

小地域推定各手法の労働力調査への適用

～ 都道府県別完全失業率の推定 ～

高部 勲[†]

労働力調査の都道府県別結果に対する要望が、近年高まってきている。しかし小地域の結果を標本調査だけから推定すると、標本数が不足して、標本誤差が大きくなってしまふ。このような標本誤差の増加を防ぎ、小地域の結果を精度よく推定する方法については、現在も活発な研究が続けられている。様々な小地域推定の手法の中でも、特に有力であると考えられる手法を労働力調査へ適用し、都道府県別月次の完全失業率を推定した結果についてまとめたものが本稿である。

Application of the Small Area Estimation Methods to the Labour Force Survey : Prefecture Level Unemployment Rate Estimation

Isao Takabe

Recently, the demand for the prefecture level labor force statistics has increased. Direct small area estimation, however, causes large variance because of its few samples. There are various methods to estimate this kind of small area statistics accurately. In this paper, we applied some of these methods to estimate prefecture level monthly unemployment rate, and the results are evaluated.

Key words : 時系列回帰モデル, EBLUP 推定量, Time Series and Cross Sectional モデル

はじめに

公表されている統計表よりも、更に細かい項目を対象とした推定は、小地域推定(Small Area Estimation)と呼ばれる。官庁の政策決定のための資料や、ビジネスにおける意思決定のための材料として、小地域推定に対する要望が、近年ますます高まってきている。

小地域の結果を標本調査だけから推定すると、各小地域に割り当てられる標本数が不足し、標本誤差が大きくなってしまふ。標本誤差の増加を防ぎ、小地域の推定を精度良く行うための統計的手法に関する研究は、数多く行われている。小地域推定の手法を広範にまとめた体系的な書籍としては、Rao(2003)がある。

様々な小地域推計の手法の中で、特に有力であると考えられる、時系列回帰モデル、EBLUP (Empirical Best Linear Unbiased Prediction) 推定量、及び Rao-Yu モデルの各手法を労働力調査へ適用して、都道府県別の完全失業率を推定した結果についてまとめたものが本稿である。

本稿は全部で四つの章から構成されている。第 1 章から第 3 章までが、時系列回帰モデル、EBLUP 推定量及び Rao-Yu モデルの各手法の説明にそれぞれ対応しており、これらの章が本稿の中心的部分である。第 4 章では、小地域推定で用いられる、標本誤差とその系列相関についての解説を行う。第 4 章は、第 1 章から第 3 章までの内容を補完するという位置づけになっている。

以下、各章の内容について紹介する。

第1章では、時系列回帰モデルによる小地域推定について解説する。この手法は実際にアメリカの州別月次完全失業率などの推定に用いられている。これは完全失業率を確率過程とみなして、回帰項、トレンド項、季節項、標本誤差項、不規則変動項にそれぞれ分解し、そこから標本誤差のみを取り除くという手法である。推定を効率的に行うために、状態空間モデルおよびカルマンフィルタの手法が適用されている。

第2章では、EBLUP 推定量を用いた小地域推定について解説する。この手法では、地域（空間）の情報が推定に利用される。推定に用いられる Fay-Herriot モデルは、単純な回帰モデルと違って、小地域ごとの効果も含んだモデルになっており、小地域ごとの特徴を推定に利用することができると考えられる。

第3章では、Rao-Yu モデルによる小地域推定を解説する。Rao-Yu モデルは、Time Series and Cross-sectional モデルと呼ばれるモデルの一つであり、第2章の Fay-Herriot モデルを多変量に拡張したものとなっている。Rao-Yu モデルでは、時系列情報と空間情報の両方を推定に利用することができる。複雑なモデルによる推定を効率よく行うために、階層的ベイズモデル（Hierarchical Bayes Model）の手法及びマルコフ連鎖モンテカルロ法（Markov Chain Monte Carlo Method）が適用されている。

第4章では、都道府県別月次の標本誤差及び標本誤差の系列相関の推定について述べている。ここで推定した結果が、第1章から第3章までのすべての手法で用いられる。

なお各章の終わりには、計算に用いた SAS および R のプログラムを収録している。

第 1 章 アメリカ BLS における小地域推定

～時系列回帰モデル及びカルマンフィルタの利用～

アメリカの BLS (Bureau of Labor Statistics) は、州別完全失業率の結果を毎月公表している。そこでは時系列回帰モデル及び状態空間モデルの手法が用いられている。本稿では、このアメリカの手法を日本の労働力調査に適用し、都道府県別の完全失業率を推定した結果についてまとめている。以下 I 章では、アメリカの小地域推定に用いられるモデルについて解説する。II 章では、推定で重要となる状態空間モデル及びカルマンフィルタの理論を紹介する。III 章ではそのほかの話題について触れる。IV 章では、我が国の労働力調査の結果に適用した場合の結果等についてまとめている。V 章には今後の課題について述べている。

I アメリカの小地域推計の手法

1 シグナル + ノイズモデル

小地域ごとの観測値データを、シグナルとノイズという独立な二つの確率過程の和で表現する。

$$y(t) = \theta(t) + e(t) \quad (1)$$

$$\left[\begin{array}{l} y(t) : \text{観測値} \quad (t=1 \cdots T) \\ \theta(t) : \text{真の値の推定値 (シグナル)} \\ e(t) : \text{標本誤差の推定値 (ノイズ)} \end{array} \right]$$

このようなモデルを、シグナル + ノイズモデルと呼ぶ。 $y(t)$ から $e(t)$ を取り除くことが、この手法の目的である。 $\theta(t)$ は、さらに次のような四つの項に分解できると仮定する。

$$\theta(t) = M(t) + T(t) + S(t) + I(t) \quad (2)$$

$$\left[\begin{array}{l} M(t) : \text{回帰項} \quad T(t) : \text{トレンド項} \\ S(t) : \text{季節項} \quad I(t) : \text{不規則変動項} \end{array} \right]$$

$M(t)$ 、 $T(t)$ 、 $S(t)$ 、 $I(t)$ 及び $e(t)$ が従うモデルについて、以下で解説する。なお $e(t)$ のモデルはやや複雑なので、節を改めて解説する。

(1) 回帰項 $M(t)$ のモデル

回帰項 $M(t)$ は、 $\theta(t)$ の中で補助情報によって説明される部分を表す。

$$M(t) = X(t)\beta(t) \quad (3)$$

$$\beta(t) = \beta(t-1) + v_{\beta}(t) \quad (4)$$

$$\left[\begin{array}{l} X(t) : \text{補助情報 (説明変数)} \\ v_{\beta}(t) \sim N[0, \sigma_{\beta}^2] \end{array} \right]$$

回帰項には単回帰を想定している。式 (4) では、回帰係数 $\beta(t)$ が、時間とともに緩やかに変化することを許容している。

(2) トレンド項 $T(t)$ のモデル

トレンド項 $T(t)$ は、 $\theta(t)$ の中の緩やかな変動を抽出するために導入される。トレンド項は、各時点での水準を表す部分 $T(t)$ と、傾きを表す部分 $R(t)$ を用いて、次のように表される。

$$T(t) = T(t-1) + R(t-1) + v_T(t) \quad (5)$$

$$R(t) = R(t-1) + v_R(t) \quad (6)$$

$$\left[v_T(t) \sim N(0, \sigma_T^2), v_R(t) \sim N(0, \sigma_R^2) \right]$$

$\sigma_R^2 = \sigma_T^2 = 0$ の場合には、傾き $R(t)$ が一定である 1 次の多項式（直線）をトレンドに仮定していることになる。 $\sigma_R^2 > 0$ 又は $\sigma_T^2 > 0$ の場合には、傾き及び水準がそれぞれ独立にランダムウォークに従って変化する、柔軟なモデルになっている。

回帰項によってトレンドを十分に説明できる場合には、トレンド項が不要となり、自動的にモデルから落ちる。この項は、説明変数で取り出しきれない残差のトレンド（residual trend）を取り出すために導入されている。

(3) 季節項のモデル

季節項は、シグナルの中の季節的な変動を抽出するために導入される。季節的な変動とは、時系列の成分のうちで周期が一年であるものを指す。季節項は、季節周波数（12 か月周期、6 か月周期、4 か月周期、3 か月周期、2.4 か月周期、2 か月周期）に対応する六つの三角関数の項の和として、次のように表現される。

$$S(t) = \sum_{j=1}^6 S_j(t) \quad (7)$$

$$\left[\begin{array}{l} S_j(t) = \cos(\omega_j) S_j(t-1) + \sin(\omega_j) S_j^*(t-1) + v_{S_j}(t) \\ S_j^*(t) = -\sin(\omega_j) S_j(t-1) + \cos(\omega_j) S_j^*(t-1) + v_{S_j^*}(t) \\ v_{S_j}(t) \sim N(0, \sigma_s^2), v_{S_j^*}(t) \sim N(0, \sigma_s^2), \omega_j = \frac{2\pi j}{12} \end{array} \right]$$

$v_{S_j}(t)$ 、 $v_{S_j^*}(t)$ はすべて互いに独立に、分散一定 (σ_s^2) の正規分布に従うとしている。回帰

項によって季節性を十分に説明できる場合には、この項が不要となり、自動的にモデルから落ちる。この項は説明変数で取り出しきれない残差の季節性（residual seasonality）を取り出すために導入される。

(4) 不規則変動項のモデル

不規則変動項は、上記の各項に入らない、残りの変動を表す。ここには景気の短期的な変動などが入る可能性がある。

$$I(t) = v_I(t) \quad (8)$$

$$\left[v_I(t) \sim NID(0, \sigma_I^2) \right]$$

$\theta(t)$ が、 $M(t)$ 、 $T(t)$ 、 $S(t)$ によってほとんど説明されてしまう場合には、この項が不要となり、自動的にモデルから落ちる。

2 シグナルのモデルの一部変更

前節で紹介したモデルのうち、 $T(t)$ 、 $S(t)$ については、日本の労働力調査への適用に当たって若干の変更を加えている。以下で、その変更点について解説する。

(1) $T(t)$ のモデルの変更

トレンド項 $T(t)$ を式 (5)、(6) で定義した場合、 $T(t)$ がかなり変動することが予想される。回帰項 $M(t)$ の係数も時間とともに変動すると仮定しているので、 $T(t)$ はもっと単純なモデルで良いと考えられる。そこで $T(t)$ には、以下のような 1 次のトレンドを仮定する。

$$T(t) = T(t-1) + v_T(t) \quad (9)$$

1 次のトレンドでも、原系列のトレンドを十分取り出すことが可能であると考えられる。

(2) $S(t)$ のモデルの変更

季節項 $S(t)$ には、以下の DECOMP 式 (北川 (1993)、北川 (1997)) の定義を採用する。

$$S(t) = - \sum_{j=1}^{11} S(t-j) + v_s(t) \quad (10)$$

式 (10) による定義の方が、モデルが簡潔であり、後に述べる状態空間モデルを用いた表現の際にも記述が簡単になる。

3 $e(t)$ のモデル

標本誤差 $e(t)$ のモデリングが、アメリカの小地域推定の中で最も重要な部分である。標本にはローテーション構造がある場合には、標本誤差に系列相関があると考えられる (第 4 章を参照)。また標本誤差の分散は、①標本数、②標本設計の変更、③真の値 $\theta(t)$ の三つの影響を受けて、時間とともに変化していくと考えられる。よって $e(t)$ のモデルに以下のような仮定をおく。

仮定 1：標本誤差項 $e(t)$ は、系列相関を持つ。

仮定 2：標本誤差 $e(t)$ の分散は、時点 t により異なる。

標本設計から自然に導かれるこれら二つの仮定を $e(t)$ のモデルに取り込むために、標本誤差項 $e(t)$ を以下のような乗法型モデルで表現する。

$$e(t) = \gamma(t)e^*(t) \quad (11)$$

$$\left[\begin{array}{l} \gamma(t) : \text{標本誤差の分散の変化を表すスカラー} \\ e^*(t) : \text{ローテーション構造による系列相関を表す AR (13)過程} \end{array} \right]$$

$e^*(t)$ は標本誤差項 $e(t)$ の変動パターンを表す AR(13)モデルである。 $\gamma(t)$ は標本誤差項 $e(t)$ の変動の振れ幅を表す。これらの動きが結びついて、標本誤差 $e(t)$ になる。 $e^*(t)$ と $\gamma(t)$ の推定方法について、以下に述べる。

(1) $e^*(t)$ の推定方法

標本誤差は 1 次・11 次・12 次・13 次において自己相関を持つと想定される（第 4 章参照）。よって AR(13)モデルで $e^*(t)$ を表現する。

$$e^*(t) = \sum_{j=1}^{13} \phi_j e^*(t-j) + v_e(t) \quad (12)$$

$$\left[v_e(t) \sim N(0, \sigma_{v_e}^2) \right]$$

標本誤差の自己相関 $\hat{\rho}(k)$ は、第 4 章で解説している方法で推定できる。よって $e^*(t)$ の AR 係数 ϕ_i ($i=1 \cdots 13$) は、以下の Yule-Walker 方程式を解くことにより推定できる（尾崎、北川（1998））。

$$\begin{bmatrix} 1 & \hat{\rho}(1) & \cdots & \hat{\rho}(12) \\ \hat{\rho}(1) & 1 & \cdots & \hat{\rho}(11) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}(12) & \hat{\rho}(11) & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{13} \end{bmatrix} = \begin{bmatrix} \hat{\rho}(1) \\ \hat{\rho}(2) \\ \vdots \\ \hat{\rho}(13) \end{bmatrix} \quad (\text{Yule-Walker 方程式}) \quad (13)$$

(2) $\gamma(t)$ の推定方法

各時点 t における標本誤差 $e(t)$ の分散を $\sigma_e^2(t)$ とする。 $\sigma_e^2(t)$ は、毎月の個票データから推定することが出来る（第 4 章を参照のこと）。 $e(t)$ の分散が $\sigma_e^2(t)$ に等しくなるように調整するための係数が、 $\gamma(t)$ である。そのためには、以下のようにして $\gamma(t)$ を構成すればよい。

まず AR 過程を MA(∞) 過程に反転させる。このとき AR 仮定の定常性の条件は満たされていると仮定する。

$$e^*(t) = \sum_{k=0}^{\infty} g_k v_e(t-k) \quad (14)$$

$$\left[\begin{array}{l} g_k : \text{インパルス応答関数} \\ g_k = 1 \text{、} g_k = \sum_{j=1}^k \phi_j g_{k-j} \text{ (} k > j \text{ のとき } \phi_j = 0 \text{)} \end{array} \right]$$

このとき $e^*(t)$ の分散 $\sigma_{e^*}^2$ は、次のようにして計算することができる。

$$\sigma_{e^*}^2 = \text{var}(e^*(t)) = \text{var}\left(\sum_{k=0}^{\infty} g_k v_e(t-k)\right) = \sigma_{v_e}^2 \sum_{k=0}^{\infty} g_k^2 \quad (15)$$

$e^*(t)$ の自己相関があまり大きくなければ g_k^2 は急速に減少するので、十分多い項数を取ることで右辺の無限級数 $\sum_{k=0}^{\infty} g_k^2$ は一定値に収束する。 $e(t)$ の分散が $\sigma_e^2(t)$ に等しくなるためには、式 (15) により $\gamma(t)$ を次のように定めればよい。

$$\gamma(t) = \sqrt{\frac{\sigma_e^2(t)}{\sigma_{v_e}^2 \sum_{k=0}^{\infty} g_k^2}} \quad (16)$$

モデルから計算した $e(t)$ の分散が、個別データから計算した $e(t)$ の分散 $\sigma_e^2(t)$ に、任意の時点 t で一致することは、以下のような簡単な計算によってわかる。

$$\begin{aligned} \text{var}(e(t)) &= \text{var}(\gamma(t)e^*(t)) = \gamma(t)^2 \text{var}(e^*(t)) \\ &= \frac{\sigma_e^2(t)}{\sigma_{v_e}^2 \sum_{k=0}^{\infty} g_k^2} \times \sigma_{v_e}^2 \sum_{k=0}^{\infty} g_k^2 = \sigma_e^2(t) \end{aligned}$$

上の計算で、 $\sigma_{v_e}^2$ は約分されて消えてしまうので、任意の値に取って良い。計算の便宜上 $\sigma_{v_e}^2 = 1$ と置くことにする。よって $\gamma(t)$ の推定値は以下ようになる。

$$\gamma(t) = \sqrt{\frac{\sigma_e^2(t)}{\sum_{k=0}^{\infty} g_k^2}} \quad (17)$$

II 状態空間モデルによるシグナル抽出

1 BLS のモデルの状態空間モデルによる表現

状態空間モデルは、以下のような 2 本の方程式から成る動的なシステムである。

$$y(t) = \mathbf{H}(t)\mathbf{a}(t) + w(t), \quad w(t) \sim N(0, \sigma_w^2) \quad : \text{観測方程式} \quad (18)$$

$$\mathbf{a}(t) = \mathbf{F}\mathbf{a}(t-1) + \mathbf{V}(t), \quad \mathbf{V}(t) \sim N(0, \mathbf{Q}) \quad : \text{遷移方程式} \quad (19)$$

$$\left[\begin{array}{l} \mathbf{F}: \text{遷移行列}、\mathbf{a}(t): \text{状態変数ベクトル}、\mathbf{H}(t): \text{観測行列}、\mathbf{G}: \text{駆動行列} \\ w(t): \text{観測ノイズ}、V(t): \text{システムノイズ} \end{array} \right]$$

初期状態 $\alpha(0)$ は正規分布に従うとする。 $\mathbf{F}$ 、 $\mathbf{G}$ 、 $\mathbf{H}(t)$ はすべて既知であるとする。 $\mathbf{Q}$ と σ_I^2 については、今のところは既知としておき、後で推定する。状態変数 $\alpha(t)$ は直接観測することが出来ず、式 (18) の観測方程式を通して観測ノイズ $w(t)$ が付加され、実際の観測値として観測される。また $\alpha(t)$ は、式 (19) の遷移方程式によって時間とともに変化していく。状態空間モデルによる分析には、次のような利点がある。

- ・ $\alpha(t)$ の動的な特性（遷移方程式）により、各状態変数が時間とともに緩やかに変化していく様子を上手く取り込んだ推定を行うことができる。
- ・ オンライン・リアルタイムの処理が可能なことである。つまり新しい観測値が観測されるたびに、その時点における最適な推定値を、効率よく求めることができる。

$\mathbf{F}$ 、 $\mathbf{G}$ 、 $\mathbf{H}(t)$ 、 $\mathbf{Q}$ を以下のように設定することにより、I 章で解説したモデルを式 (18)、(19) のような状態空間モデルの形で表すことができる。

$$\begin{aligned} \alpha(t) &= [\beta(t) \mid T(t) \mid S(t) \cdots S(t-10) \mid e_1(t) \cdots e_{13}(t)] \\ \mathbf{H}(t) &= [X(t) \mid 1 \mid 1 \mid 0 \cdots 0 \mid \gamma(t) \mid 0 \cdots 0] \end{aligned} \quad (20)$$

$$\mathbf{F} = \left[\begin{array}{c|cccc|cccc} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & -1 & -1 & \cdots & -1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & & \ddots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & & & 1 & 0 & 0 & 0 & 0 \end{array} \right], \quad \mathbf{G} = \left[\begin{array}{c|cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{array} \right], \quad \mathbf{Q} = \begin{bmatrix} \sigma_\beta^2 & 0 & 0 & 0 \\ 0 & \sigma_r^2 & 0 & 0 \\ 0 & 0 & \sigma_s^2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

このとき状態変数ベクトル $\alpha(t)$ の次元は 26 になる。 $\Theta = (\sigma_\beta^2 \quad \sigma_r^2 \quad \sigma_s^2 \quad \sigma_I^2)$ は、状態変数の変化の大きさを決めるパラメータであり、ハイパーパラメータと呼ばれる。

各時点における状態変数 $\alpha(t)$ を推定することができれば、 $\mathbf{H}(t)$ の代わりに次の $\mathbf{H}_\theta(t)$ を用いることで、観測値のデータ $y(t)$ から標本誤差 $e(t)$ を取り除くことが出来る。

$$\mathbf{H}_\theta(t) = [X(t) \mid 1 \mid 1 \mid 0 \cdots 0 \mid 0 \cdots 0] \quad (21)$$

$\alpha(t)$ の効率的な推定方法については、次節で述べる。

2 カルマンフィルタによる状態変数の推定

観測値から標本誤差を取り除くためには、モデルの選択、パラメータの推定及び時系列の分解のための計算を行う必要がある。状態空間モデルで表されたシステムに対しては、カルマンフィルタにより、これらの計算を効率良く行うことができる。カルマンフィルタは、観測方程式と遷移方程式とを結び付けて、各時点における状態変数の平均二乗誤差を最小にする推定値を、効率よく推定する一連の方程式のことである。入手できる観測値の程度により、状態変数の平均を推定する問題は、以下のように区別される（図 1 を参照のこと）。

- ・ 一期先予測・・・時点 $t-1$ までの観測値が手に入った場合
- ・ フィルタ・・・時点 t までの観測値が手に入った場合
- ・ 平滑化・・・時点 T ($T > t$) までの観測値が手に入った場合

遷移方程式と観測方程式は線形であり、また $\alpha(0)$ 、 $w(t)$ 、 $V(t)$ はすべて正規分布に従う。このことによって、観測値が与えられたときの状態変数の条件付分布は正規分布に従い、それは平均と分散により完全に決定される。よって状態変数の推定の際には、平均と分散の推定だけを考えればよい。カルマンフィルタを適用すると、状態変数の平均と分散の予測値及びフィルタの値を、逐次的に効率よく推定することができる。そしてそれらの値を用いた逆方向の逐次的計算により、平滑化の値を推定することが出来る。

各成分の確率分布の分散を既知とした場合、カルマンフィルタによる状態変数の平均と分散の予測値およびフィルタの値は、次のようになる。

$$\text{一期先予測} \quad \begin{cases} \alpha(t|t-1) = \mathbf{F}\alpha(t-1|t-1) \\ \mathbf{P}(t|t-1) = \mathbf{F}\mathbf{P}(t-1|t-1)\mathbf{F}' + \mathbf{G}\mathbf{Q}\mathbf{G}' \\ y(t|t-1) = \mathbf{H}\alpha(t|t-1) \\ \tilde{y}(t) = y(t|t) - y(t|t-1) \cdots (\text{一期先予測誤差}) \end{cases} \quad (22)$$

$$\text{フィルタ} \quad \begin{cases} f(t|t-1) = \mathbf{H}(t)\mathbf{P}(t|t-1)\mathbf{H}(t)' + \sigma_t^2 \cdots (\text{一期先予測誤差の分散}) \\ \mathbf{K}(t) = \mathbf{P}(t|t-1)\mathbf{H}(t)' / f(t|t-1) \cdots (\text{カルマンゲイン}) \\ \alpha(t|t) = \alpha(t|t-1) + \mathbf{K}(t)\tilde{y}(t) \\ \mathbf{P}(t|t) = (\mathbf{I} - \mathbf{K}(t)\mathbf{H}(t)')\mathbf{P}(t|t-1) \end{cases} \quad (23)$$

なおここでは、 t_1 期までの観測値が得られた場合の、 t_2 期における状態変数の平均及び分

散共分散行列の推定値を、それぞれ $\alpha(t_2|t_1)$ 及び $\mathbf{P}(t_2|t_1)$ で表している。式の導出等の詳細については、北川（1993）及び Harvey（1989）を参照のこと。

状態変数の平均のフィルタ $\alpha(t|t)$ は、一期前のデータに基づく予測 $\alpha(t|t-1)$ と、今期の観測値 $y(t)$ との加重和となっている。今期の観測データが観測されたとき、カルマンゲイン $\mathbf{K}(t)$ に基づいて一期先予測 $\alpha(t|t-1)$ を修正したものが、 $\alpha(t|t)$ であると解釈できる。ま

た状態変数の分散共分散行列のフィルタ $\mathbf{P}(t|t)$ は今期の観測値が観測されたことにより、右辺の $\mathbf{K}(t)\mathbf{H}(t)'$ の分だけ分散が改善された（小さくなった）と解釈できる。

