023
2024
サービス産業計
パターン①
パターン②
AIC

差率



	パターン①	パターン②	AIC
MAE(平均絶対誤差)	0.007094	0.006869	0.007086

予測結果の概要

予測精度

- ✓ MAEを計算すると、パターン①：0.007094、パターン②：0.006869、AIC：0.007086となり、予測精度としては高いと思われる。
- ✓ 予測結果の精度は、パターン②＞AIC＞パターン①の順であった。推定に利用する変数を増せば、より精度を上げられる可能性が示唆された。

重要度の高い（PIPの高い）変数

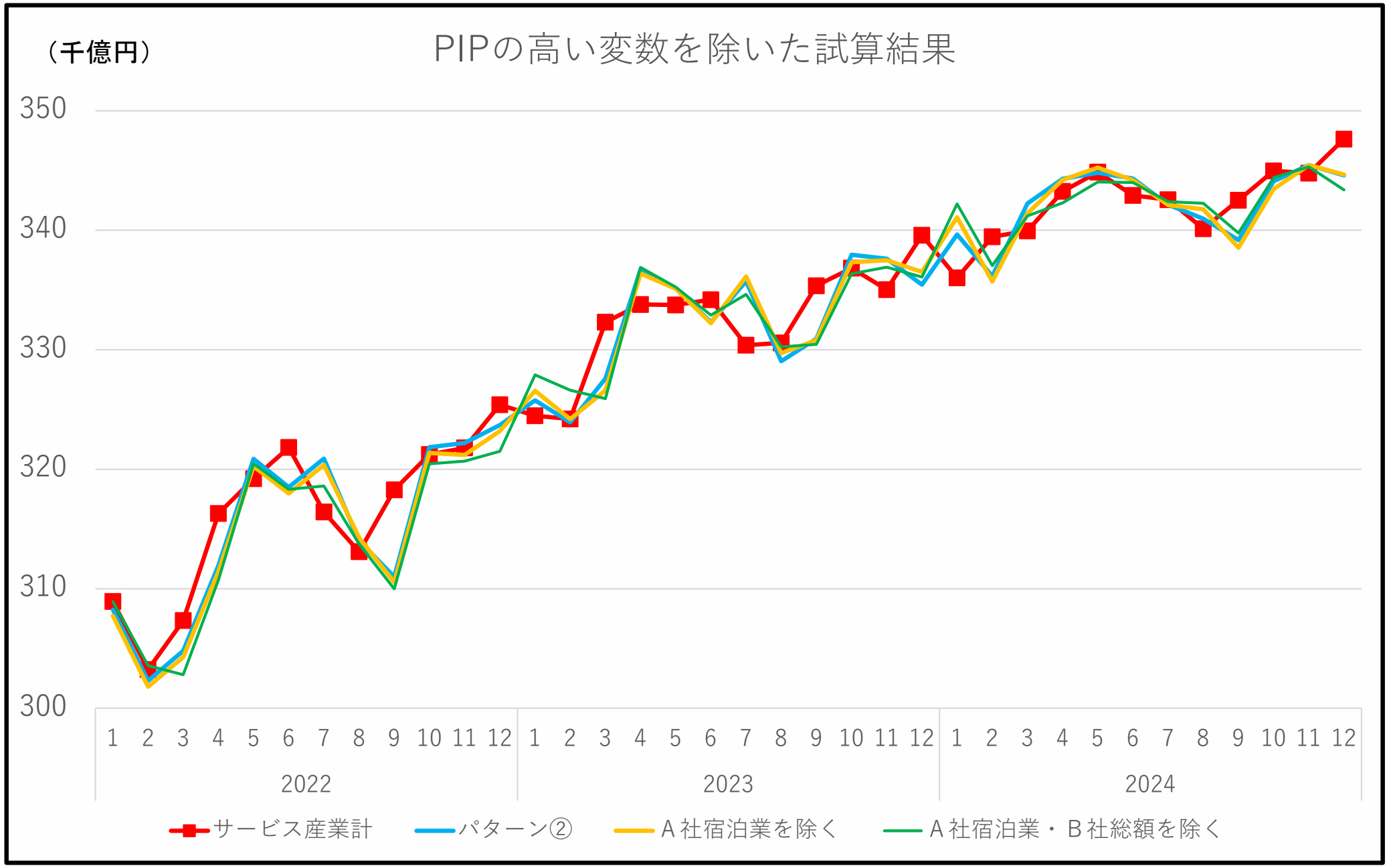
✓ パターン①
- A社総額データ
- B社総額データ
- サービス産業計の一期ラグ
- 日経平均