時点 T までの観測値が手に入った場合、一期先予測とフィルタの結果を利用して、以下のよう後ろ向きの逐次的な計算により平滑化が行われる。

$$\text{平滑化} \quad \begin{cases} \mathbf{A}(t) = \mathbf{P}(t|t)\mathbf{F}'\mathbf{P}(t+1|t) \\ \boldsymbol{\alpha}(t|T) = \boldsymbol{\alpha}(t|t) + \mathbf{A}(t)(\boldsymbol{\alpha}(t+1|T) - \boldsymbol{\alpha}(t+1|t)) \\ \mathbf{P}(t|T) = \mathbf{P}(t|t) + \mathbf{A}(t)(\mathbf{P}(t+1|T) - \mathbf{P}(t+1|t))\mathbf{A}(t)' \end{cases} \quad (24)$$

3 ハイパーパラメータ及び初期状態の推定

これまでの議論は、ハイパーパラメータ $\Theta = (\sigma_\beta^2 \quad \sigma_\tau^2 \quad \sigma_s^2 \quad \sigma_i^2)$ が既知であるという仮定のもとで進められてきた。 Θ のとり方によって時系列の分解もかなり変わってくるので、観測値のデータから何らかの方法で合理的に Θ を決定する必要がある。そこで観測されたデータをもとにして対数尤度を計算し、最尤法によって Θ を推定する。一期先予測誤差 $\tilde{y}(t)$ は平均 0、分散 $f(t|t-1)$ の正規分布に従う。よって期間 $1 \leq t \leq T$ における観測値に対する対数尤度関数 $L(\Theta)$ は、次のように計算される (Harvey(1989))。

$$\begin{aligned} L(\Theta) &= \log \left[\prod_{t=1}^T \frac{1}{\sqrt{2\pi f(t|t-1)}} \exp \left(-\frac{\tilde{y}(t)^2}{f(t|t-1)} \right) \right] \\ &= -\frac{T}{2} \log 2\pi - \frac{T}{2} \sum_{t=1}^T \log(f(t|t-1)) - \frac{T}{2} \sum_{t=1}^T \frac{\tilde{y}(t)^2}{f(t|t-1)} \end{aligned} \quad (25)$$

式(26) の $L(\Theta)$ を最大化する Θ が、ハイパーパラメータの最尤推定量になる。

カルマンフィルタの公式と式 (25) を用いることで、様々な Θ に対する尤度関数を、非常に効率よく計算することができる。式(25) は複雑な非線形の関数であり、その最大値をとるパラメータを解析的に求めることは困難である。そこで数値計算による非線形最適化を行う。非線形最適化には様々な方法がある。今回の分析では、R の非線形最適化パッケージ `optim` を用いて、対数尤度関数の最大化を行っている。 Θ の各成分は分散なので、正の値である必要がある。そこで各成分を例えば $\sigma_\beta^2 = \exp(\tau)$ のように変換して、 τ の推定を行うこととする。

状態空間モデル及びカルマンフィルタを用いた分解を行うためには、状態変数の平均及び分散の初期値を与える必要がある。Tiller(1989)では、最初の数時点分の観測値から初期状態を作っている。状態変数ベクトルの次元が D の場合には、 $t \leq D$ における観測値に対してカルマンフィルタを適用し、 $\boldsymbol{\alpha}(D|D)$ 及び $\mathbf{P}(D|D)$ を状態変数の初期状態の平均及び分散とする。そして対数尤度は $t > D$ に対して計算する。このとき $\boldsymbol{\alpha}(0|0)$ 及び $\mathbf{P}(0|0)$ を定める必要がある。 $\mathbf{P}(0|0)$ を大きく取ることによって初期状態 $\boldsymbol{\alpha}(0|0)$ の影響を少なくすることができるものの、安定した初期状態を得るためには十分長い時間カルマンフィルタを作動させる必要がある。しかしもっともらしいと思われる $\boldsymbol{\alpha}(0|0)$ の値を決められる場合には、比較的短時間で初期状態を安定させることができる (柏木 (1997))。今回の分析では、 $\boldsymbol{\alpha}(0|0)$ 及び $\mathbf{P}(0|0)$ を次のように設定する。なお事前の情報が乏しい状態を表現するために、初期状態の分散を大きく設定している。

$$\begin{aligned}
 \mathbf{P}(0|0) &= 10000 * V_y I \\
 \boldsymbol{\alpha}(0|0) &= [\tilde{\beta}_1 | \tilde{\beta}_2 | 0 \dots 0] \\
 \left[\begin{array}{l} \mathbf{V}_y: \text{最初の 2 年分の } y(t) \text{ から求めた標本分散} \\ \mathbf{I}: \text{状態変数と同じ次元 (D) の単位行列} \\ \tilde{\beta}_1, \tilde{\beta}_2: \text{最初の 2 年分の } y(t) \text{ と } X(t) \text{ から求めた、回帰部分の定数及び係数} \end{array} \right]
 \end{aligned} \tag{26}$$

III モデルの診断及び推定値の精度の評価

1 自己相関、正規性、不均一分散の検定

基準化した一期先予測誤差 $\tilde{y}(t)$ を、以下のように定義する。

$$\tilde{y}(t) = \frac{y(t) - y(t|t-1)}{\sqrt{\mathbf{H}(t)\mathbf{P}(t|t-1)\mathbf{H}(t)' + \sigma_t^2}} \tag{27}$$

これはカルマンフィルタを用いた計算により、効率よく求めることが出来る。

仮定したモデルがデータに良く適合しており、かつハイパーパラメータ Θ が既知の場合には、 $\tilde{y}(t)$ は平均 0、分散 1 の互いに独立な正規分布に従うと考えられる。これを利用して、推定したモデルの診断を行うことができる。診断方法としては、自己相関、正規性、不均一分散の検定が考えられる (Durbin and Koopman (2001) を参照のこと)。

以下、各種の診断方法について解説する。なお式中の D は、状態変数ベクトルの次元を表しており、この期間の状態変数は初期状態を求めるのに使用しているので、診断の際には利用しないこととする (II 節の 3 を参照のこと)。

(1) 自己相関の検定 (Ljung-Box 統計量)

$\tilde{y}(t)$ の自己相関の検定には、以下の Ljung-Box 統計量を用いる (Durbin and Koopman (2001))。

$$L.B(m) = (n-D)(n-D+2) \sum_{l=1}^m \frac{\hat{r}_l^2}{n-D-l} \tag{28}$$

$$\left[\hat{r}_l^2 = \frac{\sum_{t=D+1+l}^T \tilde{y}(t)\tilde{y}(t-l)}{\sum_{t=D+1}^k \tilde{y}(t)^2} \right]$$

予測誤差が自己相関を持たないとき、上記の検定統計量は自由度 k の χ^2 分布に従う。5% で有意であれば、 $\tilde{y}(t)$ には自己相関があるものとみなす。今回の分析では、次数 k として 12 と 24 を選んでいる。

(2) 正規性の検定 (Jarque - Bera 統計量)

誤差の正規性の検定には以下の Jarque - Bera 統計量を用いる (Durbin and Koopman (2001))。

$$J.B = n \left(\frac{S^2}{6} + \frac{(K-3)^2}{24} \right) \quad (29)$$

$$\left[\begin{array}{l} S = \frac{\frac{1}{n} \sum_{t=D+1}^T \tilde{y}(t)^3}{\left[\sqrt{\frac{1}{n} \sum_{t=D+1}^T \tilde{y}(t)^2} \right]^3}, \quad K = \frac{\frac{1}{n} \sum_{t=D+1}^T \tilde{y}(t)^4}{\left[\frac{1}{n} \sum_{t=D+1}^T \tilde{y}(t)^2 \right]^2} \end{array} \right]$$

ここで S は標本歪度、 K は標本尖度である。予測誤差の分布が正規分布であれば、Jarque - Bera 統計量は自由度 2 の χ^2 分布に従う。5%で有意であれば、 $\tilde{y}(t)$ は正規分布に従っていないとみなす。

(3) 不均一分散の検定

F 検定統計量により、不均一分散の検定を行う (Durbin and Koopman (2001))。

$$F_{m|m} = \frac{\sum_{t=T-m+1}^T \tilde{y}(t)^2}{\sum_{t=D+1}^m \tilde{y}(t)^2} \quad (30)$$

$$\left[m = \frac{T-D}{3} \right]$$

$\tilde{y}(t)$ の最初の 1/3 と最後の 1/3 とを用いて、分散不均一性の検定を行う。一期先予測誤差の分散が均一ならば、検定統計量 H は自由度 (m, m) の F 分布に従うと考えられる。そこでこれを利用して検定を行う。5%で有意であれば、 $\tilde{y}(t)$ の分散は不均一であるとみなす。

2 平滑化後の分散による精度の評価

平滑化により推定したシグナル $\theta(t)$ の分散は、以下のようにして推定することができる。

$$\begin{aligned} \text{var}(\theta(t)) &= \mathbf{H}_\theta(t) \mathbf{P}(t|t) \mathbf{H}_\theta(t) + \sigma_f^2 \\ \left[\mathbf{H}_\theta(t) &= [X(t) | 1 | 1 \ 0 \ \dots \ 0 | 0 \ \dots \ 0] \right] \end{aligned} \quad (31)$$

ここで $\mathbf{H}_\theta(t)$ は、シグナルを抽出するための観測行列である (II 節 1 参照)。

式 (28) から以下のようにして $\theta(t)$ の誤差率 $CV_\theta(t)$ を求めることができる。

$$CV_\theta(t) = \sqrt{\text{var}(\theta(t))} / \theta(t) = \frac{\sqrt{H_\theta(t)P(t|t)H_\theta(t) + \sigma_t^2}}{\theta(t)} \quad (32)$$

これを観測値の誤差率と比較することで、精度の向上について調べることができる。

3 ベンチマーキング

BLS では、時系列回帰モデルによって求めた推定値の精度を更に高めるために、ベンチマーキングと呼ばれる手法を適用している。これはモデルによる推定値の年平均と、もとの観測値の年平均とを一致させるように調整する方法である。観測値を月次で見た場合には、標本数が少なく不安定である。しかし 1 年間分の観測値のデータをプールすることで推定の精度が増し、より正確な推定値になると考えられる。よって観測値の年平均にモデルの年平均を合わせることで、精度の向上が期待される。Tiller (1989) では、Denton-Method (Denton (1971)) と呼ばれる、代数的な操作による方法を紹介している。今回の分析では、この方法によってベンチマーキングを行う。

モデルにより推定された系列 $\mathbf{z} = (z_1, z_2, \dots, z_{12k})^t$ の変動となるべく近い動きをするような時系列で、その年平均が観測値 $\mathbf{y} = (y_1, y_2, \dots, y_k)^t$ の年平均と近くなるような時系列 $\mathbf{x} = (x_1, x_2, \dots, x_{12k})^t$ を推定する問題として、ベンチマーキングをとらえることができる。

$\mathbf{z}$ となるべく近い動きをする時系列 $\mathbf{x}$ をどのように数学的に定義するかが問題である。Denton (1971) ではこれを、階差同士の差の 2 乗和が最小になる時系列として、次のように定義している。

$$\sum_{i=1}^{12k} (\Delta x_i - \Delta z_i)^2 (= \sum_{i=1}^{12k} (\Delta(x_i - z_i))^2) \text{ を最小にする } \mathbf{x} \quad (33)$$

このとき上記のベンチマーキングの問題を、次のような線形制約付きの 2 次形式の最小化問題に置き換えることができる。

$$\text{線形制約: } \frac{1}{12} \sum_{i=12T-11}^{12T} x_i = \frac{1}{12} \sum_{i=12T-11}^{12T} 12y_i \quad (\Leftrightarrow \sum_{i=12T-11}^{12T} x_i = \sum_{i=12T-11}^{12T} 12y_i) \quad (1 \leq T \leq k) \quad (34)$$

$$\text{目的関数: } \mathbf{u} = (\mathbf{x} - \mathbf{z})^t \mathbf{D}^t \mathbf{D} (\mathbf{x} - \mathbf{z}) - 2\lambda^t (\mathbf{y} - \mathbf{B}^t \mathbf{x})$$

$$\left[\begin{array}{l} z_i : \text{最初から数えて } i \text{ 月目の時系列モデルによる推定値。} (1 \leq i \leq 12k) \\ x_i : \text{最初から数えて } i \text{ 月目のベンチマーキング後の値。} (1 \leq i \leq 12k) \\ \lambda = (\lambda_1, \lambda_2, \dots, \lambda_{12k})^t : \text{Lagrange の未定乗数} \end{array} \right]$$

$$\left[\begin{array}{l} \mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \mathbf{j} & \mathbf{o} & \cdots & \mathbf{o} \\ \mathbf{o} & \mathbf{j} & \cdots & \mathbf{o} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{o} & \mathbf{o} & \cdots & \mathbf{j} \end{bmatrix} \\ \mathbf{j} = (1, 1, \dots, 1) : 1 \times 12 \text{ 列ベクトル} \text{ , } \mathbf{o} = (0, 0, \dots, 0) : 1 \times 12 \text{ 列ベクトル} \end{array} \right]$$

式 (34) の解は Lagrange の未定乗数法により求めることができ、それは次の式で表される。

$$\begin{bmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{B}' & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{z} \\ \mathbf{y} - \mathbf{B}'\mathbf{z} \end{bmatrix} \quad (35)$$

$$\left[\begin{array}{l} \mathbf{A} = \mathbf{D}'\mathbf{D} \end{array} \right]$$

よってベンチマーキング後の値は、部分行列の逆行列に関する公式を式 (35) に適用することにより、以下のように推定出来る。

$$\mathbf{x} = \mathbf{z} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1}(\mathbf{y} - \mathbf{B}'\mathbf{z}) \quad (36)$$

IV 労働力調査への適用

1 推定方法及び結果の比較方法

BLS の小地域推定の手法を労働力調査へ適用し、都道府県別の完全失業率を推定する。推定の枠組みは以下のとおりである。

- ・平成 3 年 1 月～平成 15 年 12 月の観測値のデータを用いて推定を行う。
- ・都道府県別の標本誤差 ψ_i には、第 4 章の方法で個票データから推定したものをを用いる。
- ・最尤法で推定する際の非線形最適化では、 $\hat{\sigma}_v^2 = \exp(\tau)$ と変換して、BFGS 公式による準ニュートン法を用いた推定を行う（I 節 3 参照）。このとき、まず Nelder-Mead 法によって大まかな推定値を求め、それを初期値として更に準ニュートン法を適用する。この一連の計算は R の非線形最適化パッケージ optim を用いて行う。
- ・回帰の部分の補助情報については、職業安定業務統計における都道府県別月次の有効求人倍率を用いる。
- ・推定した $\theta(t)$ に対して Denton Method によるベンチマーキングを適用し、調整を行う。

- ・推定した結果に対して各種検定統計を用いた分析などを行い、結果を考察する。

なお、今回の推定に用いたプログラムについては、すべてこの章の最後に示してある。

2 推定対象となる小地域の選定

すべての都道府県について分析を行うことは困難であるので、推定対象となる小地域を絞ることとする。時系列は標本誤差に大きく影響されるものと考えられる。また標本誤差は標本数（標本調査区数）に影響を受ける。よって平成 12 年～平成 14 年の年平均標本調査区数（表 1 参照）に応じて、推定を行う県を選定する。

標本調査区数を都道府県別に見たものが図 2-1、2-2 である。かなり偏った分布をしていることが分かる。この中から標本数に応じて、以下の八つの小地域を選定する。

- ・調査区数が少ないところの代表として、山梨県、鳥取県、佐賀県を選定する。
- ・調査区数が多いところの代表として、北海道、東京都、大阪府、沖縄県を選定する。
- ・ヒストグラムによれば、標本数が 30、40、50 程度の調査区数のところが多くなっている。よってこれらの代表としてそれぞれ、香川県、岐阜県、宮城県を選定する。

これらの各県について推定を行い、結果を考察する。

3 完全失業率と補助情報との相関関係

完全失業率と関係が深いと思われる補助情報で、都道府県別に月次で入手できる統計のうち、主なものとして、有効求人倍率が考えられる。これは職業安定事務所における職業紹介業務をもとにした業務統計であり、以下の式で定義される。

$$\text{有効求人倍率} = 100 * \frac{\text{月間有効求人数}}{\text{月間有効求職者数}} \quad (37)$$

月間有効求職者数とは、前月から繰り越された有効求職者の数に、当月の新規求職申し込み件数を加えたものである。また月間有効求人数とは、前月から繰り越された有効求人数に、当月の新規求人数を加えたものである。有効求人倍率の増加は、完全失業率の低下にある程度つながるものと考えられる。よって有効求人倍率は、完全失業率と負の相関があると考えられる。

まず小地域の完全失業率の分析に先立って、各説明変数と完全失業率との相関についてまとめる。前節で選んだ八つの小地域について、完全失業率と補助情報の関係を時系列グラフおよび散布図にまとめたものが、図 3 である。これにより、以下のことが分かる。

- ・時系列のグラフを見ると、有効求人倍率はほぼ、完全失業率と逆の動きをしているように見える。散布図でも負の相関が認められる。ただし決定係数は全体的に低く、高いところでも 0.5 を超すところはない。

- ・平成5年以前と以降で、両者の関係が変化していると思われる地域がいくつかある。多くの小地域では、平成5年以前の急激な落ち込みに対して完全失業率の上昇は緩やかである。しかしその後の完全失業率の大幅な上昇と比較して、有効求人倍率の変動は幾分緩やかである。完全失業率は遅行指標であり、有効求人倍率に比べて景気への反応が若干遅れることが原因ではないかと推測される。これに伴う相関関係の変化も、状態空間モデルであればある程度とらえることができると考えられる。
- ・北海道では補助情報の季節性がかなり高い。季節労働者の多いことが示唆される。しかも観測値との間に、横方向への若干のずれが見られる。北海道では補助情報の強い季節性のために、回帰部分がかなりの影響を受けると予想される。

4 結果の考察

(1) 時系列のグラフによる比較

初期状態が安定したと考えられる平成5年1月から平成15年12月にかけて、各項の推定値および誤差率を時系列のグラフに表したものが図4-1～4-8である。これらの結果から、以下のことが分かる。

- ・ほとんどすべての小地域で、標本誤差が比較的良く取り除かれているように見える。
- ・岐阜県、鳥取県、佐賀県では、ベンチマーキングによる改定が大きかった。佐賀県については、平成10年における外れ値の影響が、ベンチマーキング後もなお残っているように感じられる。
- ・佐賀県の外れ値と思われる時点（平成6年、平成10年）では、モデルによる推定値の誤差率が、観測値の誤差率を上回ってしまっている。それ以外のほとんどの小地域では、観測値よりもモデルの誤差率の方が、かなり低い値となっている。しかもモデルによる推定値の誤差率は、時間にかかわらずほぼ一定であり、推定値はかなり安定していると考えられる。
- ・ほとんどの小地域について、回帰項とトレンド項により、大まかな傾向が上手く取り出されている。ただし北海道については、回帰項の部分に補助情報の季節性がかなりの程度現れており、しかもその変動が観測値の季節性と微妙にずれている。しかしこのずれは、季節項によって上手く取り出されているようである。
- ・季節項の動きは、すべての小地域で非常に安定したものとなっている。

(2) 各種統計量による比較

推定した値に対する各種統計量については、表2にまとめてある。これにより、以下のことが分かる。

- ・各種検定統計量を用いた診断を見ても、診断をパスしなかったのは、以下の小地域である。

正規性の検定をパスしなかった小地域：宮城県、佐賀県

不均一分散の検定をパスしなかった小地域：岐阜県、鳥取県

これ以外のほぼすべての地域でほとんどの診断をパスしており、その結果も良いようである。

- ・ハイパーパラメータを見ると、どの小地域でも季節項のパラメータがかなり低く、安定した季節性となっている。このことはグラフからも読みとれる。それに対してトレンド項のパラメータは高い値となっている。完全失業率は推定期間中には上昇基調であり、この動きを補助情報で説明できていないことが考えられる。佐賀県で不規則変動項のパラメータが大きな値を示している。これは外れ値の影響であると考えられる。外れ値が回帰項、トレンド項、季節項で説明できず、不規則変動項に入ったためと思われる。それ以外の地域では不規則変動項のパラメータは低くなっており、観測値のデータが各項によって、よく説明されていることが分かる。

(3) まとめ

鳥取県や佐賀県などの一部の小地域で不安定な変動をする時系列が若干あるものの、ほぼすべての地域でシグナルが安定して取り出されているように見える。グラフで見る限りでは推定が上手く行われていると思われる。外れ値については、何らかの対策が必要である。例えば外れ値をそのまわりの時期の移動平均で置き換えるといった処置を行うことで、外れ値の影響を避けた、よりロバストな推定を行うことができると考えられる。

V 今後の課題

1 地域の値を用いた回帰モデル

今回の分析では、有効求人倍率を補助情報とした時系列回帰モデルを用いて、都道府県別月次の完全失業率の推計を行った。これに対して、近隣の小地域の完全失業率を補助情報として用いることも考えられる。近隣の小地域同士のトレンドは連動していると考えられる。この場合には、近隣の定義が問題になる。労働力調査の 10 地域区分を便宜的に利用することが考えられる。

2 外れ値の処理

外れ値が原因で、推定値が不安定であるような小地域があった。外れ値の処理を行うことで、推定値が安定すると考えられる。外れ値の処理については、以下のような方法が考えられる。

- ・移動平均や前年同月値で外れ値を補正する。
- ・外れ値を欠測値とみなして、一期先予測により補正する。

U.S.Bureau of Census (1996) では、一期先予測誤差の絶対値が個別データによる標準誤差の推定値の 3 倍を超えた時点を外れ値としている (3 σ ルール)。

3 非線形、非正規モデルによる推定

今回は、線形な状態空間モデルを用いて分析を行った。またすべての確率変数の分布に正規分布を仮定していた。しかしこれらの仮定が成り立たないと考えられる場合も多くありうる。このような場合に非線形モデルや非正規モデルを用いることで、より精度の高い推定を行うことが出来ると考えられる。

非線形、非正規モデルを用いた分析については様々な研究があり、その一部は Durbin and Koopman(2001)に解説されている。この場合推定には、対数尤度を解析的に評価するのではなく、リサンプリングなどの手法を用いた、シミュレーションによる方法が用いられる。この手法により、非線形なシステムを直接分析することができる。また外れ値及びレベルシフトの処理もカルマンフィルタの中で自動的に行うことが出来るので、よりロバストな推定を行うことが出来ると期待される。ただし計算機にかかる負担が大きくなるので、計算にははるかに多くの時間がかかるものと予想される。リサンプリングを多用するために、ループやソートの利用も多くなる。よってこの場合には今回のように R でプログラムを組むのではなく、FORTRAN のようなコンパイラを用いた方が、効率が良いと考えられる。

VI 図表

図 1 予測、フィルタ、平滑化の概念

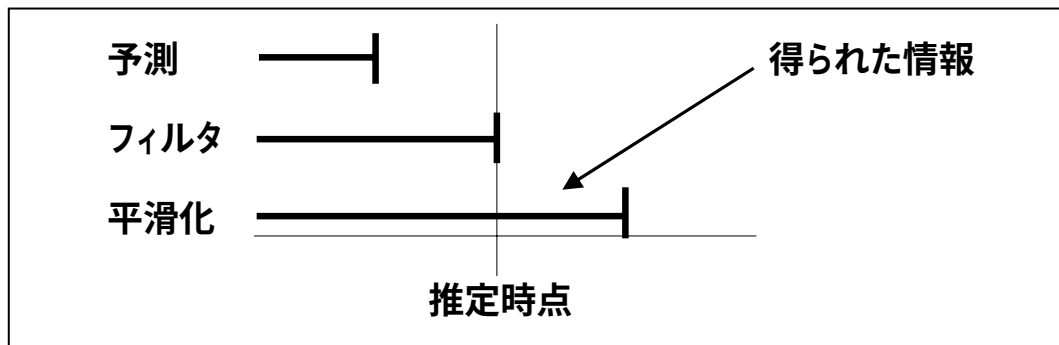


図 2-1 都道府県別標本調査区数（昇順）

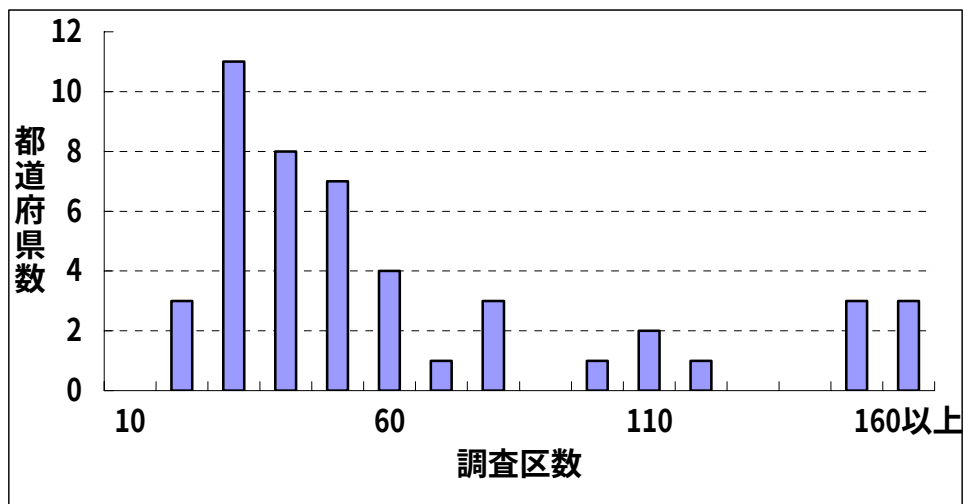


図 2-2 都道府県別標本調査区数（ヒストグラム）

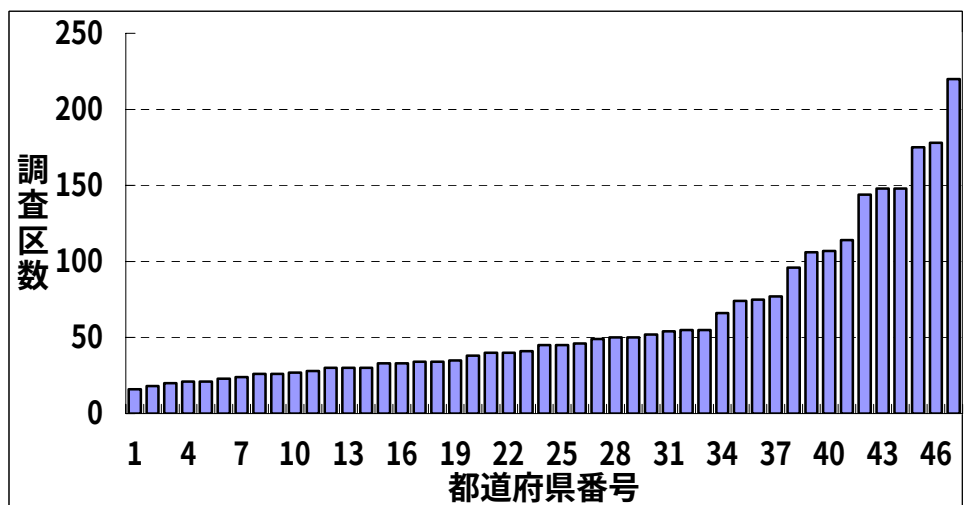
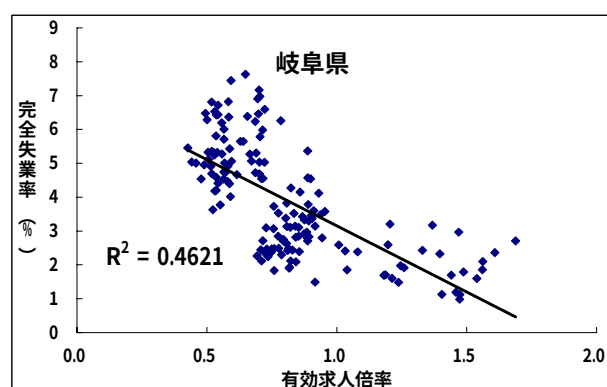
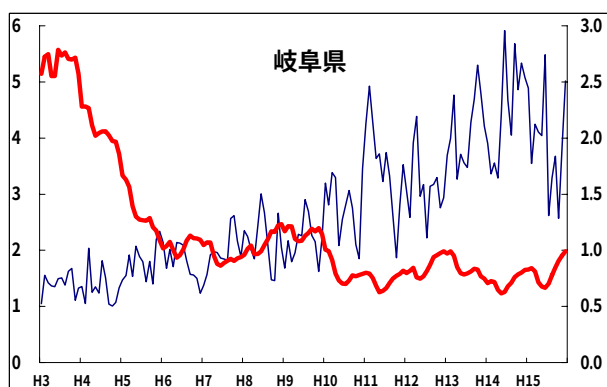
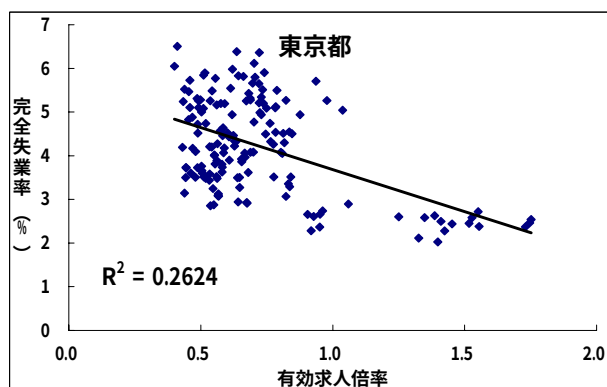
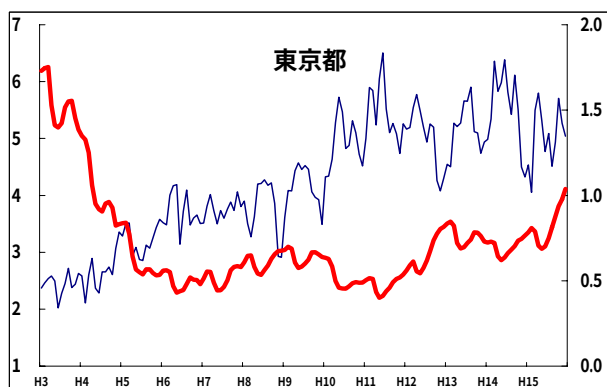
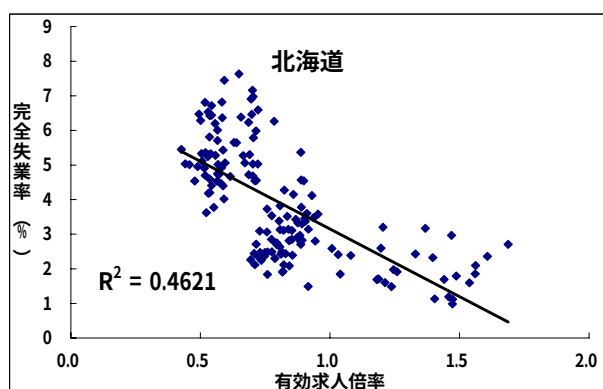
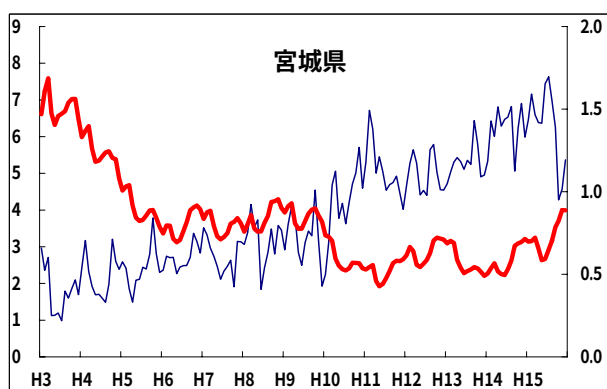
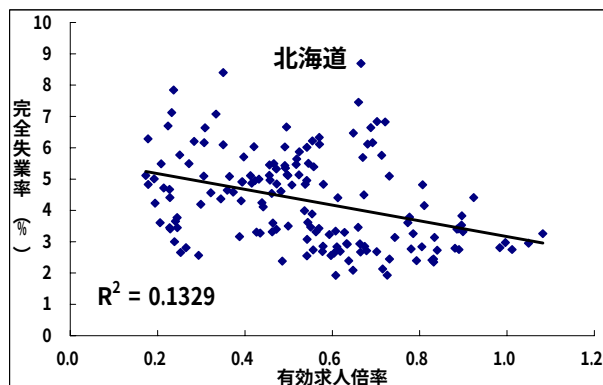
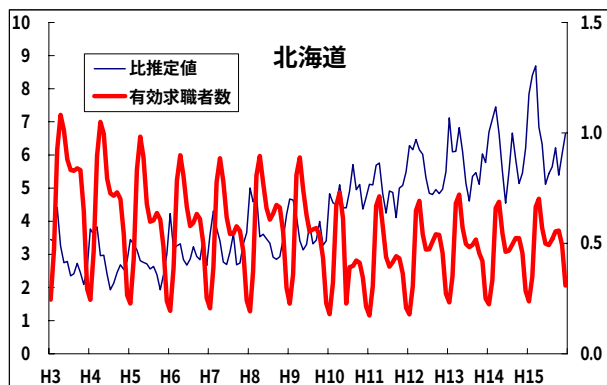


図3 完全失業率と有効求人倍率の相関関係



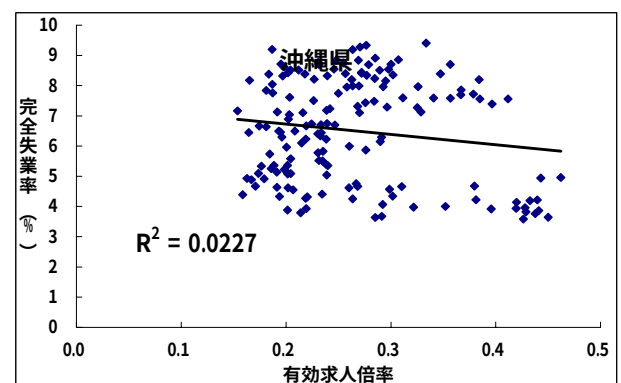
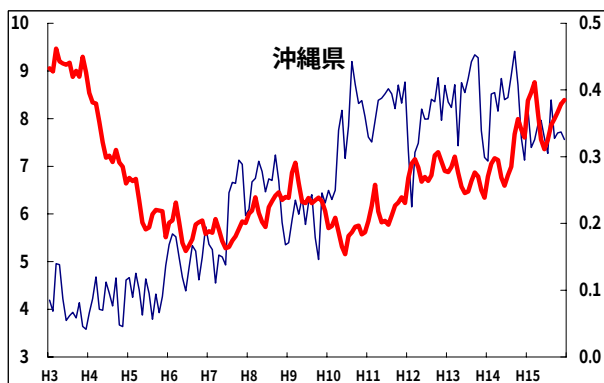
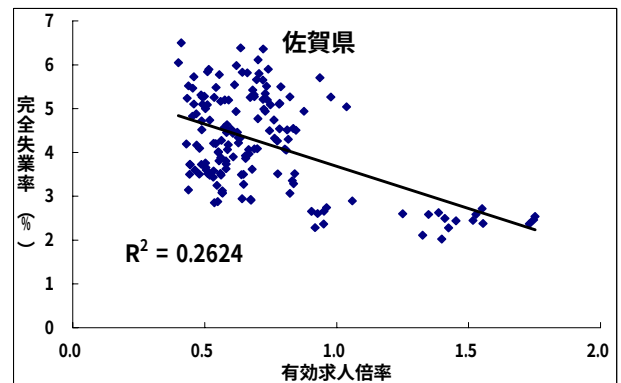
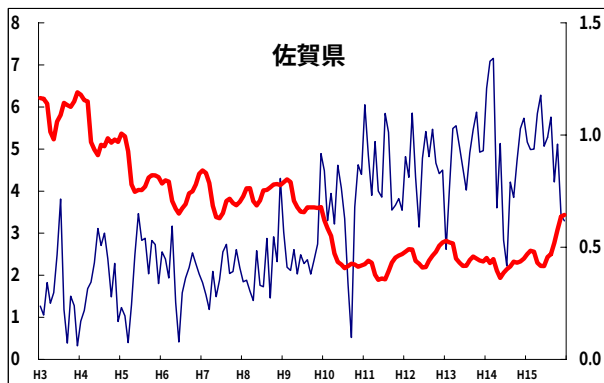
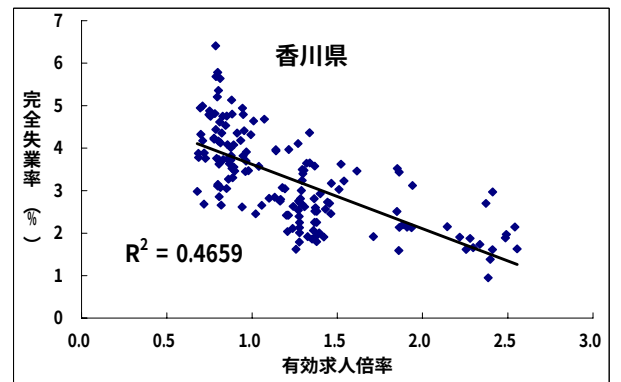
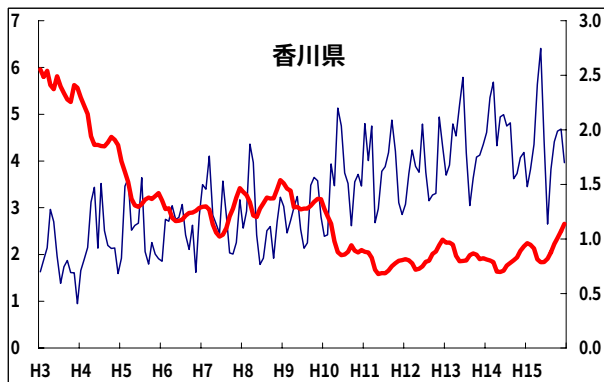
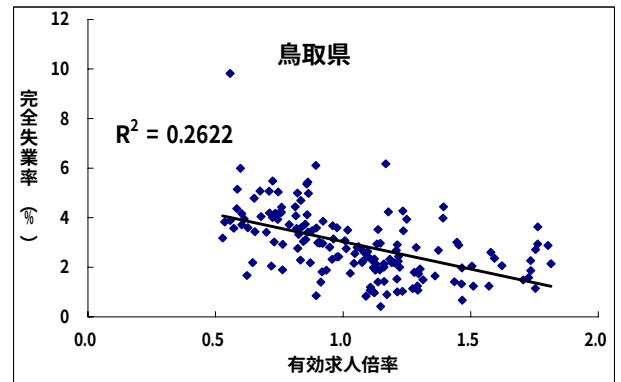
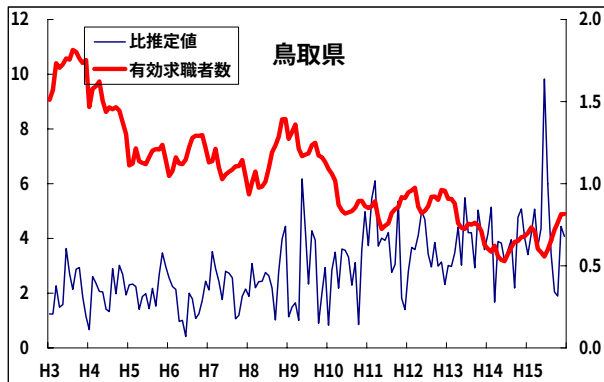


図 4-1 北海道推定結果のグラフ

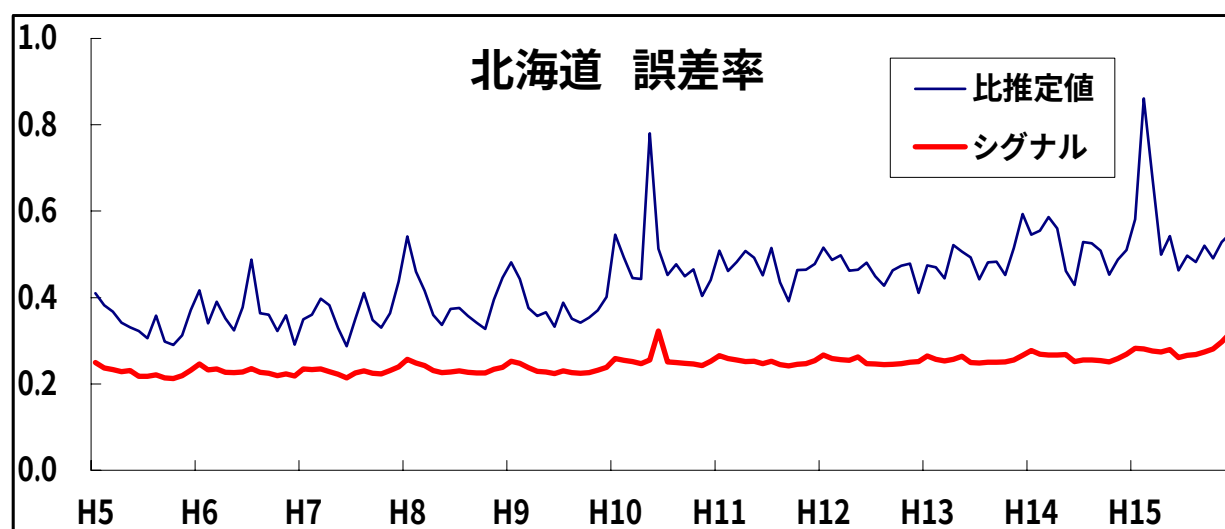
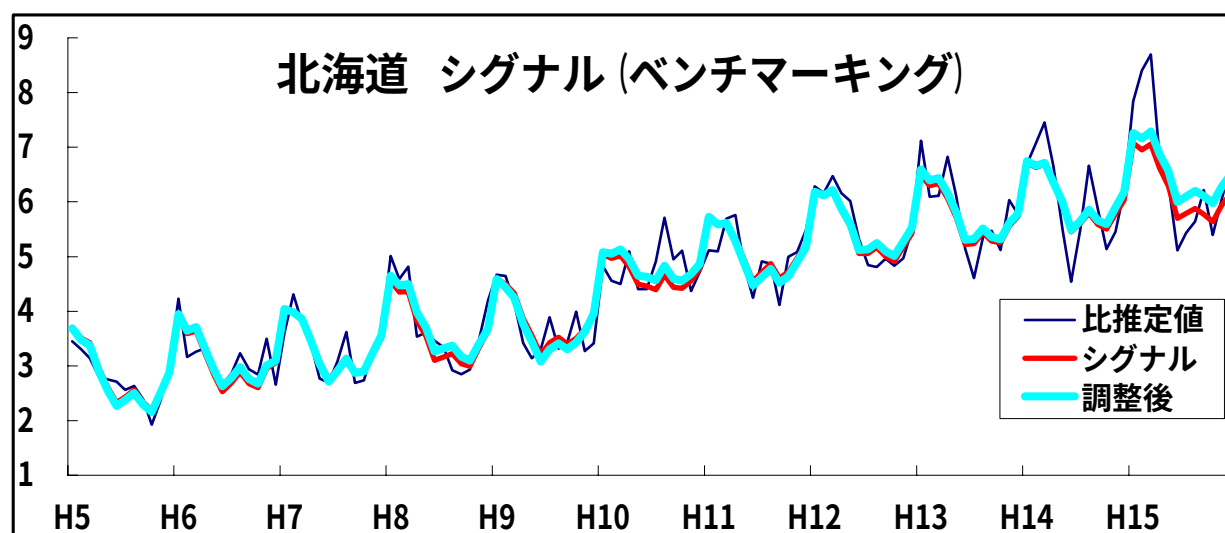
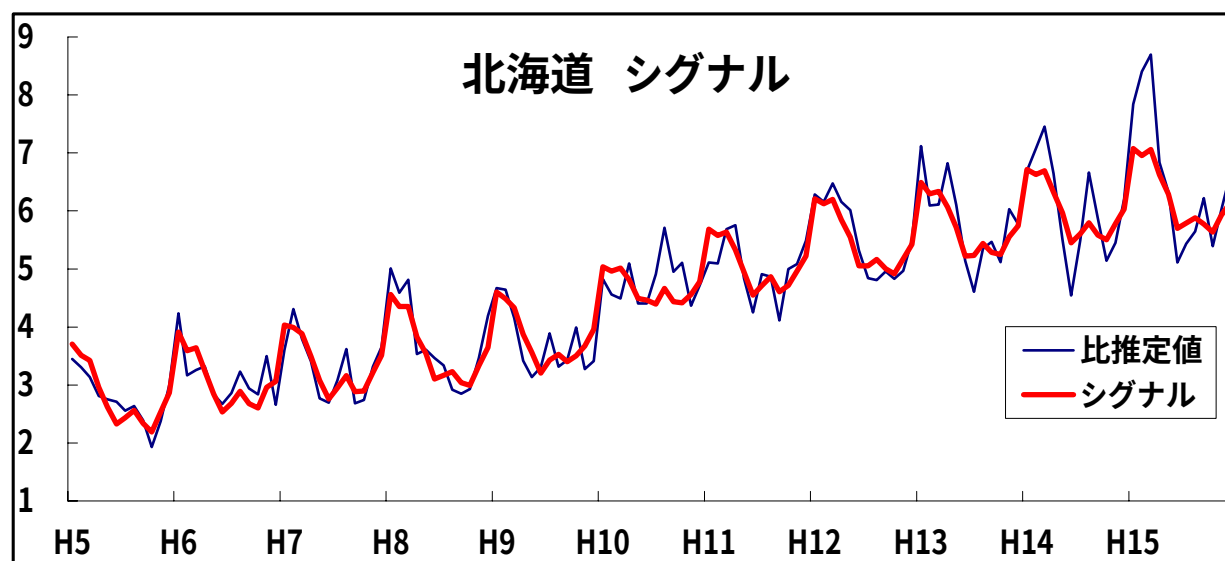


図 4-1 北海道推定結果のグラフ (続き)