✓ パターン②
- A社宿泊業データ
- A社医療介護データ
- B社総額データ

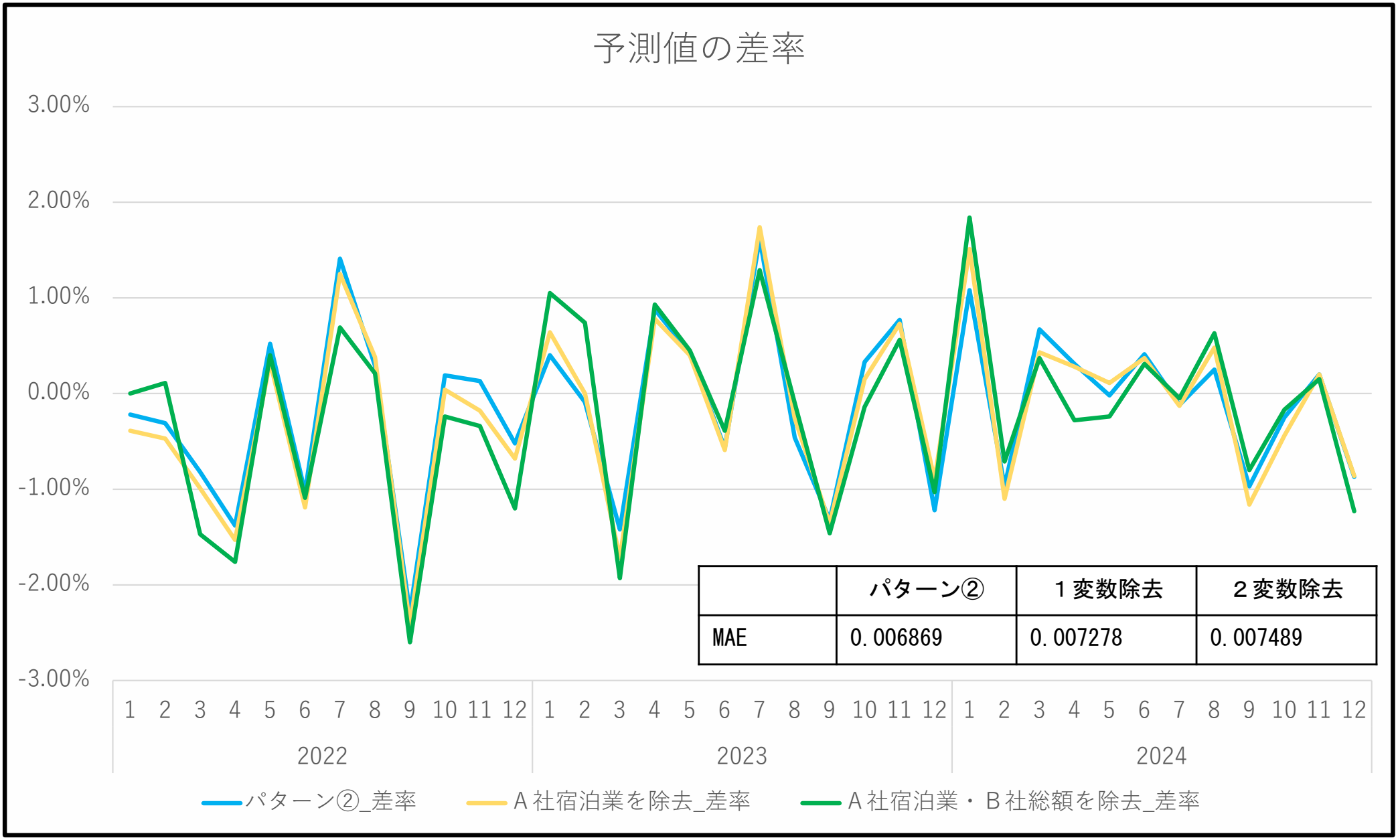
データに特殊な変動が生じた場合のシミュレーション

✓ パターン②において、PIPの高い変数に特殊な変動が生じたと仮定。
- ➢ A社宿泊業データの１変数を除いたパターンと、A社宿泊業データ・B社総額データの２変数を除いたパターンで予測を行った。

PIPの高い変数を除いた試算結果



PIPの高い変数を除いた試算結果 差率



	パターン②	１変数除去	２変数除去
MAE	0.006869	0.007278	0.007489

PIPの高い変数を１～２個除いても、一定の予測精度が確保できた。

５．まとめと課題

まとめと課題

まとめ

- ✓ 今回の予測試算の結果、ベイジアンモデル平均化法は利用するデータを増やすことで、予測精度を上げる可能性があることが示唆された。
- ✓ 仮に民間データ等に特殊な変動が生じた際も、一旦そのデータを外して予測試算を行うという対応が可能。

課題

- ✓ ベイジアンモデル平均化法は、（恐らく）日本の公的統計で初めて検討される手法で、本手法に対する理解が不十分。
- ✓ 今回はあくまでも試算にとどまっており、推定に利用する事前分布等も現段階では検討が不十分。
- ✓ 他のモデル平均化法との比較やダイナミック・ファクター・モデル等の検討。
- ✓ 実装を目指すための課題等の抽出と検討。

参考文献

Bencivelli, L., Marcellino, M., & Moretti, G. (2017). Forecasting economic activity by Bayesian bridge model averaging. *Empirical Economics*, 53(1), 21-40.

Fernandez, C., Ley, E., & Steel, M. F. (2001). Model uncertainty in cross - country growth regressions. *Journal of applied Econometrics*, 16(5), 563-576.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and EI George, and a rejoinder by the authors). *Statistical science*, 14(4), 382-417.

Koop, G., & Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3), 867-886.

Ley, E., & Steel, M. F. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of applied econometrics*, 24(4), 651-674.