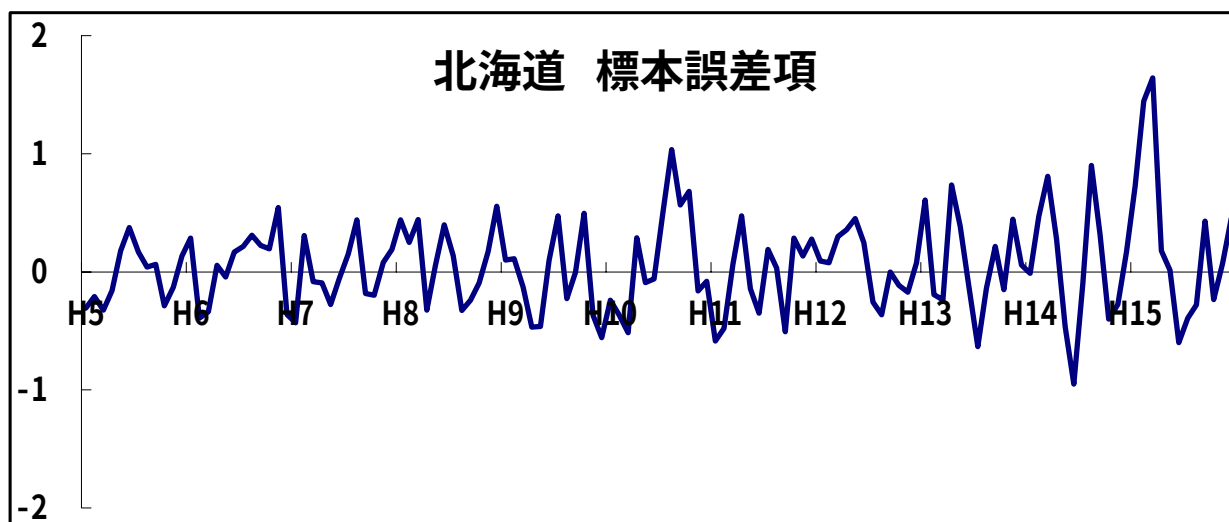
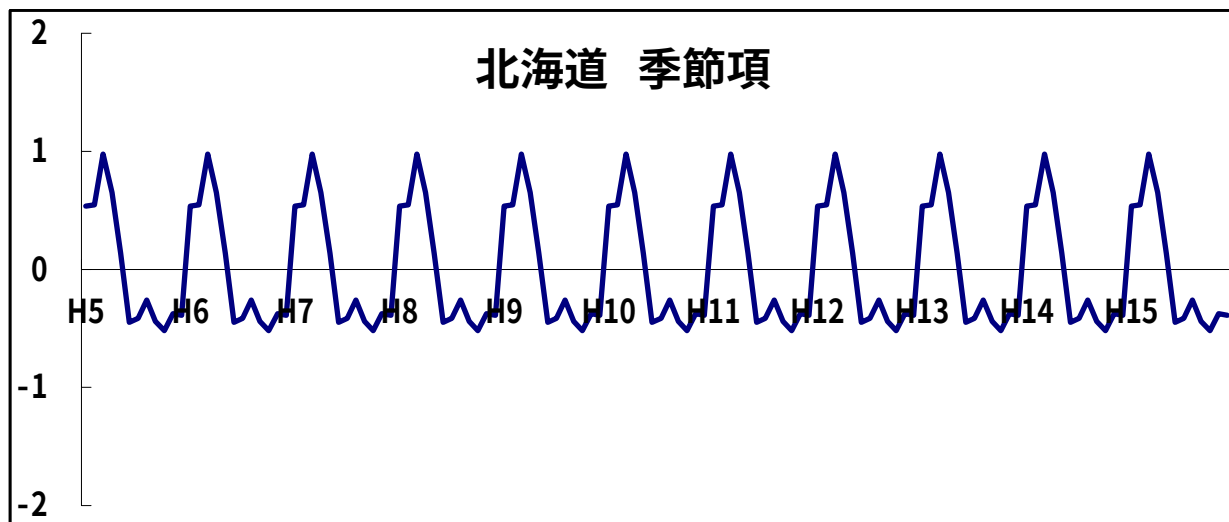
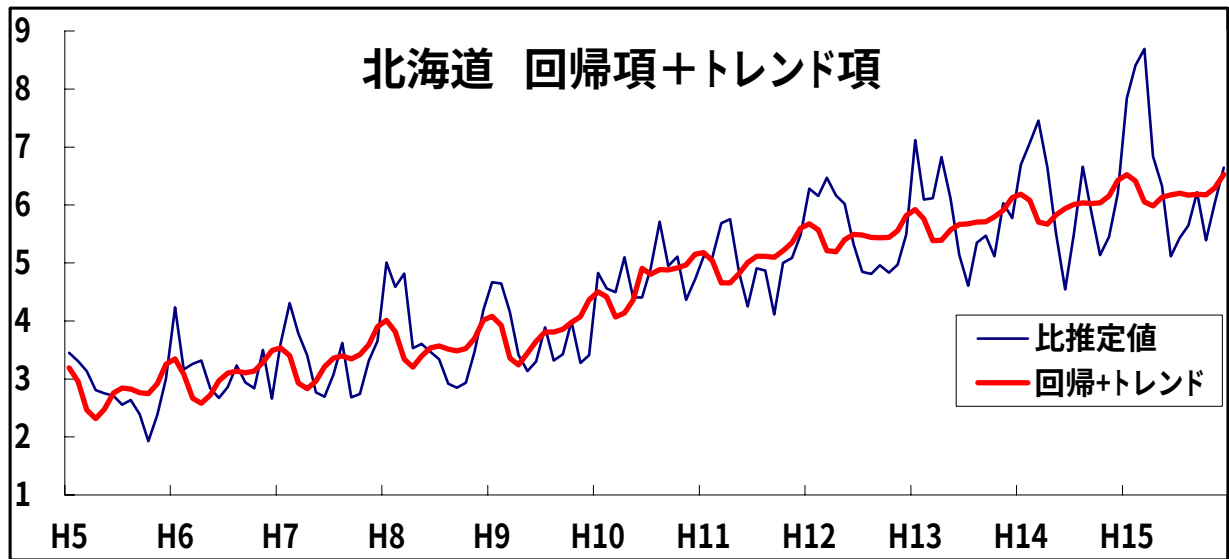


図 4-2 北海道推定結果のグラフ

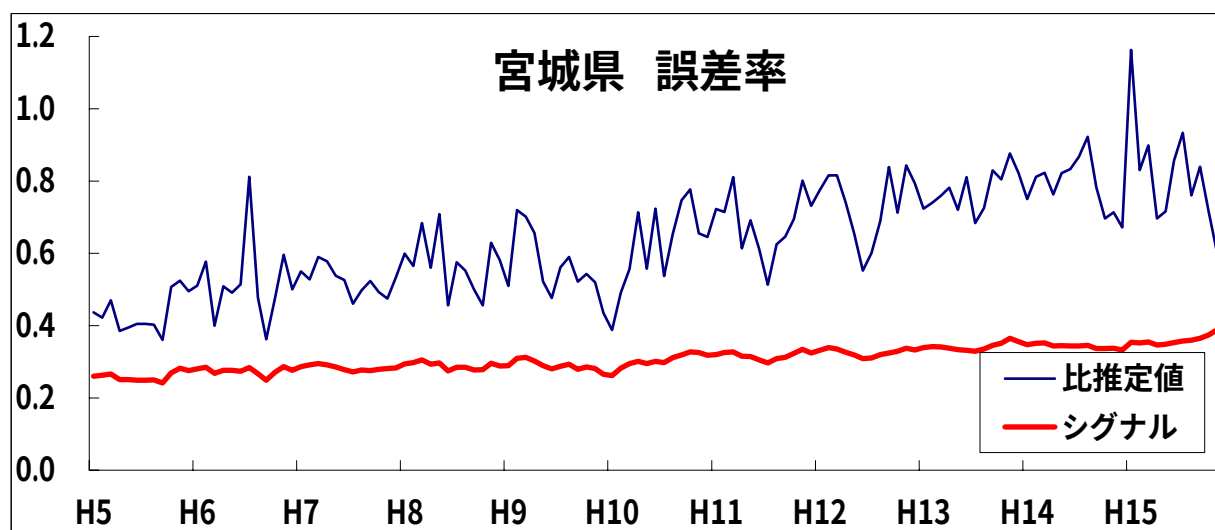
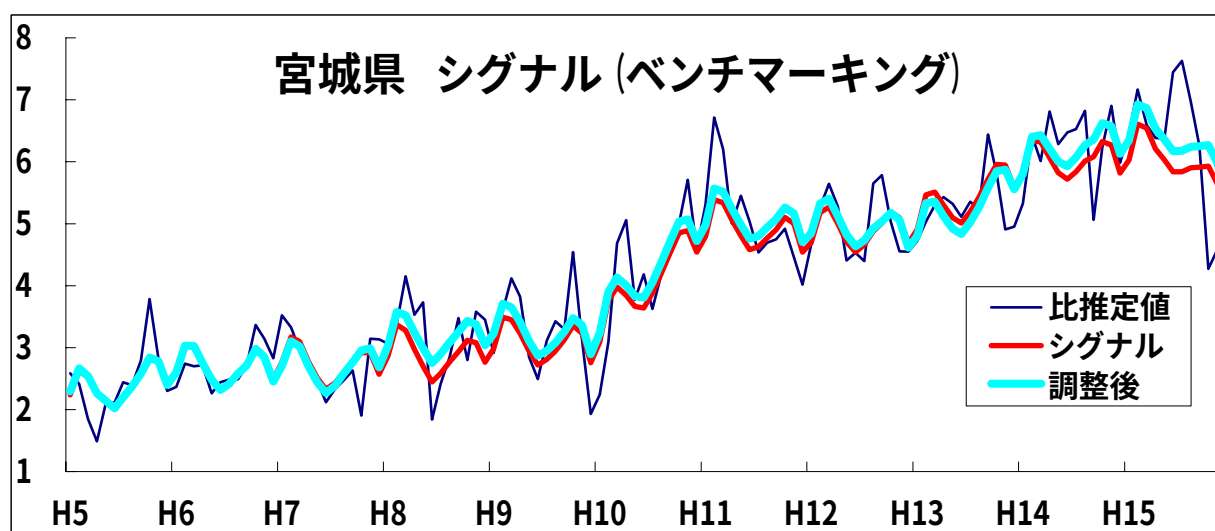
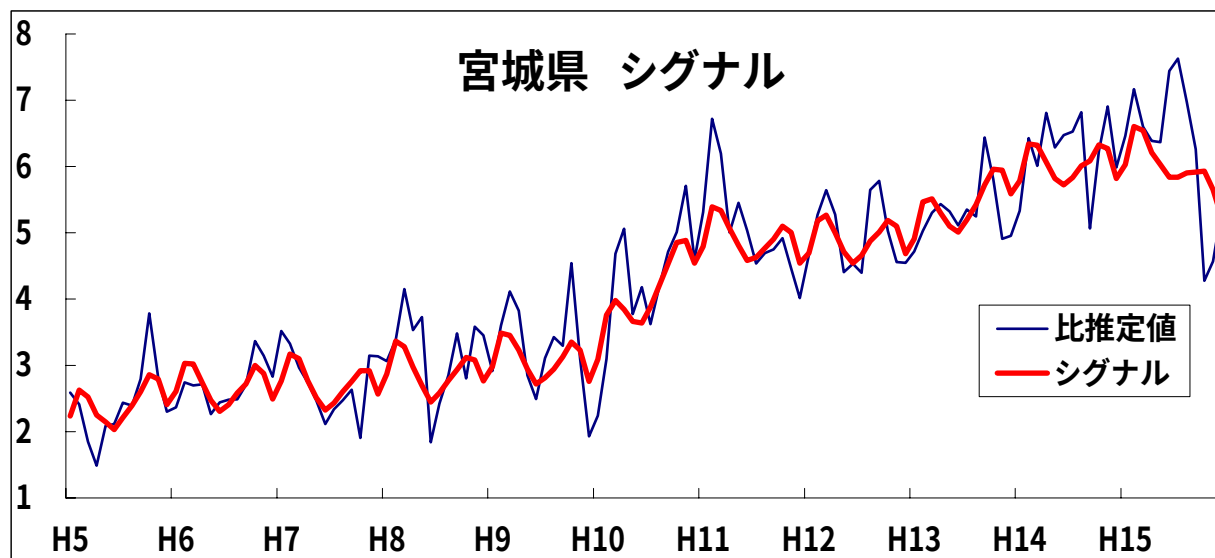


図 4-2 北海道推定結果のグラフ (続き)

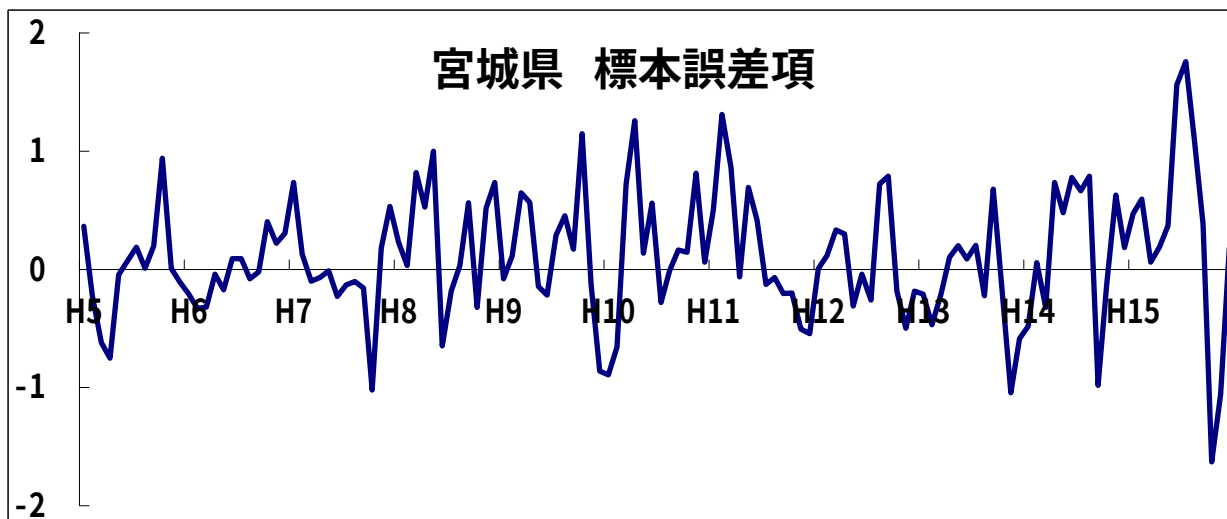
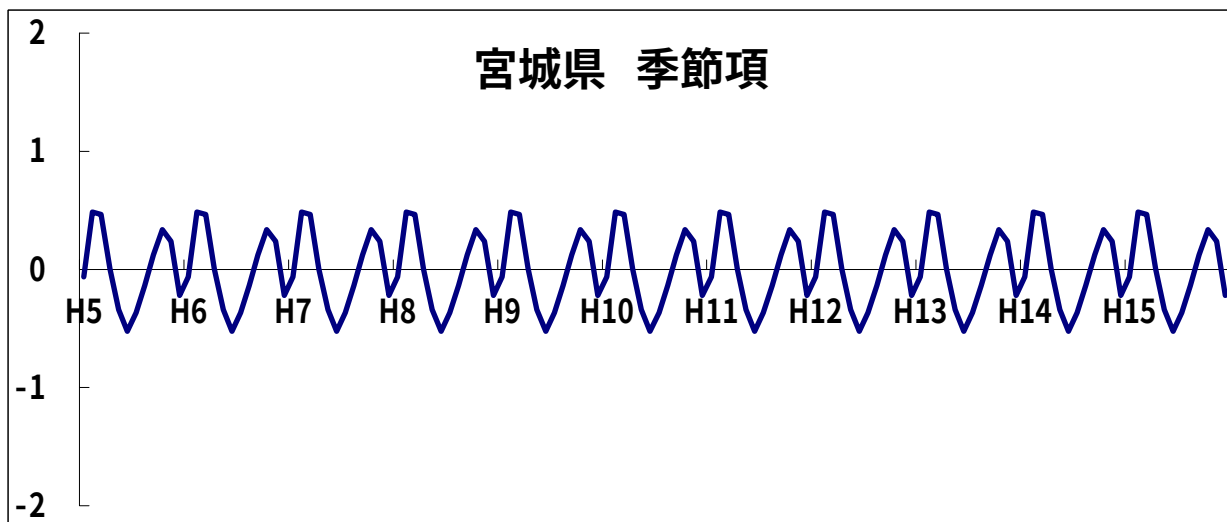
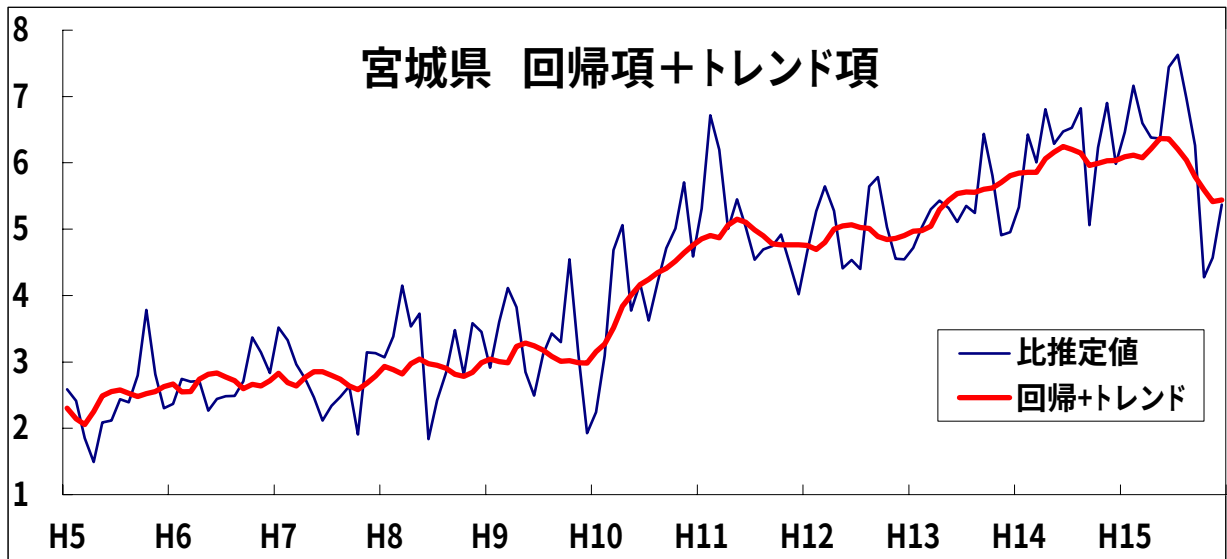


図 4-3 東京都推定結果のグラフ

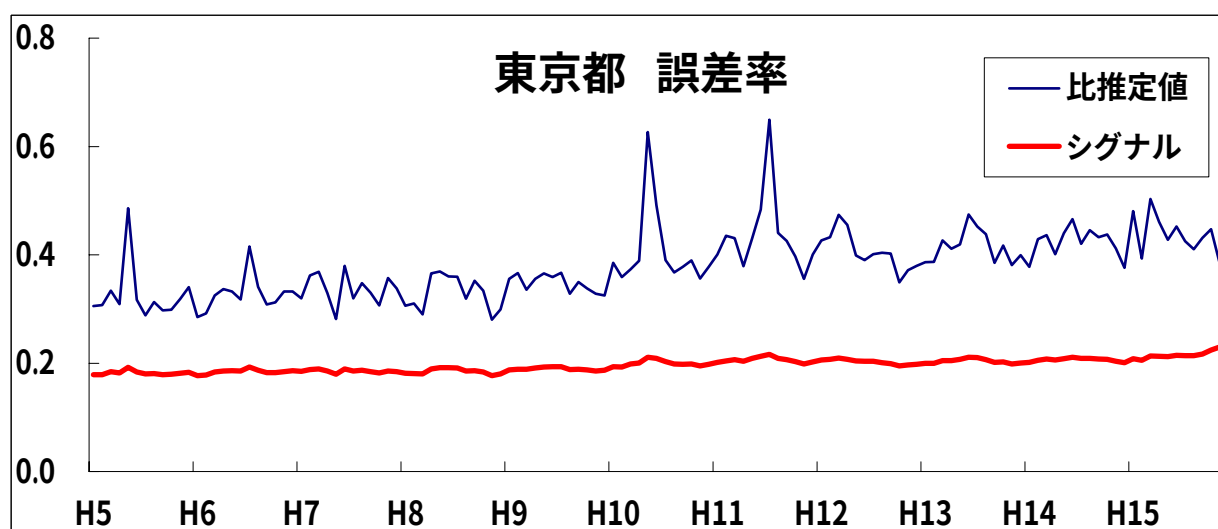
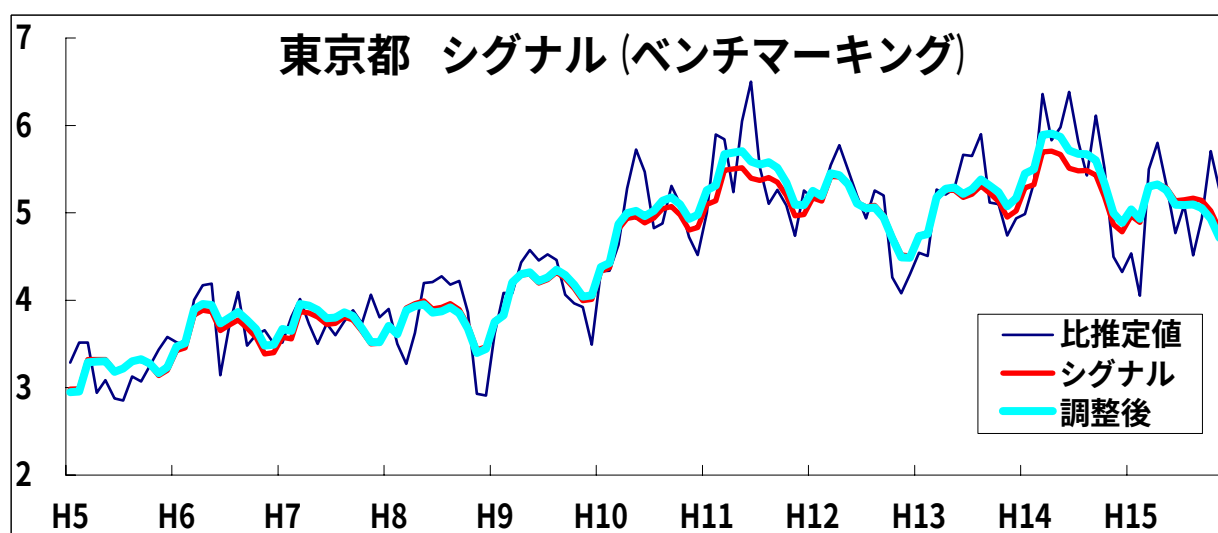
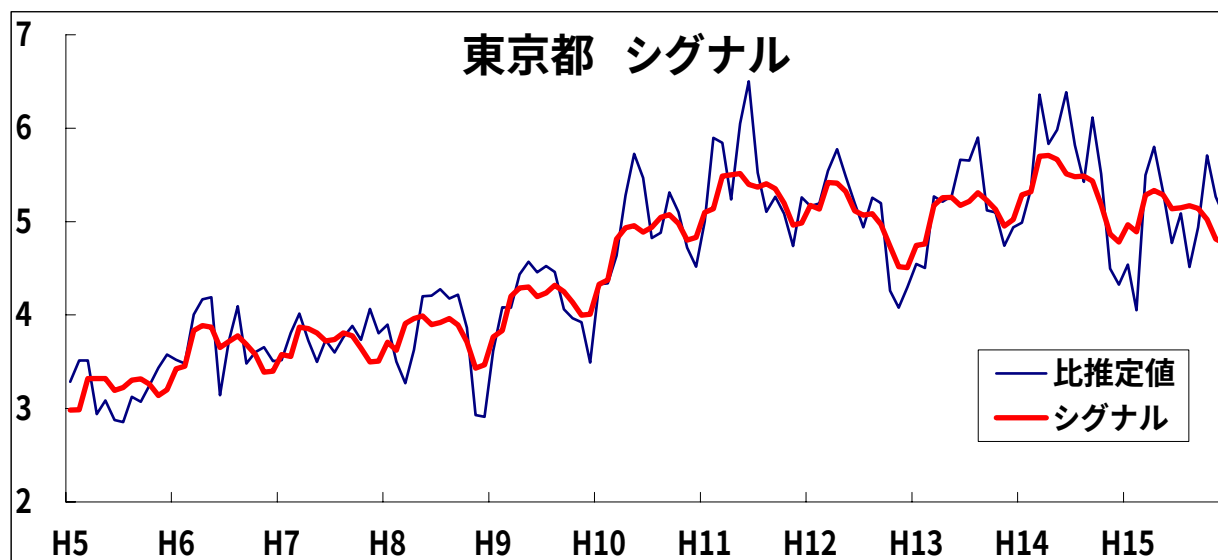


図 4-3 東京都推定結果のグラフ (続き)

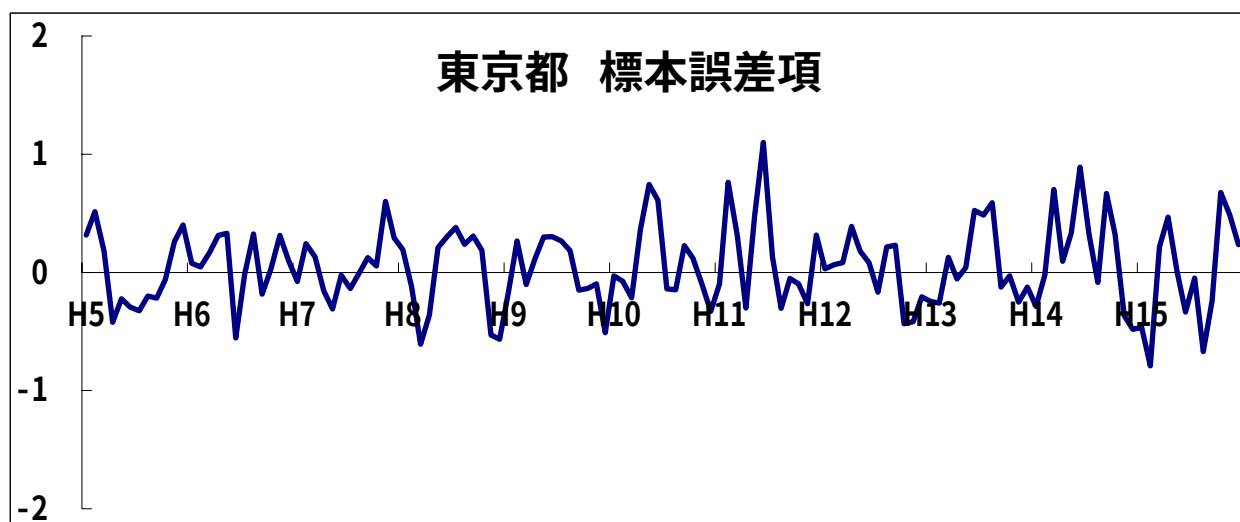
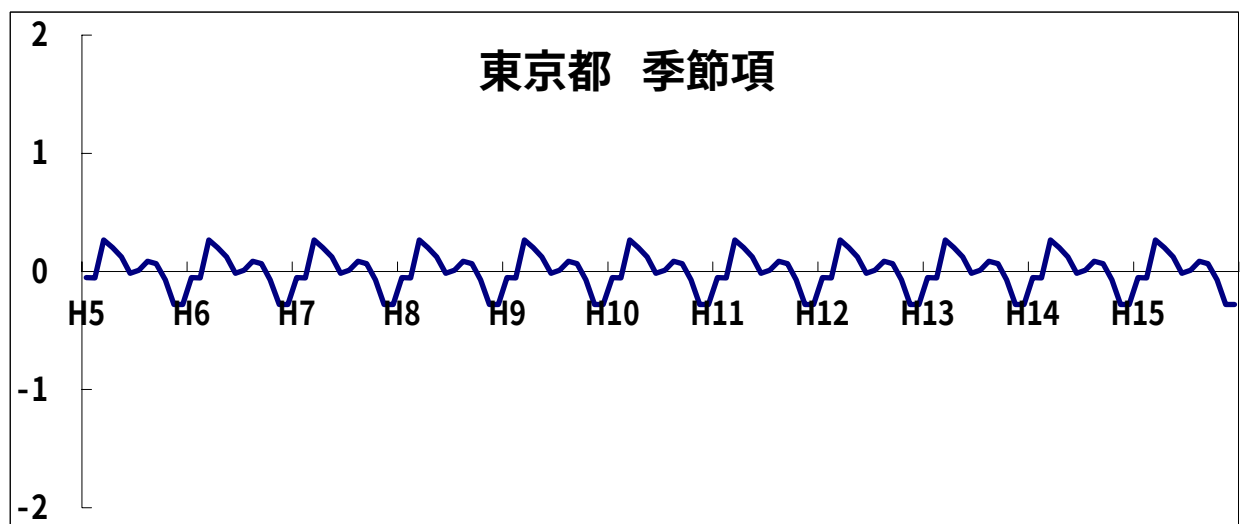
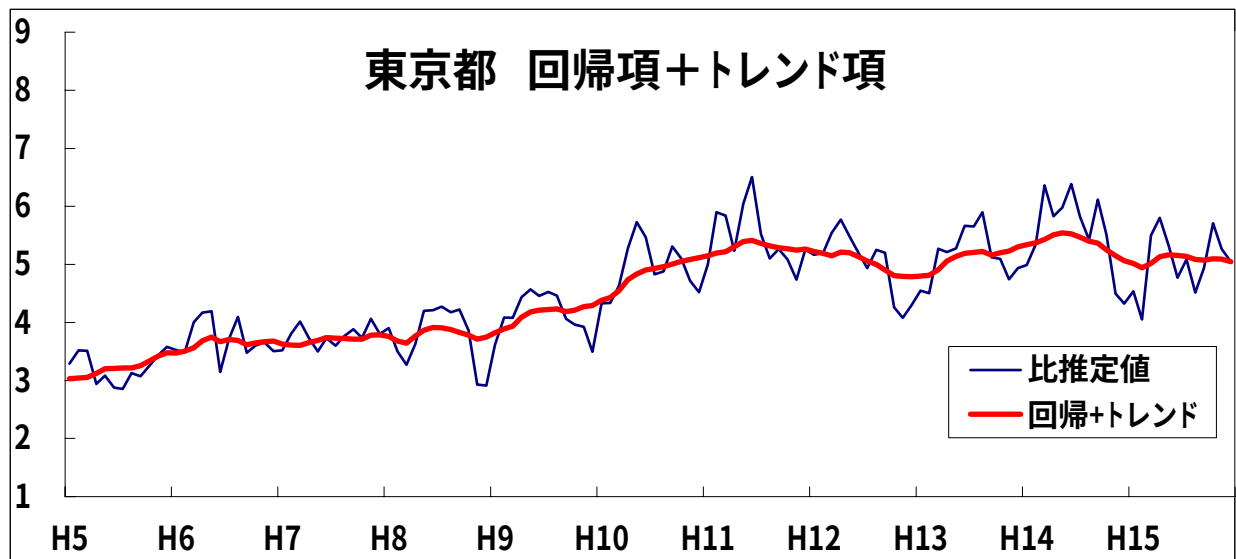


図 4-4 岐阜県推定結果のグラフ

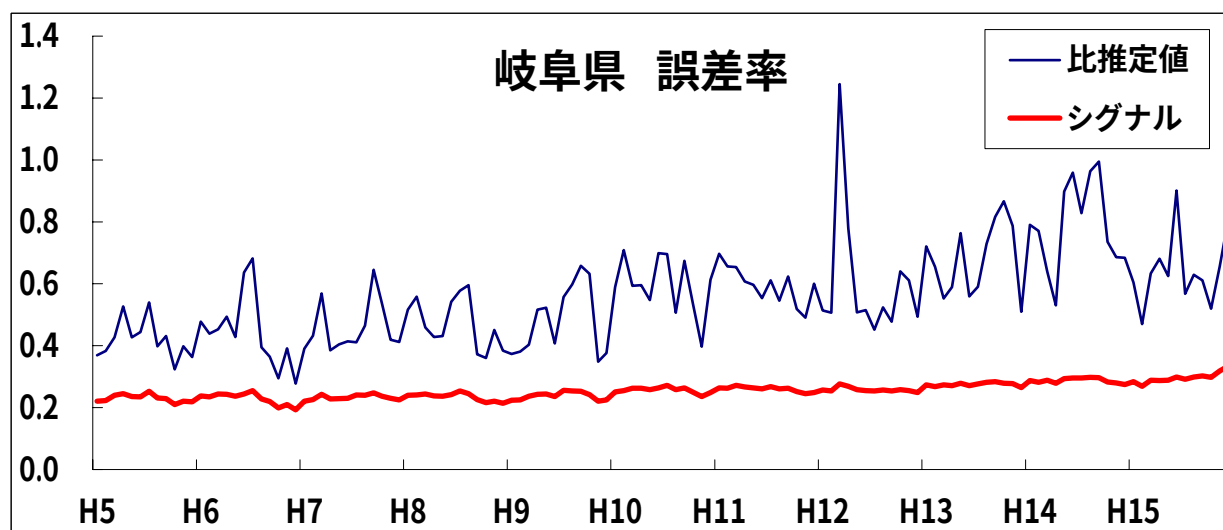
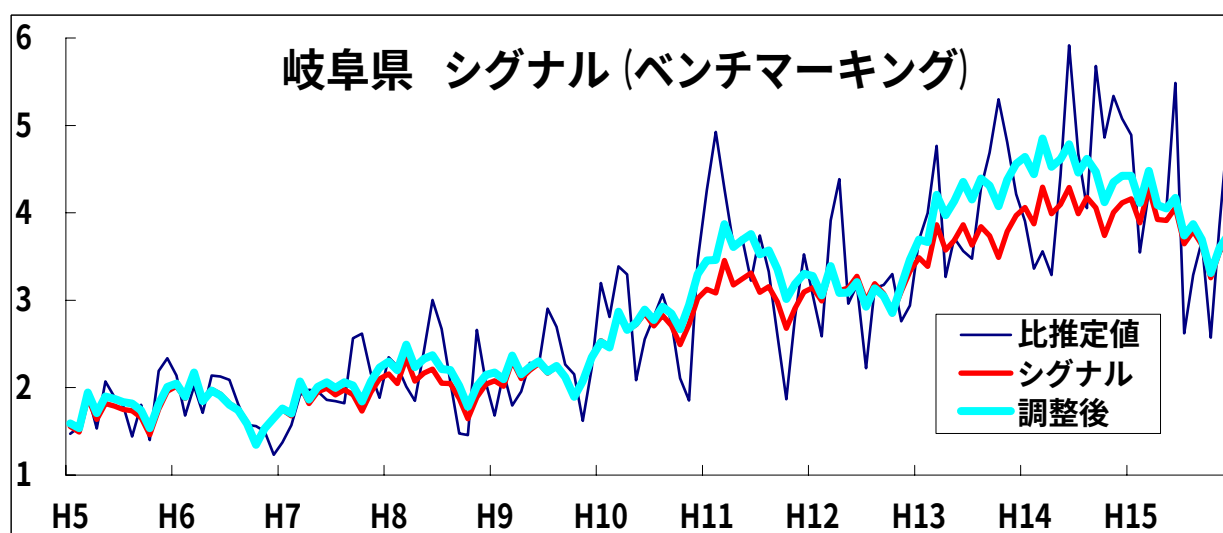
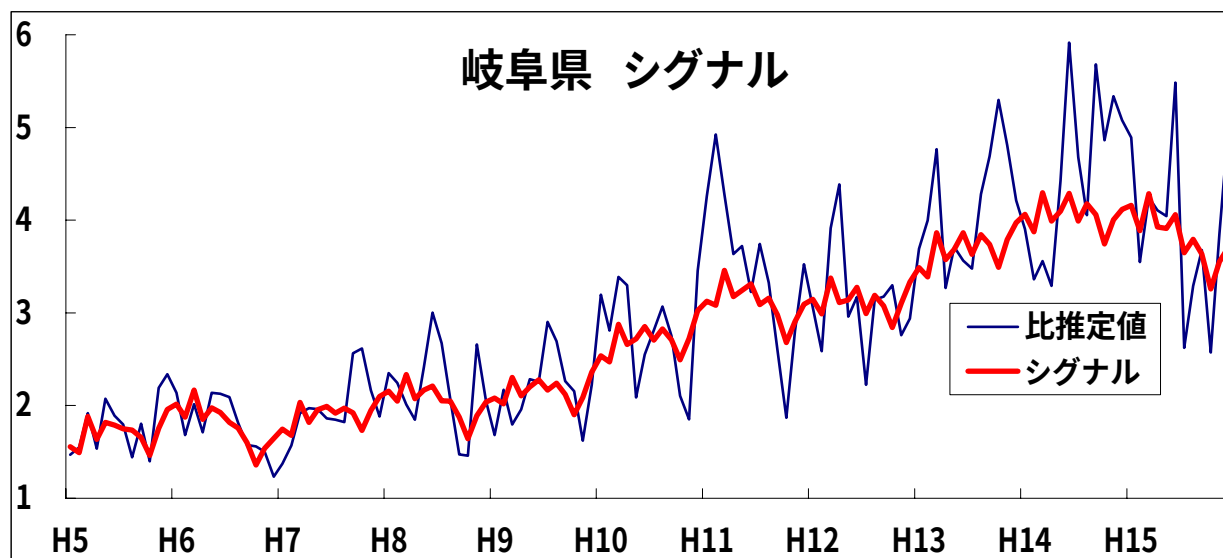


図 4-4 岐阜県推定結果のグラフ (続き)

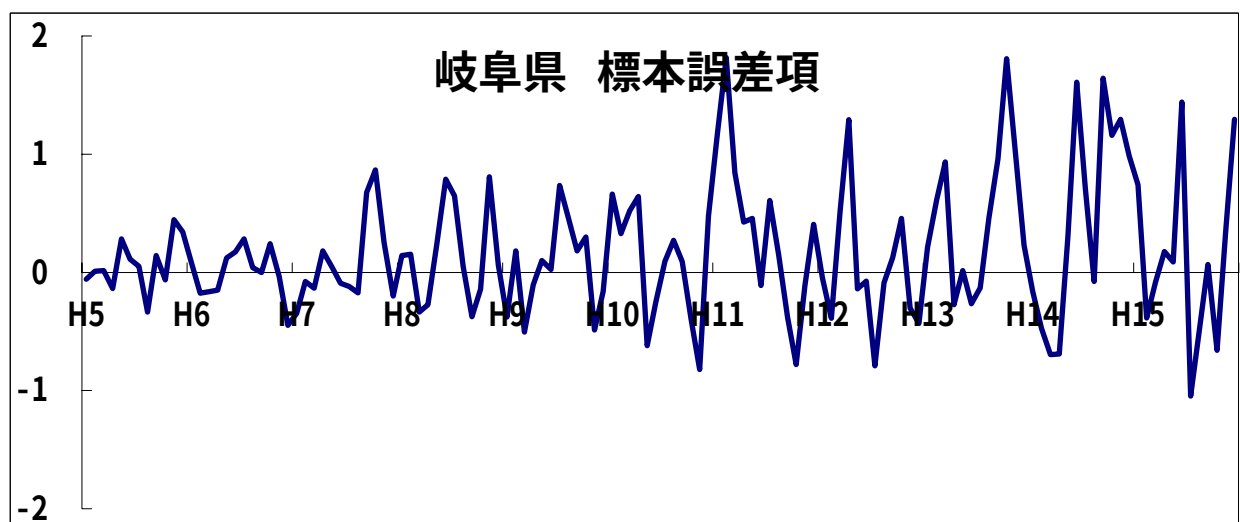
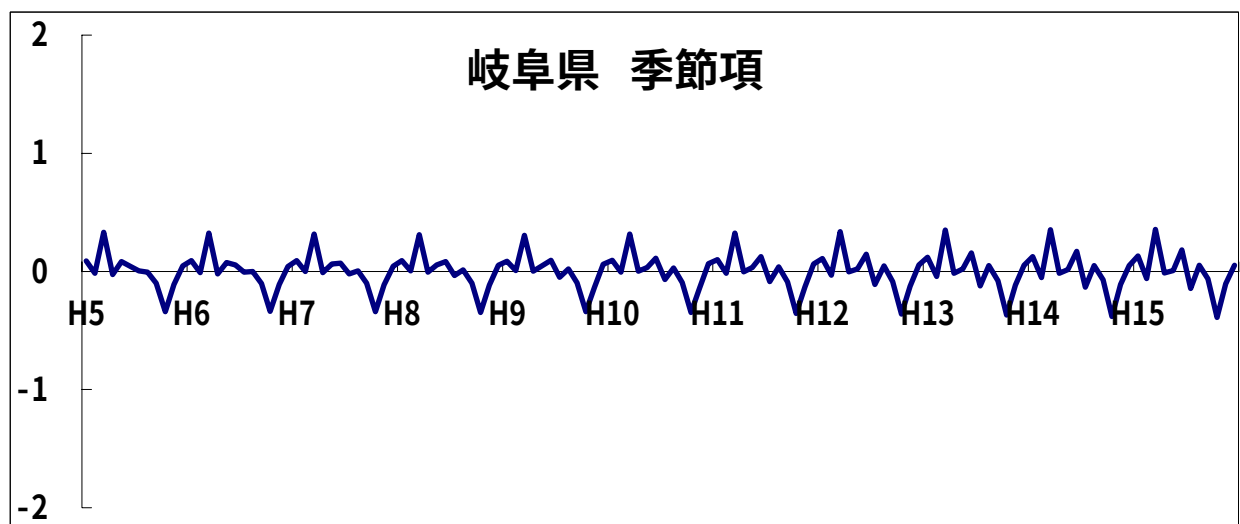
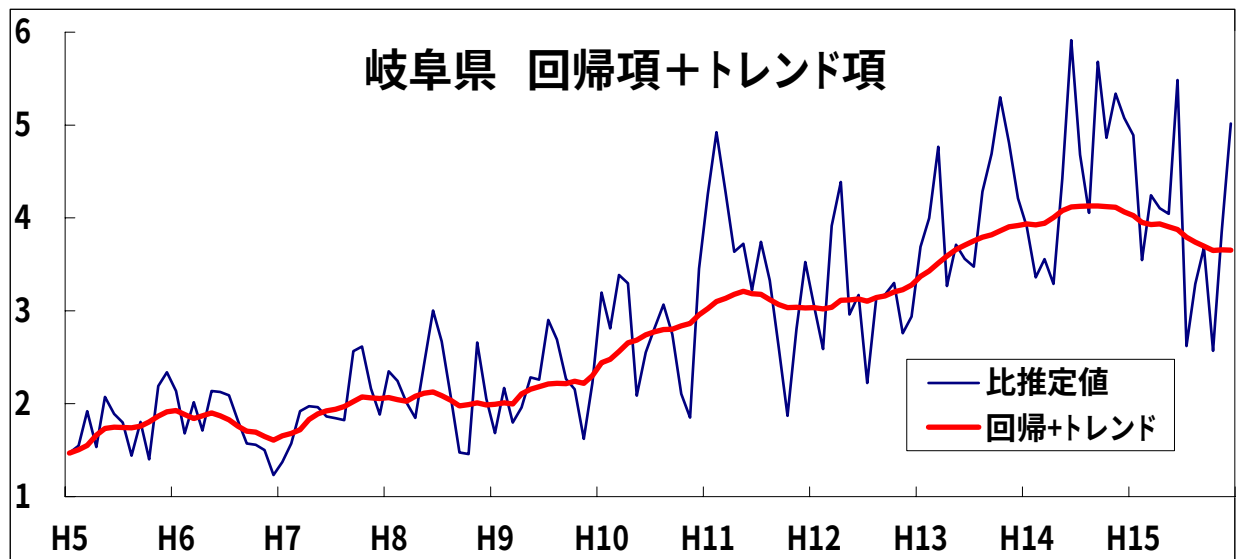


図 4-5 鳥取県推定結果のグラフ

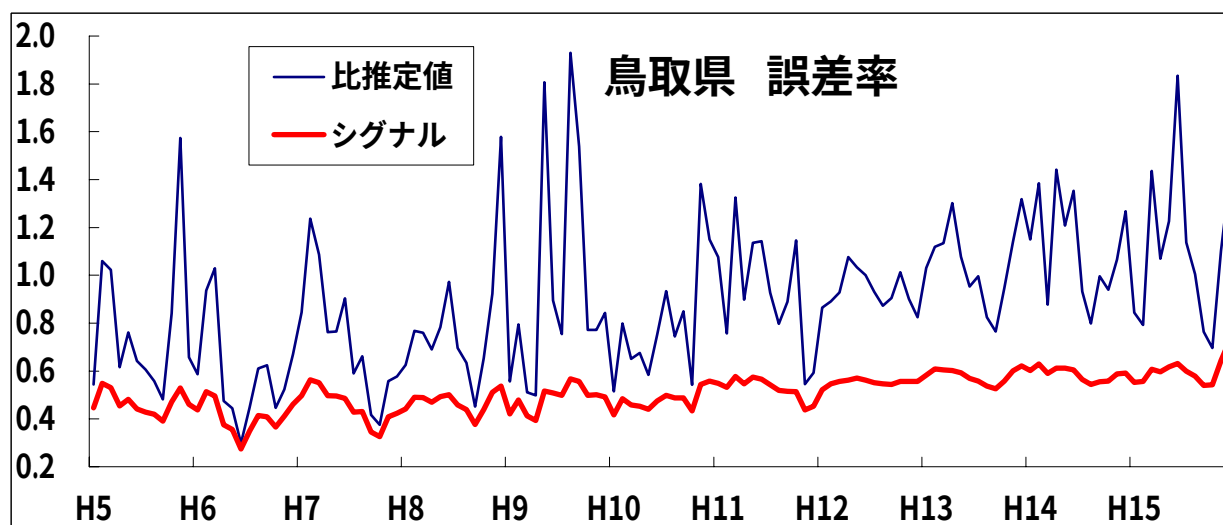
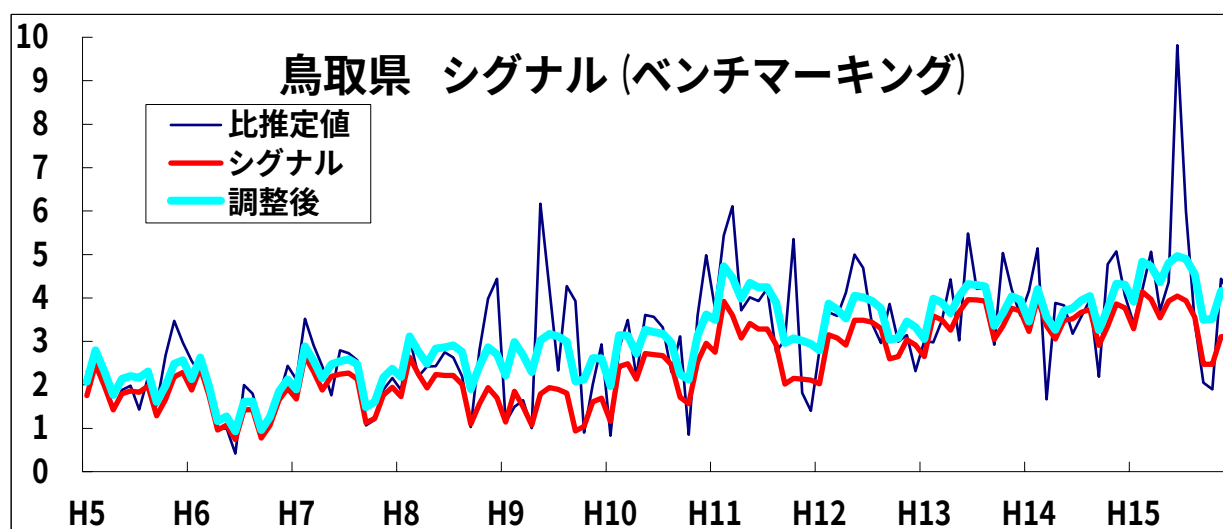
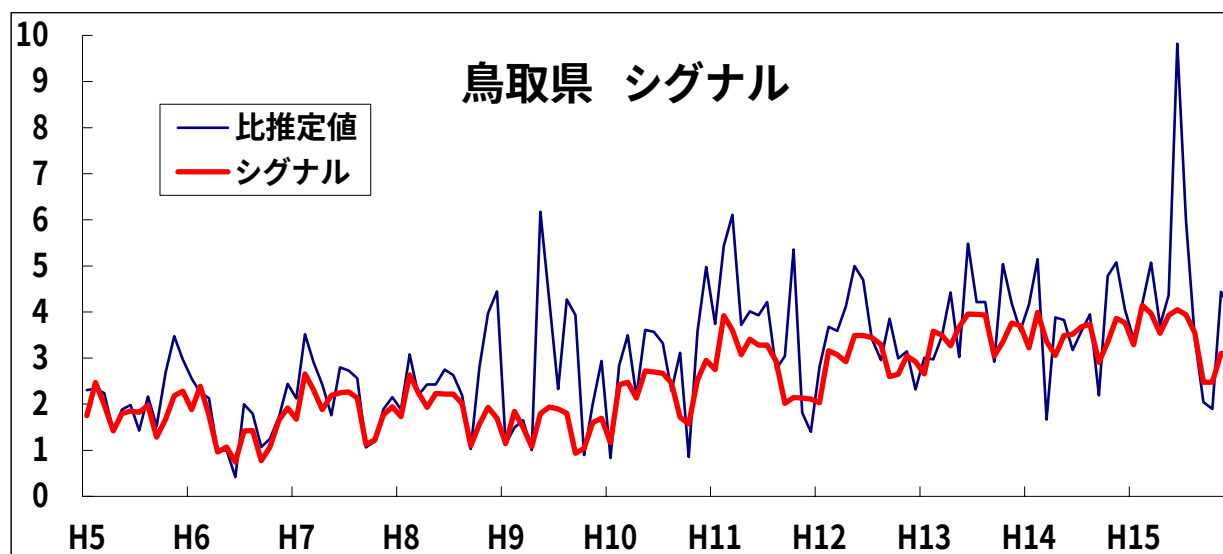


図 4-5 鳥取県推定結果のグラフ (続き)

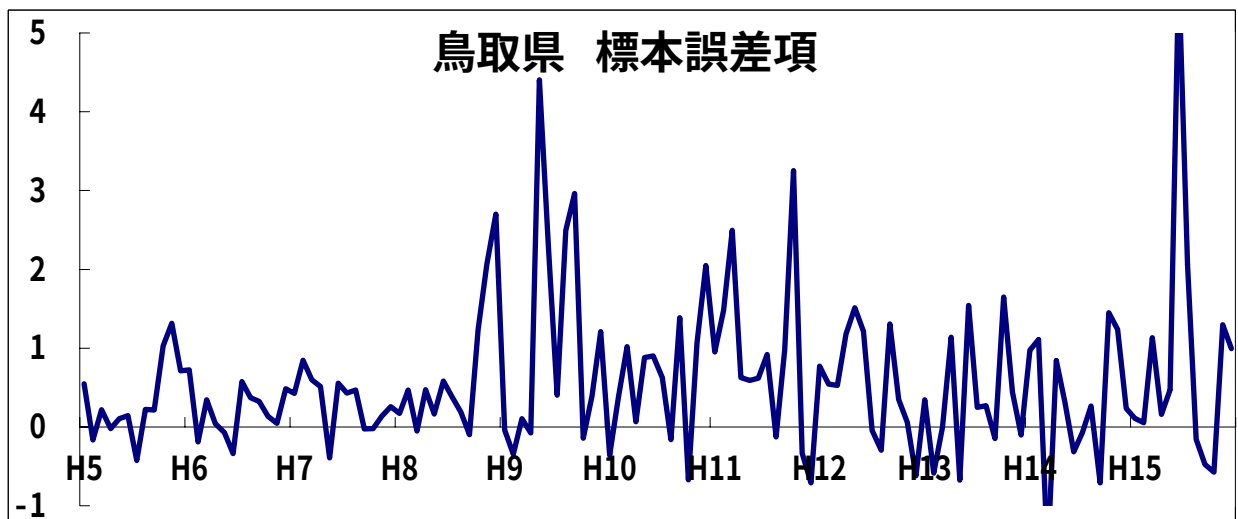
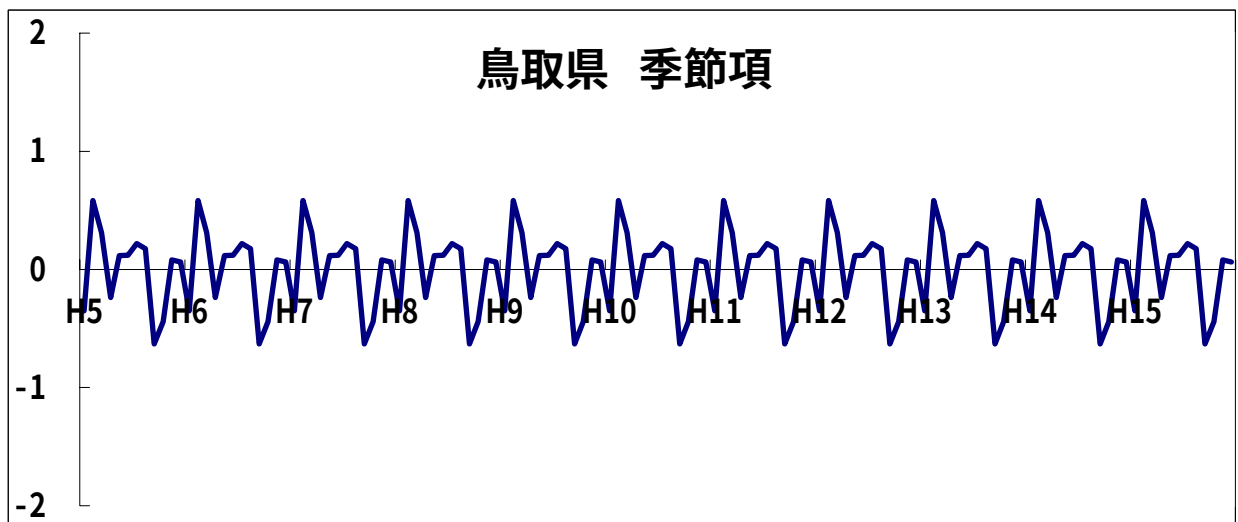
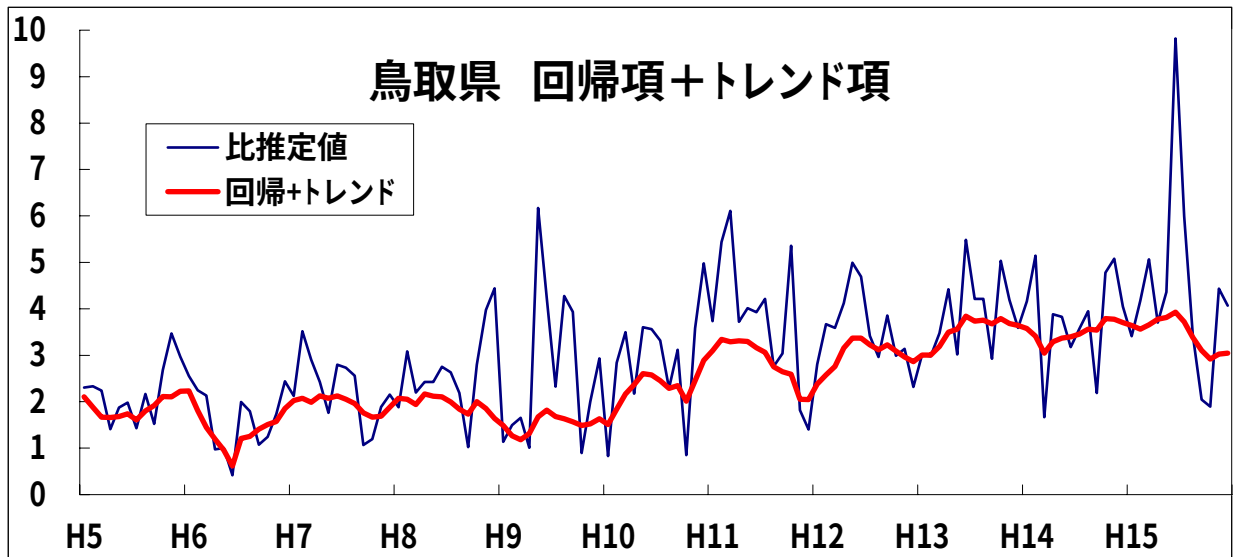


図 4-6 香川県推定結果のグラフ

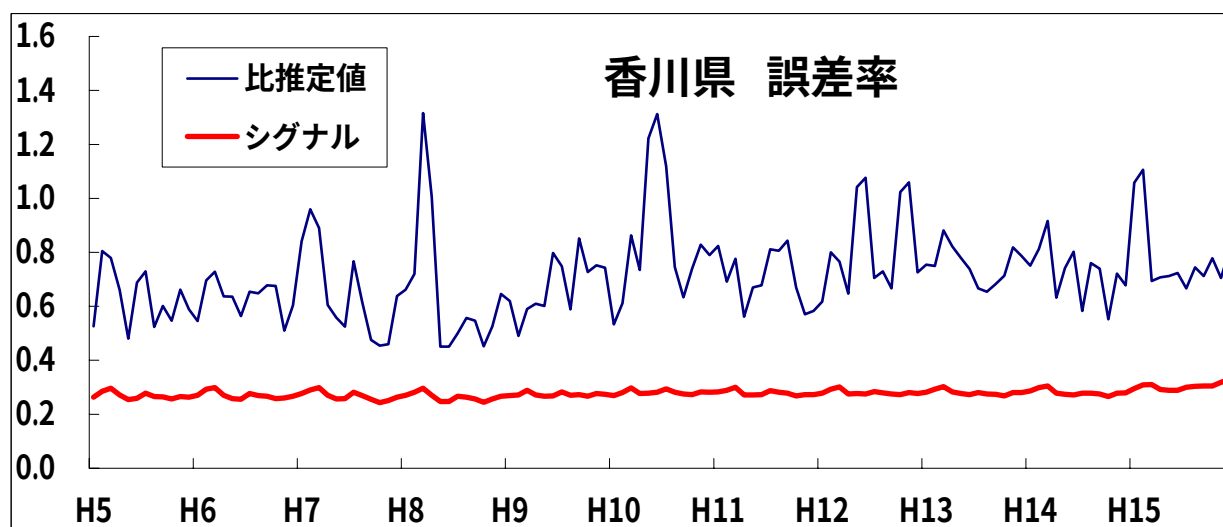
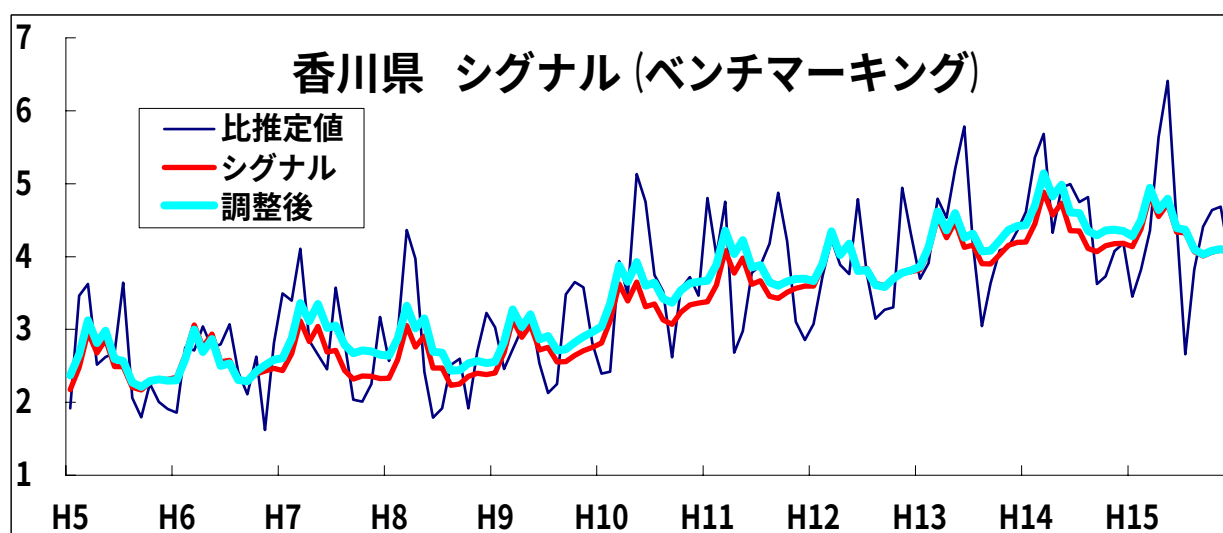
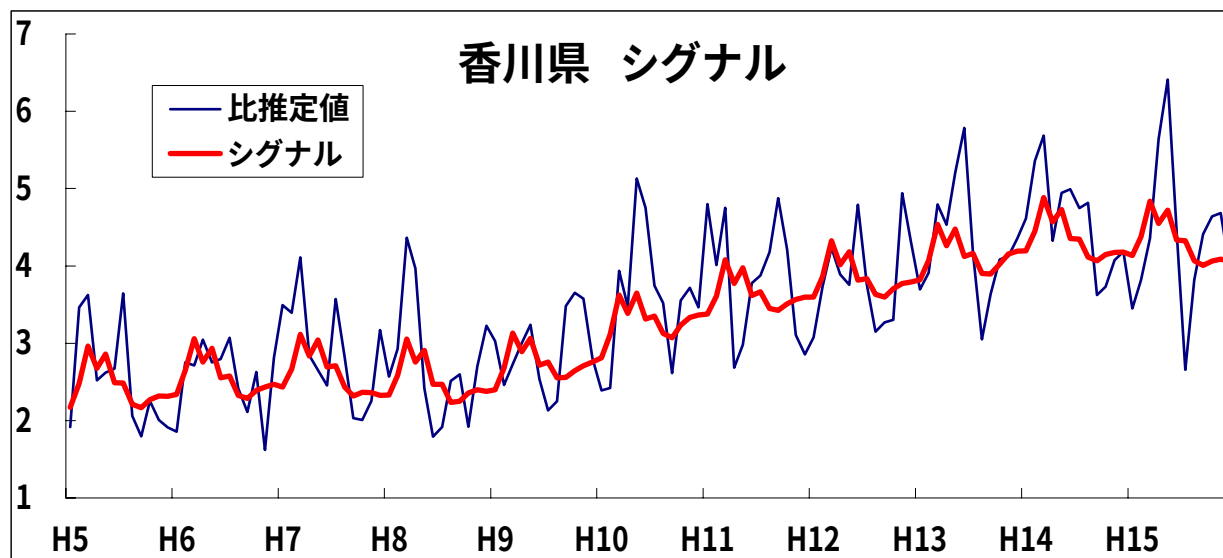


図 4-6 香川県推定結果のグラフ (続き)

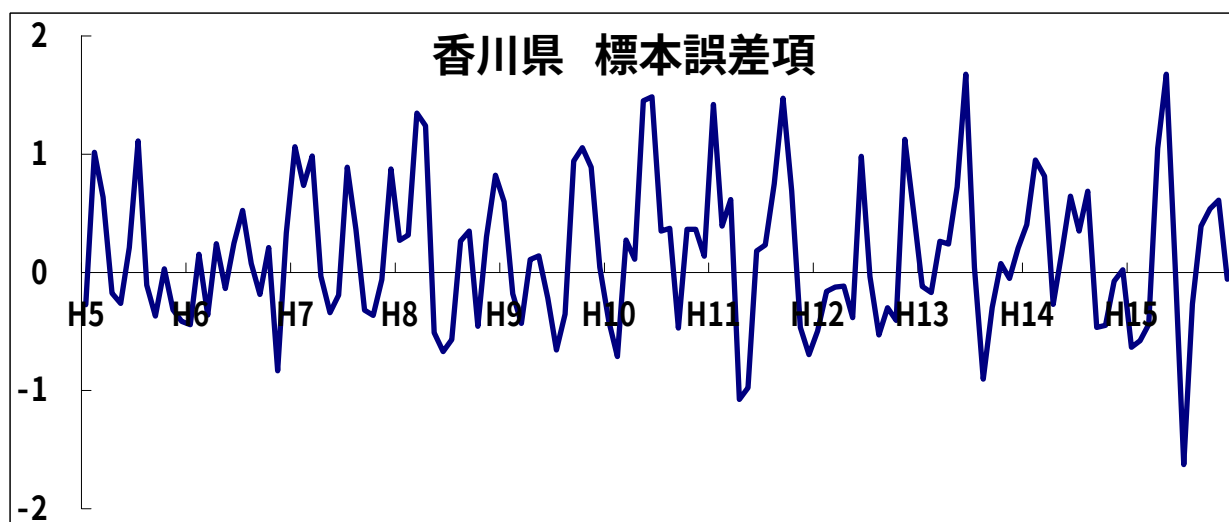
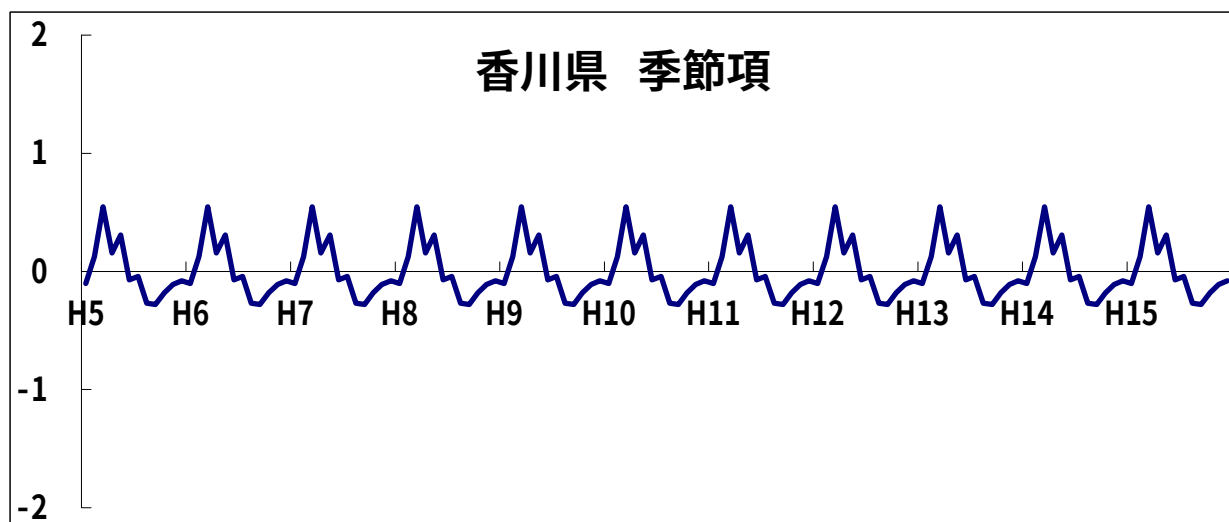
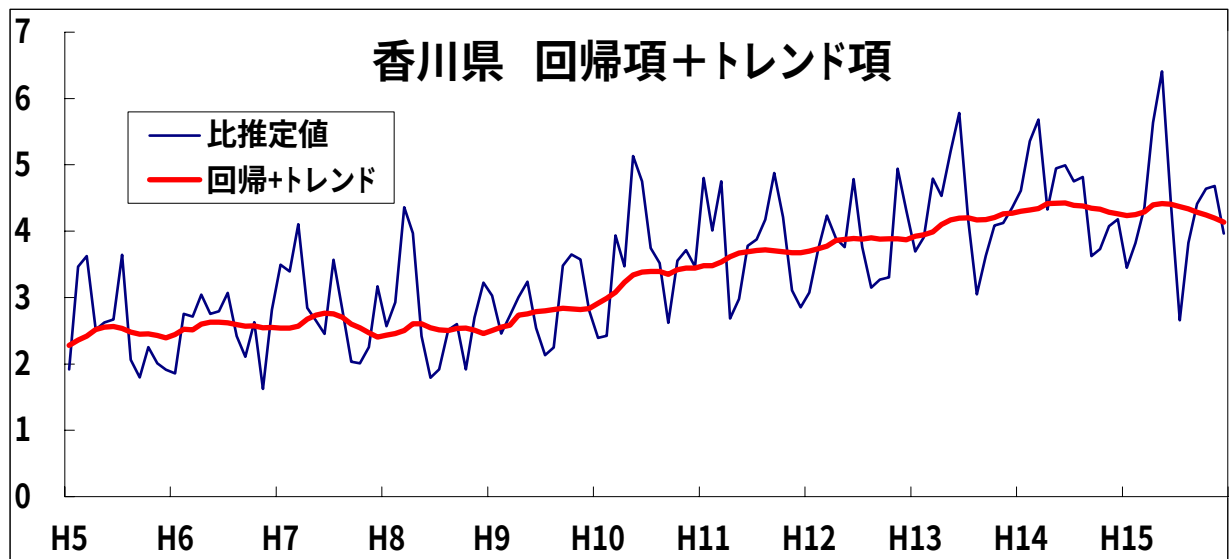


図 4-7 佐賀県推定結果のグラフ

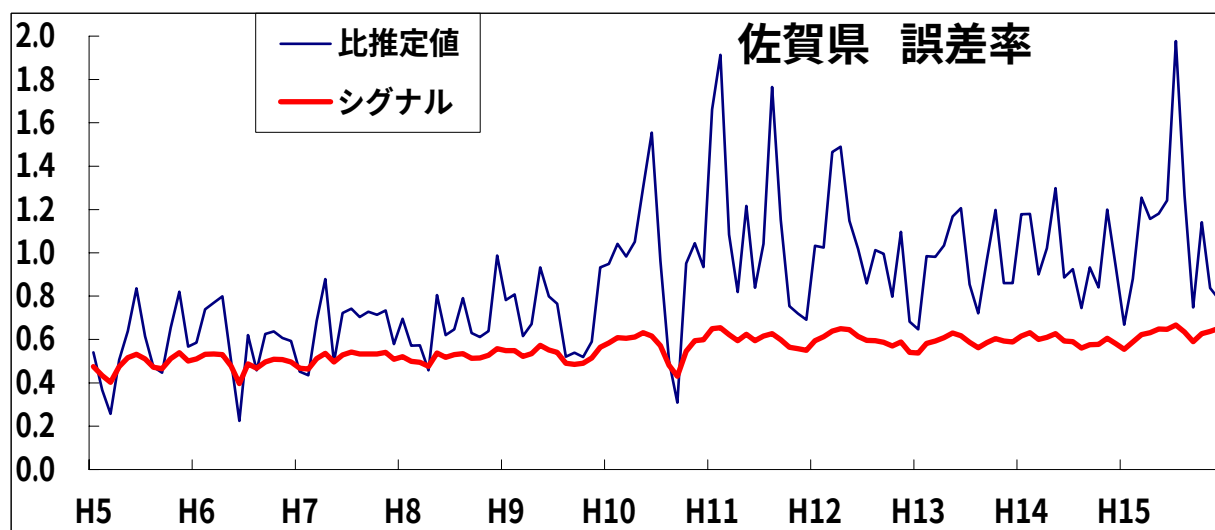
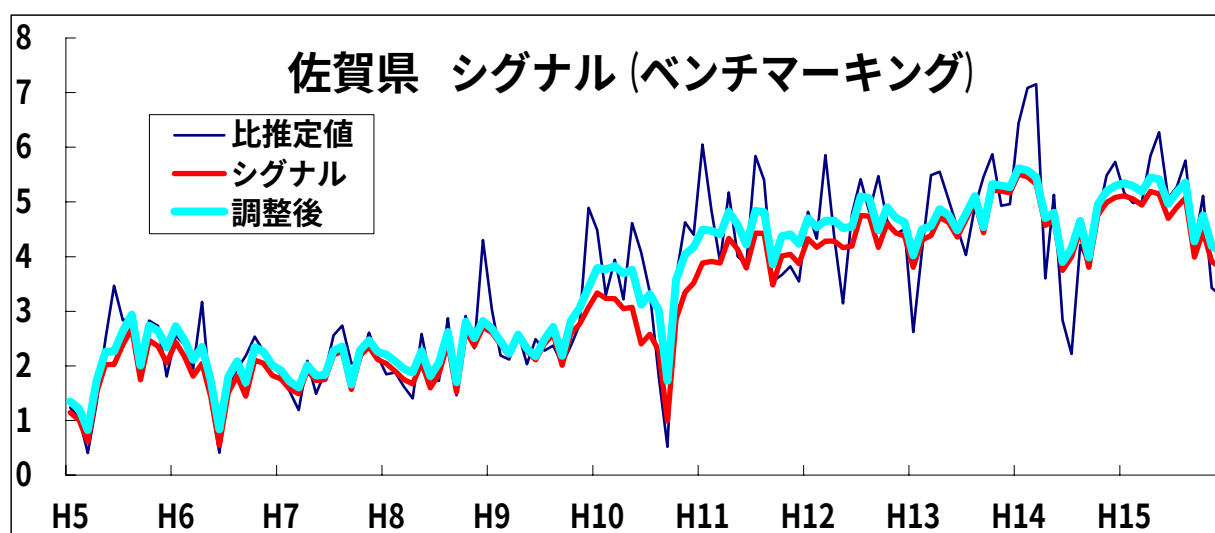
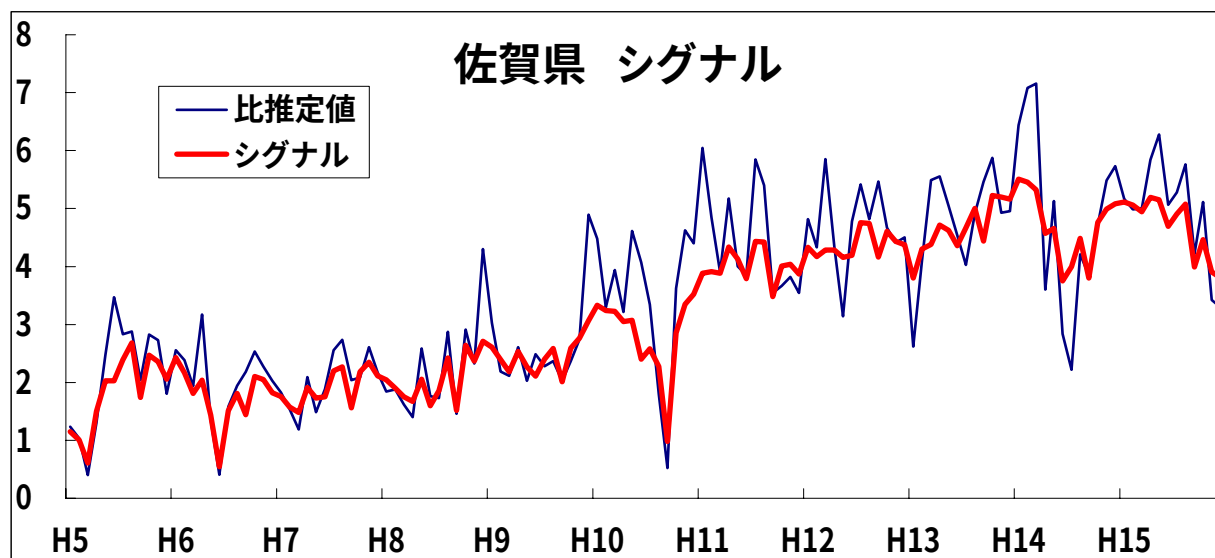


図 4-7 佐賀県推定結果のグラフ (続き)

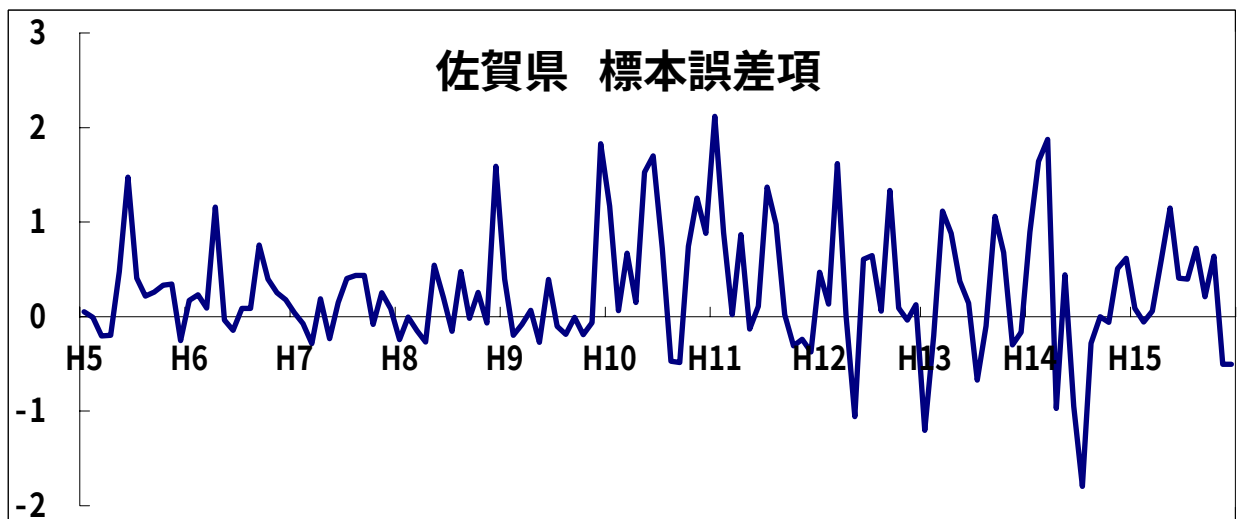
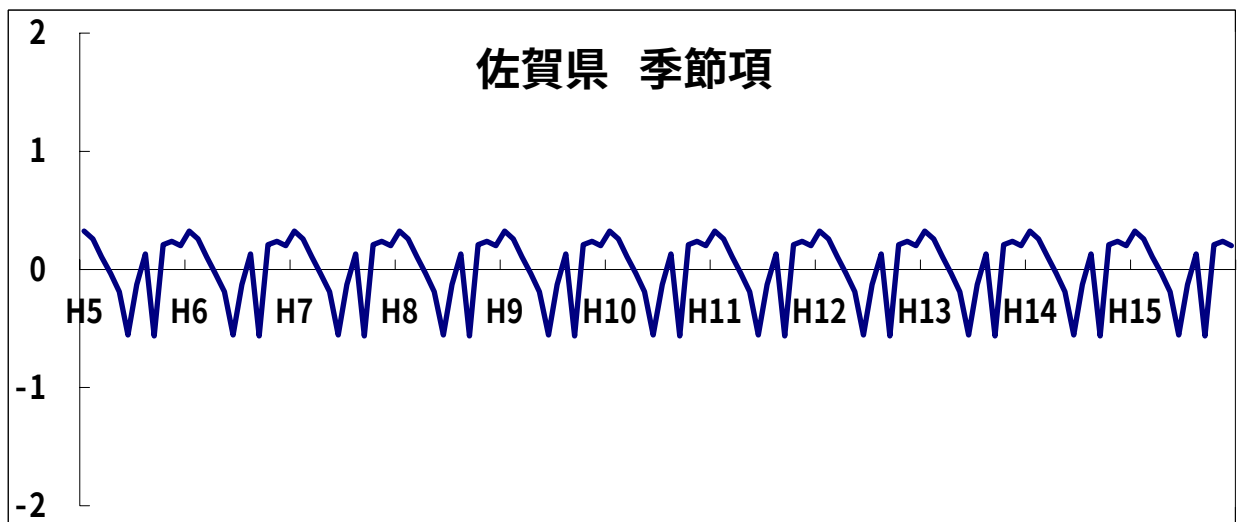
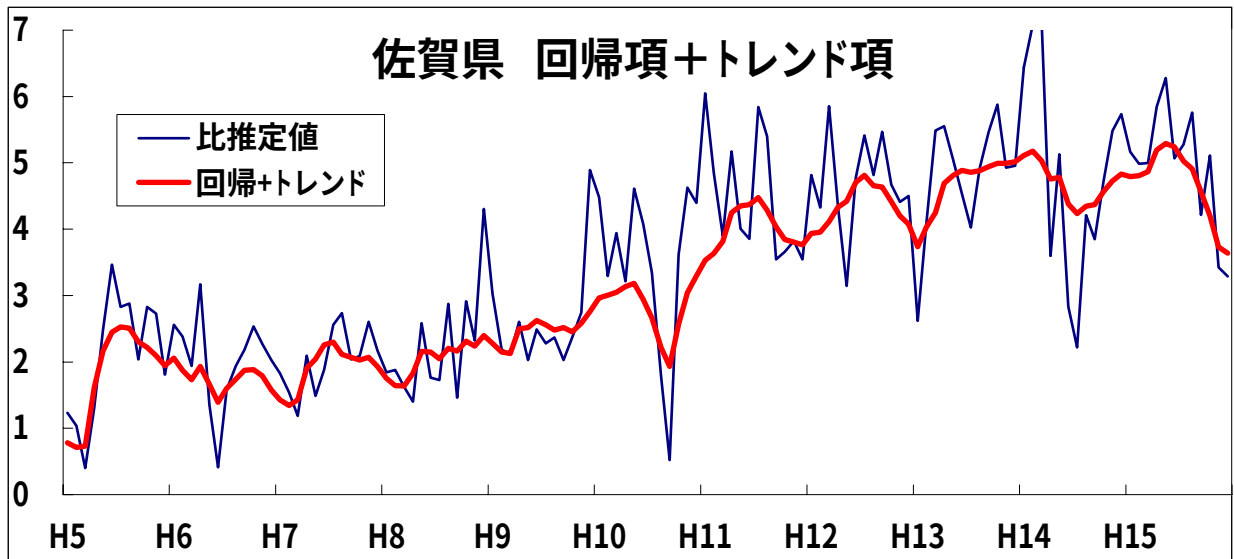


図 4-8 沖縄県推定結果のグラフ

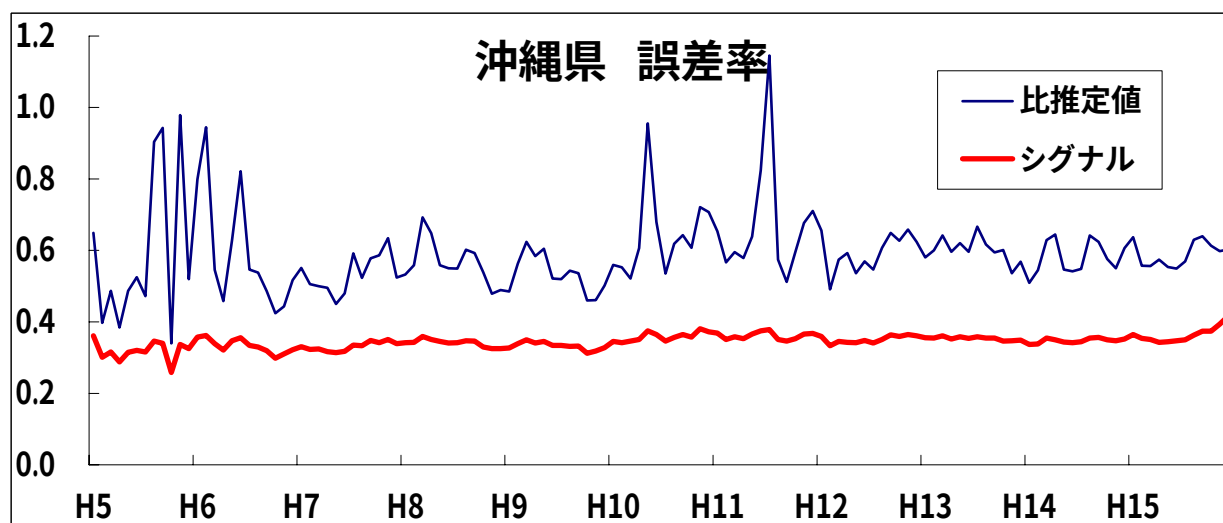
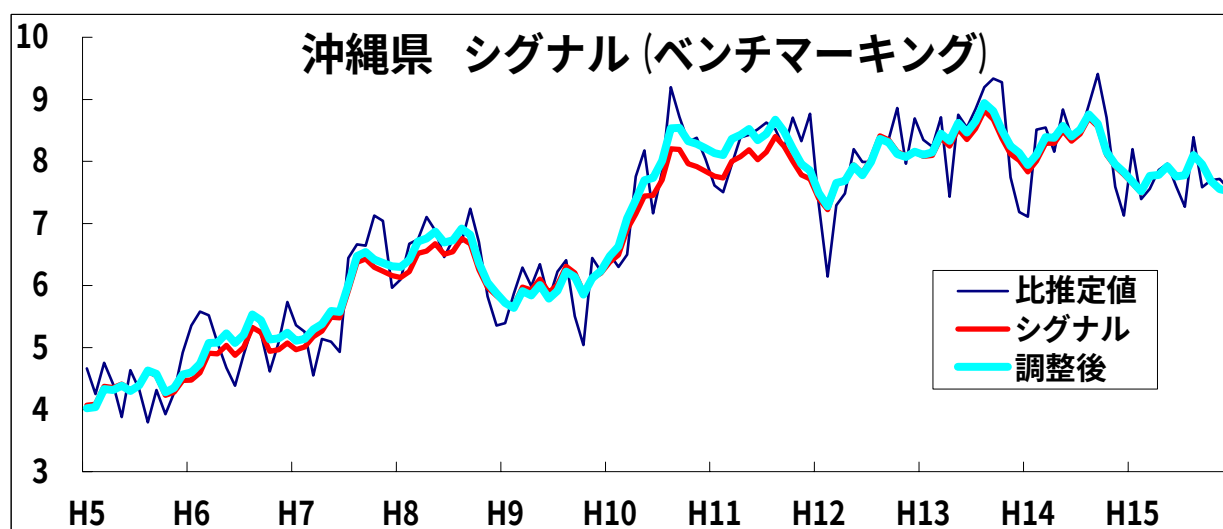
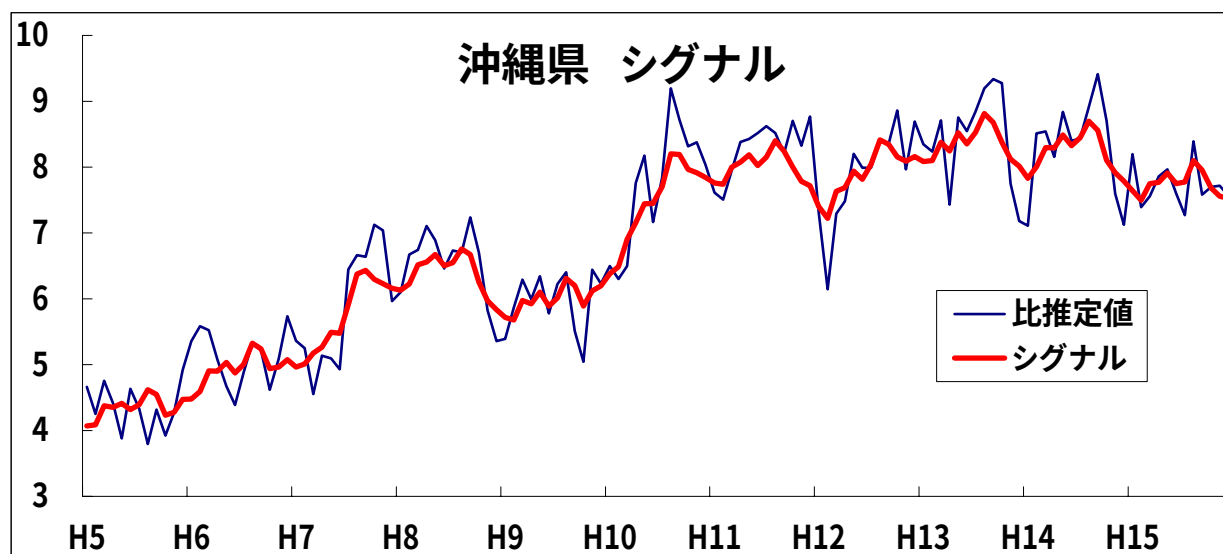


図 4-8 沖縄県推定結果のグラフ (続き)

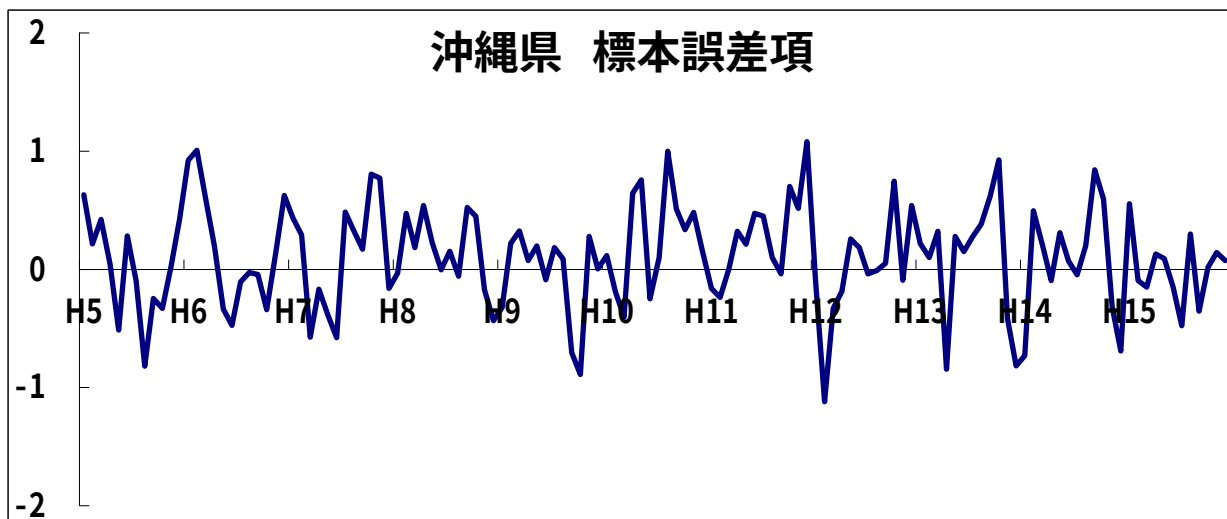
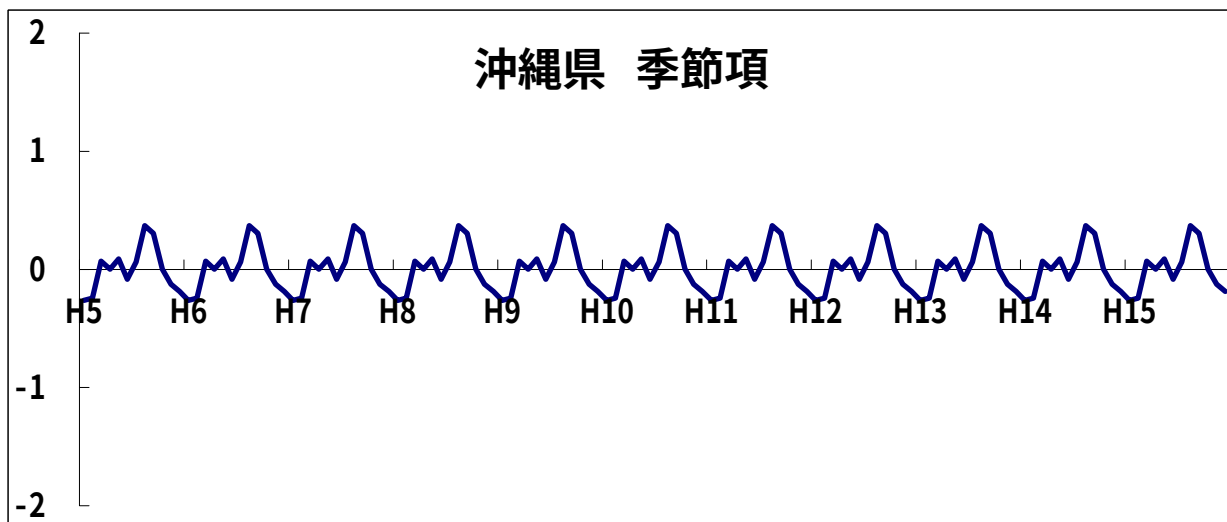
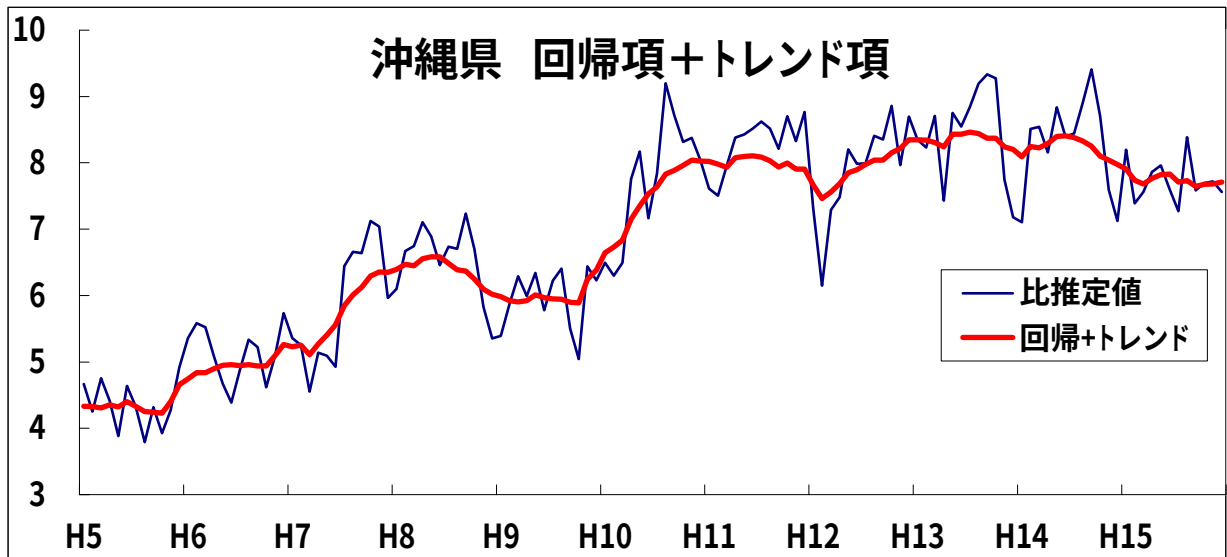


表 1 各種推定値の表

	都道府県名	10地域区分	H12～H14 平均調査区数
1	北海道	北海道	178
2	青森県	東北	34
3	岩手県		35
4	宮城県		54
5	秋田県		26
6	山形県		30
7	福島県		55
8	茨城県	北関東・甲信	66
9	栃木県		46
10	群馬県		50
11	埼玉県	南関東	114
12	千葉県		96
13	東京都		220
14	神奈川県		148
15	新潟県	北陸	75
16	富山県		30
17	石川県		41
18	福井県		23
19	山梨県	北関東・甲信	21
20	長野県		52
21	岐阜県	東海	45
22	静岡県		74
23	愛知県		148
24	三重県		40
25	滋賀県		24
26	京都府	近畿	49
27	大阪府		175
28	兵庫県		106
29	奈良県		28
30	和歌山県		20
31	鳥取県	中国	16
32	島根県		21
33	岡山県		50
34	広島県		77
35	山口県		45
36	徳島県	四国	30
37	香川県		33
38	愛媛県		55
39	高知県		33
40	福岡県	九州	107
41	佐賀県		18
42	長崎県		33
43	熊本県		38
44	大分県		26
45	宮崎県		27
46	鹿児島県		39
47	沖縄県		144

表 2 都道府県別平均標本調査区数 (平成 12 年~14 年)

対数尤度

1	3	13	21	31	37	41	47
北海道	宮城県	東京都	岐阜県	鳥取県	香川県	佐賀県	沖縄県
-117.94	-151.65	-92.10	-142.95	-229.74	-159.30	-215.24	-149.61

診断統計量

	1	4	13	21	31	37	41	47
	北海道	宮城県	東京都	岐阜県	鳥取県	香川県	佐賀県	沖縄県
J.B	2.03	23.35	3.05	4.29	1.13	1.23	19.02	1.73
S	-0.18	-0.59	-0.32	-0.38	-0.01	-0.22	-0.68	-0.07
K	2.50	4.70	3.39	3.45	3.46	3.19	4.28	3.55
LB12	10.21	8.96	12.49	10.13	14.48	16.03	17.42	8.77
LB24	17.13	22.29	27.15	24.65	23.59	27.58	32.60	22.71
F	0.89	0.88	0.71	0.41	0.46	1.01	0.68	0.73

p値

	1	4	13	21	31	37	41	47
	北海道	宮城県	東京都	岐阜県	鳥取県	香川県	佐賀県	沖縄県
J.B	0.36	0.00	0.22	0.12	0.57	0.54	0.00	0.42
LB12	0.60	0.71	0.41	0.60	0.27	0.19	0.13	0.72
LB24	0.84	0.56	0.30	0.42	0.49	0.28	0.11	0.54
F	0.72	0.67	0.26	0.00	0.01	0.97	0.22	0.32

ハイパーパラメータ

	1	4	13	21	31	37	41	47
	北海道	宮城県	東京都	岐阜県	鳥取県	香川県	佐賀県	沖縄県
回帰項	2.0E-06	4.7E-09	7.0E-08	2.4E-11	4.4E-03	1.4E-09	1.1E-14	4.2E-07
トレンド項	1.5E-02	2.6E-02	1.3E-02	1.3E-02	1.3E-01	6.3E-03	1.5E-01	4.9E-02
季節項	2.3E-17	4.1E-12	3.1E-09	1.0E-03	3.3E-11	5.5E-15	1.9E-06	1.9E-13
不規則変動項	7.2E-03	1.9E-08	5.7E-08	5.2E-14	4.4E-11	4.4E-13	6.7E-02	8.1E-08

III 推定に用いたプログラム

今回の推定に用いたプログラムを、以下に示す。**Est** という関数により、分析を行うことが出来る。**Est** の引数 **pref** は、都道府県番号である。

プログラム中のデータの内容は、以下のとおりである。

Y.dt：観測値 T 行 × 47 列の行列
X.dt：補助情報 T 行 × 47 列の行列
E.dt：標準誤差 T 行 × 47 列の行列
SEA：誤差の系列相関（0 次を除く） 13 行 × 47 列の行列

ここで T は時系列の長さを表す。計算結果の出力は、以下のとおりである。

sgnl：シグナル（真の値）の推定値。
be：シグナルにベンチマーキングを適用した結果。
MT：回帰項 + トренд項の推定値。
S：季節項の推定値。
I：不規則変動項の推定値。
e：標本誤差項の推定値。
std：シグナル（ベンチマーキングをする前）の標準誤差。
rsd：基準化した一期先予測誤差の推定値。
par：ハイパーパラメータの推定値。内訳は以下のとおり。
 M：回帰項の分散
 T：トレンド項の分散
 S：季節項の分散
 I：不規則変動項の分散
stat：診断に用いる各種統計量。内訳は以下のとおり。
 JB：Jarque - Bera 統計量。
 S：歪度。
 K：尖度。
 LB12：Ljung-Box 統計量（次数：12）。
 LB24：Ljung-Box 統計量（次数：24）。
 F：F 統計量。
p：stat の各統計量の p 値。

```
#####
#   Small Area Estimation by the State Space model
#   (Using Kalman Filter)
#####
```

```
library(ts)
library(MASS)
```

```
Y.dt <- matrix(scan("D:/uemp.r.txt"), ncol = 47, byrow = TRUE)
X.dt <- matrix(scan("D:/job.r.txt"), ncol = 47, byrow = TRUE)
E.dt <- matrix(scan("D:/uemp.r.err.TXT"), ncol = 47, byrow = TRUE)
SEA <- matrix(scan("D:/uemp.r.SEA.TXT"), ncol = 47, byrow = TRUE)
leng <- nrow(Y.dt)
```

```
Y.W <- function(acov)
{
  work <- diag(13)
  for (i in 1:12) {
    work[abs(row(work) - col(work)) == i] <- acov[i]
  }
  solve(work, acov)
}
```

```
Set.model <- function (pref)
{
  Y <- Y.dt[, pref]
  X <- X.dt[, pref]
  AR <- Y.W(SEA[, pref])
  inp <- c(1, ARMAtoMA(ar = AR, lag.max = 1000))
  gm <- E.dt[, pref] / sqrt(crossprod(inp))
  F <- matrix(0, 26, 26)
  G <- matrix(0, 26, 4)
  H0 <- t(rep(0, 26))
  F[row(F) == col(F) + 1] <- F[1, 1] <- F[2, 2] <-
    G[1, 1] <- G[2, 2] <- G[3, 3] <- G[14, 4] <-
    H0[, 2] <- H0[, 3] <- 1
  F[2, 1] <- F[3, 2] <- F[14, 13] <- 0
  F[3, 3:13] <- - 1
  F[14, 14:26] <- AR
}
```

```
JB.test <- function(r){
  m1 <- mean(r)
  m2 <- mean((r - m1) ^ 2)
```

```
m3 <- mean((r - m1) ^ 3)
m4 <- mean((r - m1) ^ 4)
wrk1 <- m2 ^ (3 / 2)
wrk2 <- m2 ^ 2
S <- mean(m3 / wrk1)
K <- mean(m4 / wrk2)
wrk3 <- (S ^ 2) / 6
wrk4 <- ((K - 3) ^ 2) / 24
JB <- length(r) * (wrk3 + wrk4)
return(list(JB = JB, S = S, K = K))
}
```

```
Diags <- function(rsd){
  JB.res <- JB.test(rsd)
  JB <- JB.res$JB
  JB.p <- 1 - pchisq(JB, df = 2)
  S <- JB.res$S
  K <- JB.res$K
  L.B12 <- Box.test(rsd, lag = 12, type = "Ljung")
  L.B24 <- Box.test(rsd, lag = 24, type = "Ljung")
  LB12 <- as.vector(L.B12$statistic)
  LB12.p <- as.vector(L.B12$p.value)
  LB24 <- as.vector(L.B24$statistic)
  LB24.p <- as.vector(L.B24$p.value)
  n1 <- length(rsd)
  n2 <- floor(n1 / 3)
  het <- var.test(rsd[1:n2], rsd[(n1 - n2 + 1):n1])
  F <- as.vector(het$statistic)
  F.p <- as.vector(het$p.value)
  list(JB = JB, JB.p = JB.p, S = S, K = K, LB12 = LB12,
       LB12.p = LB12.p, LB24 = LB24, LB24.p = LB24.p,
       F = F, F.p = F.p)
}
```

```
Out <- function (t.s, be, rsd, par, lik, test)
{
  sgnl <- t.s$sgnl
  std <- t.s$std
  plot(Y, type = "l")
  points(sgnl, type = "l", col = "red", lwd = 2)
  points(be, type = "l", col = "green", lwd = 3)
  param <- exp(par[1:4])
}
```

```

names(param) <- c("M", "T", "S", "I")
stat <- c(test$JB, test$S, test$K, test$LB12,
          test$LB24, test$F)
names(stat) <- c("JB", "S", "K", "LB12", "LB24", "F")
p <- c(test$JB.p, test$LB12.p, test$LB24.p, test$F.p)
names(p) <- c("JB.p", "LB12.p", "LB24.p", "F.p")
list(sgnl = sgnl, be = be, MT = t.s$MT, S = t.s$S,
      I = t.s$I, e = t.s$e, std = std, rsd = rsd,
      par = param, log.lik= lik,
      stat = stat, p = p)
}

Symt <- function (A)
{
  temp <- t(A)
  A[lower.tri(A)] <- temp[lower.tri(temp)]
  A
}

K.Filter <- function(par, type, Z, P)
{
  H <- H0
  Q <- diag(c(exp(par[1:3]), 1))
  sig <- exp(par[4])
  if (type == "sm") {
    Zw1 <- Zw2 <- matrix(NA, 26, leng)
    Pw1 <- Pw2 <- array(NA, c(26, 26, leng))
    rsd <- rep(NA, leng)
  }
  work <- 0
  for (iter in 1:leng) {
    H[, 1] <- X[iter]
    H[, 14] <- gm[iter]
    Z1 <- F %%% Z
    P1 <- F %%% P %%% t(F) + G %%% Q %%% t(G)
    P1 <- Symt(P1)
    r <- as.vector(Y[iter] - H %%% Z1)
    d <- as.vector(H %%% P1 %%% t(H) + sig)
    K <- P1 %%% t(H) / d
    Z <- Z1 + K * r
    P <- P1 - K %%% H %%% P1
    P <- Symt(P)
    if (type == "sm") {
      Zw1[, iter] <- Z1
    }
  }
}

```

```

        Pw1[, , iter] <- P1
        Zw2[, iter] <- Z
        Pw2[, , iter] <- P
        rsd[iter] <- r / sqrt(d)
    }
    if (iter > 26) work <- work + log(d) + (r ^ 2) / d
}
lik <- -0.5 * (leng * log(2 * pi) + work)
if (type == "lk") return(-1 * lik)
if (type == "sm") {
    t.s <- Smoother(F, sig, Zw1, Zw2, Pw1, Pw2)
    test <- Diags(rsd[27:leng])
    be <- Bench(t.s$sgnl)
}
Out(t.s, be, rsd, par, lik, test)
}

```

```

Smoother <- function(F, sig, Zw1, Zw2, Pw1, Pw2)
{
    sgnl <- MT <- e <- S <- I <- std <- rep(NA, leng)
    H <- H0
    Zs <- Zw2[, leng]
    Ps <- Pw2[, , leng]
    for (iter in leng:1) {
        H[, 1] <- X[iter]
        if (iter < leng) {
            Z1 <- Zw1[, iter + 1, drop = FALSE]
            Z2 <- Zw2[, iter, drop = FALSE]
            P1 <- Pw1[, , iter + 1]
            P2 <- Pw2[, , iter]
            A <- P2 %%% crossprod(F, ginv(P1))
            Zs <- Z2 + A %%% (Zs - Z1)
            Ps <- P2 + A %%% (Ps - P1) %%% t(A)
            Ps <- Synt(Ps)
        }
        MT[iter] <- wrk1 <- Zs[1] * X[iter] + Zs[2]
        S[iter] <- wrk2 <- Zs[3]
        e[iter] <- wrk3 <- Zs[14] * gm[iter]
        I[iter] <- Y[iter] - (wrk1 + wrk2 + wrk3)
        sgnl[iter] <- Y[iter] - wrk3
        std[iter] <- sqrt(H %%% Ps %%% t(H) + sig)
    }
}

```

```
list(sgnl = sgnl, MT = MT, S = S, I = I, e = e, std = std)
}
```

```
Bench <- function (Z)
{
  yr <- leng / 12
  B <- matrix(0, leng, yr)
  for(i in 1:yr) B[(12 * i - 11):(12 * i),i] <- 1
  D <- matrix(0, leng, leng)
  D[leng, leng] <- 1
  for(i in 1:(leng - 1)){
    D[i, i] <- 1
    D[i + 1, i] <- -1
  }
  iA <- solve(t(D) %*% D)
  wrk1 <- iA %*% B %*% solve(t(B) %*% iA %*% B)
  wrk2 <- t(B) %*% (Y - Z)
  as.vector(Z + wrk1 %*% wrk2)
}
```

```
Est <- function(pref) {
  Set.model(pref)
  Z0 <- matrix(rep(0, 26))
  reg <- lm(Y ~ X)
  Z0[1] <- reg$coef["X"]
  Z0[2] <- reg$coef["(Intercept)"]
  P0 <- diag(rep(var(Y[1:24]) * 10000, 26))
  fun <- function(par) K.Filter(par, "lk", Z0, P0)
  init <- c(5e-3, 5e-3, 1e-10, 1e-10)
  res <- optim(log(init), fun, method = "Nelder-Mead",
    control = list(trace = 10, REPORT = 10, maxit = 1000))
  res <- optim(res$par, fun, method = "BFGS",
    control = list(trace = 10, REPORT = 10, maxit = 1000))
  cat("¥n", res$par, "¥n")
  K.Filter(res$par, "sm", Z0, P0)
}
```

```
result <- Est(47)
```

第 2 章 EBLUP(Empirical Best Linear Unbiased Prediction) 推定量

を用いた小地域推定 ～単純な Fay-Herriot モデルの利用～

この章では、EBLUP 推定量 (Empirical Best Linear Unbiased Prediction: 経験的最良線形不偏予測量) を用いた都道府県別完全失業率の推定方法についてまとめる。

EBLUP 推定量は、BLUP 推定量 (Best Linear Unbiased Prediction: 最良線形不偏予測量) 中の未知パラメータを、最尤法などで観測値データから推定した値で置き換えたものである。BLUP 推定量ではデータの背後にある構造として、線形回帰モデルを想定する。ただし単純な線形回帰モデルではなく、小地域ごとの特徴も考慮したモデルとなっている。このようなモデルを線形混合効果モデルと呼ぶ。このときモデルから導かれる BLUP 推定量は、もとの観測値と回帰モデルによる予測とを、それぞれの分散の比からなるウェイトを用いて加重平均した形になっている。そして回帰の予測が信頼できない場合には観測値にかかるウェイトが増し、逆の場合は回帰にかかるウェイトが増すという、それぞれの推定値の変動を考慮した、自然な理解しやすい形の推定量となっている。

推定した結果に対して、大規模調査及びセンサス (H7、H12 国勢調査、H14 就業構造基本調査) をベンチマークとした分析、MSE (Mean Square Error: 平均二乗誤差) の分析、及び時系列のグラフによる分析を行い、もとの観測値と比較した。その結果、もとの観測値に比べて EBLUP 推定量を用いた場合の方が、誤差率やベンチマークの観点で、良好であるという結論が得られた。ただし時系列のグラフで比較した場合には、沖縄県や大阪府など一部の地域で、EBLUP 推定量による推定値がもとの観測値と比較して常に過小又は過大評価となることが確認された。

今後の課題として、上記のような、過大又は過小な評価となった地域に関する先見的な情報をもとに、このような偏りを調整することが考えられる。また今回のモデルを多変量に拡張することや、Unit Level の補助情報を用いたモデルによる推定も今後の課題である。

以下、I 節では Fay-Herriot モデル、EBLUP 推定量及び EBLUP 推定量の MSE に関する理論について述べる。II 節では、I 節で述べた EBLUP 推定量を労働力調査へ適用し、都道府県別・月次の完全失業率を推定する方法について述べる。III 節では、今後の課題について述べる。

I Fay-Herriot モデルと EBLUP 推定量

EBLUP 推定量の理論について、Rao(2003) の 5 章～7 章を適宜引用しつつ解説する。推定式の導出などの理論的な詳細については省略する。詳細については Rao(2003) の該当する部分を参考にしたい。

小地域ごとに何らかの特徴があるデータを扱う場合を考える。単純な線形回帰モデルでは、通常ダミー変数を入れることで、このような問題に対処する。しかしそれでは小地域一つ一つの特徴を取り出すことは難しく、また自由度が低下するという問題も生ずる。線形混合効果モデルをもとにした BLUP 推定量を用いることで、このような問題点に対処することができ、精度の向上が期待される。

BLUP 推定量はもともと畜産関係の分野で用いられてきた手法である。実際の応用例としては、アメリカの 1 世帯当たり収入のメディアン分析や、カナダセンサスの調査漏れ率の推定などがある。これらの例については、Rao(2003) に詳しく解説されている。EBLUP 推定量については、現在も活発な研究が行われている。

1 Fay-Herriot モデル

総数 m の小地域があり、Area Level の補助情報が手に入ると仮定する。このとき、データの背後にある構造として回帰モデルを考えて、観測値が以下のようなモデルに従うと仮定する。

$$y_i = \mathbf{x}_i^t \boldsymbol{\beta} + b_i v_i + e_i, \quad i = 1, 2, \dots, m \quad (1)$$

$$\left[\begin{array}{l} y_i : \text{第 } i \text{ 小地域における観測値} \\ \mathbf{x}_i : \text{第 } i \text{ 小地域における補助情報のベクトル (定数項を含む)} \\ b_i : \text{既知の定数、} e_i \sim N(0, \psi_i), v_i \sim N(0, \sigma_v^2) \end{array} \right]$$

e_i 及び v_i は、すべて独立であると仮定する。また σ_v^2 及び ψ_i はそれぞれ既知であると仮定する。各 i について y_i は独立であると仮定する。式 (1) は、このモデルを最初に考えた人たちの名前から、Fay-Herriot モデルと呼ばれる (Fay and Herriot(1979))。式 (1) のモデルは、以下のような二つのモデルを組み合わせたものである。

$$\theta_i = \mathbf{x}_i^t \boldsymbol{\beta} + b_i v_i, \quad i = 1, 2, \dots, m \quad (2)$$

$$y_i = \theta_i + e_i, \quad i = 1, 2, \dots, m \quad (3)$$

$$\left[\theta_i : \text{各小地域における真の値} \right]$$

式(2)では、真の値 θ_i が線形回帰モデルに従うことを表している。式 (3) は、真の値 θ_i に標本誤差 e_i が付加されて、実際の観測値 y_i として観測されることを表している。

通常の線形回帰モデルでは、式 (2) の $b_i v_i$ の項が無く、真の値 θ_i が回帰直線 $\mathbf{x}_i^t \boldsymbol{\beta}$ の上に完全に乗っていると仮定している。このような考え方は、実際のデータを考えた場合には現実的でないことが多い。Fay-Herriot モデルでは線形回帰モデルと異なり、真の値 θ_i について、回帰直線から $b_i v_i$ の分だけ幅を持たせたモデルを考えている。これにより、地域の特徴を反映させた、より現実的な推定が可能になると考えられる。確率変数が複数あるこのような線形モデルのことを、一般に線形混合効果モデル (Mixed Random Effect Model) と呼ぶ。Fay-Herriot モデルは線形混合効果モデルの特殊な例である。

e_i 及び v_i は、ともに平均がゼロの確率変数である。しかしその分散に関する仮定には、以下のように大きな違いがある。

- ・ v_i は、地域 i によらない一定の分散 σ_v^2 を持つと仮定されている。小地域間の効果のばらつきは、既知の定数 b_i により v_i の外部から与えられている。本稿の分析では、全ての場合で $b_i = 1$ としている。しかし分析者が地域ごとの変動 v_i に関して先見的な情報を持っている場合には、それを基にして b_i の値を定めることもある。
- ・ e_i の分散 ψ_i には添え字 i がついている。よって σ_v^2 の場合と異なり、標本誤差 e_i の分散が小地域ごとに異なることを許容している。標本誤差の分散 ψ_i は既知であると仮定している。しかし実際は ψ_i の値は未知なので $\hat{\psi}_i$ として、個別データから推定した $\hat{\psi}_i$ を用いることとする。 $\hat{\psi}_i$ の推定方法については、第4章で詳しく解説している。

2 BLUP 推定量

θ_i の予測量 $\hat{\theta}_i = \mathbf{x}_i^t \hat{\boldsymbol{\beta}} + \hat{v}_i$ を線形不偏なクラスに限定した場合に、平均二乗誤差が最小となるものは、BLUP (Best Linear Unbiased Prediction: 最良線形不偏予測量) と呼ばれる。Henderson(1950) は、より一般的な線形混合モデルに対する BLUP を導出した。Henderson の理論を式(1)のモデルに当てはめることで、以下のようにして Fay-Herriot モデルの BLUP 推定量 θ_i^B を求めることができる。

$$\theta_i^B = \gamma_i \hat{y}_i + (1 - \gamma_i) \mathbf{x}_i^t \hat{\boldsymbol{\beta}} \quad (4)$$

$$\left[\begin{array}{l} \gamma_i = \frac{b_i^2 \sigma_v^2}{\psi_i + b_i^2 \sigma_v^2} \\ \hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^m \frac{\mathbf{x}_i^t \mathbf{x}_i}{\psi_i + b_i^2 \sigma_v^2} \right]^{-1} \left[\sum_{i=1}^m \frac{\mathbf{x}_i^t y_i}{\psi_i + b_i^2 \sigma_v^2} \right] \end{array} \right]$$

回帰係数 $\boldsymbol{\beta}$ の推定量 $\hat{\boldsymbol{\beta}}$ は、確率変数の分散を考慮した場合の一般化最小二乗回帰推定量となっている。

θ_i^B は、観測値 y_i と、単純な回帰モデルによる予測値 $\mathbf{x}_i'\boldsymbol{\beta}$ とを、ウェイトを γ_i として加重平均した形になっている。つまり BLUP 推定量は回帰直線と観測値を $\gamma_i : (1-\gamma_i)$ に内分する点になっている。よって先にも述べたとおり、真の値は回帰直線上に完全に乗ると仮定するのではなく、回帰直線と観測値の間にあると考えていることになる(図1参照)。

加重平均のウェイト γ_i は、 e_i と v_i の分散の比によって決まってくる。 γ_i の形から、以下のことが分かる。

- ・観測値の標本誤差 ψ_i が大きくなるに従って($\psi_i \rightarrow \infty$)、 γ_i はゼロに近づき、モデルによる予測値にかかるウェイトが大きくなる。
- ・反対に地域の効果の変動 σ_v^2 が大きくなるに従って($\sigma_v^2 \rightarrow \infty$)、 γ_i は1に近づき、観測値にかかるウェイトが大きくなる。
- ・ $\psi_i = 0$ のとき、加重平均のウェイト γ_i は σ_v^2 の値にかかわらず1となって、 θ_i^B は常に観測値 y_i に一致する。

これにより θ_i^B は、観測値の分散が小さく観測値が信頼できる場合には観測値のウェイトを大きくとり、そうでない場合はその程度に応じて適当な割合でモデルの情報を取り込むという、観測値の誤差に応じた非常に柔軟な推定値となっている。また、母集団すべてを調査したとき($\psi_i = 0$)には、 θ_i^B は母集団の値に一致することも分かる。

観測値 y_i は、例えば分散を安定させるための対数変換のような、適当な変換が行われていてもよいものとする。そのような場合には、EBLUP 推定量により推定された値を逆変換して、推定値のスケールをもとの観測値と合わせたものを、最終的な推定値とする。

EBLUP 推定量は、実は適当な事前分布を設定したときの経験的ベイズ推定量として導出できる。また EBLUP 推定量は、回帰予測へ向けて推定値を縮小させるスタインタイプの縮小推定量の特殊な場合であるととらえることも出来る(下平 他(2004))。

3 EBLUP 推定量

前節では σ_v^2 が既知であると仮定して、線形不偏なクラスの中で最良(分散が最小)な予測量 θ_i^B を導いた。しかし実際の推定では、 σ_v^2 の値が未知であり、その値を観測値のデータから何らかの方法で推定する必要がある。BLUP 推定量 θ_i^B において、 σ_v^2 の値を観測値のデータによ

る推定値 $\hat{\sigma}_v^2$ で置き換えたものを、EBLUP 推定量と呼ぶ。Fay-Herriot(1979)では、単純なモーメント法と簡略化したニュートン法を組み合わせ、 $\hat{\sigma}_v^2$ を推定している。

ここでは最尤法を用いた $\hat{\sigma}_v^2$ の推定について述べる。その場合には、対数尤度関数 $l(\hat{\sigma}_v^2)$ が最大値になるときの $\hat{\sigma}_v^2$ が、 σ_v^2 の推定値となる。対数尤度関数 $l(\hat{\sigma}_v^2)$ の具体的な形は以下のとおりである。

$$\begin{aligned}
 l(\hat{\sigma}_v^2) &= \log \left[\prod_{i=1}^m \frac{1}{\sqrt{2\pi (b_i^2 \hat{\sigma}_v^2 + \psi_i)}} \exp \left[-\frac{1}{2} \frac{(y_i - \mathbf{x}_i^t \boldsymbol{\beta}(\hat{\sigma}_v^2))^2}{b_i^2 \hat{\sigma}_v^2 + \psi_i} \right] \right] \\
 &= -\frac{m}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^m \log (b_i^2 \hat{\sigma}_v^2 + \psi_i) - \frac{1}{2} \sum_{i=1}^m \frac{(y_i - \mathbf{x}_i^t \boldsymbol{\beta}(\hat{\sigma}_v^2))^2}{b_i^2 \hat{\sigma}_v^2 + \psi_i} \\
 &\quad \left[\boldsymbol{\beta}(\hat{\sigma}_v^2) = \left[\sum_{i=1}^m \frac{\mathbf{x}_i^t \mathbf{x}_i}{\psi_i + b_i^2 \hat{\sigma}_v^2} \right]^{-1} \left[\sum_{i=1}^m \frac{\mathbf{x}_i^t y_i}{\psi_i + b_i^2 \hat{\sigma}_v^2} \right] \right]
 \end{aligned} \tag{5}$$

式(5)は複雑な非線形の方程式なので、最小値を解析的に計算することは不可能である。そこで数値計算による非線形最適化の方法を用いて、 $\hat{\sigma}_v^2$ の推定を行う。関数の最小化のための数値計算には、様々な方法がある。

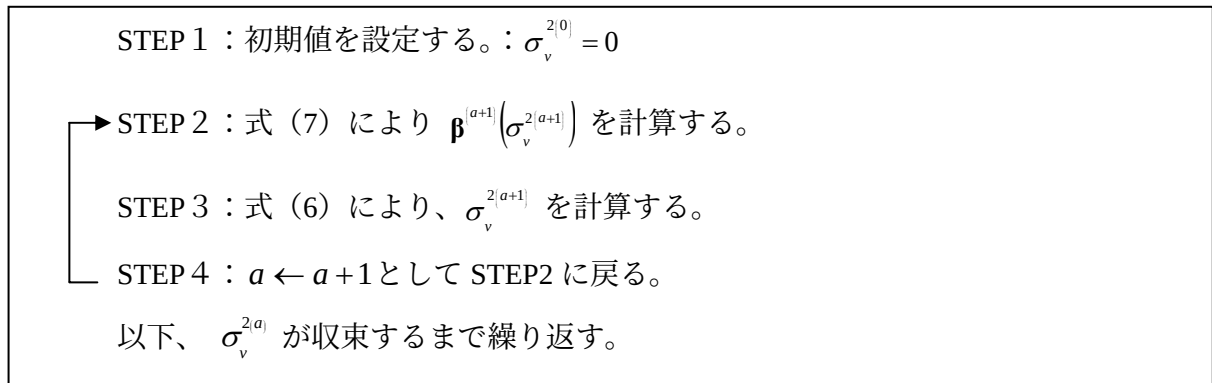
非線形最適化の方法として、スコア法を紹介する。これはニュートン・ラフソン法において、傾きを表す Hessian 行列をその期待値で置き換えた手法である。期待値をとることで被微分関数が簡単な形になり、計算が簡略化される。Hessian 行列の期待値は、Fisher の情報行列に等しくなる。この方法は Fisher によって初めて用いられたので、Fisher のスコア法とも呼ばれる (鈴木、山田 (2000))。式 (5) を最小化するためのスコア法のアルゴリズムは、次のようにして与えられる。詳しくは Rao(2003)を参照のこと。

$$\sigma_v^{2(a+1)} = \sigma_v^{2(a)} + \left[I(\sigma_v^{2(a)}) \right]^{-1} s(\boldsymbol{\beta}^{(a)}, \sigma_v^{2(a)}) \tag{6}$$

$$\left[\begin{aligned}
 I(\sigma_v^{2(a)}) &= \frac{1}{2} \sum_{i=1}^m \frac{1}{(\psi_i + \sigma_v^{2(a)})^2} \quad : \text{Fisher の情報量} \\
 s(\boldsymbol{\beta}^{(a)}, \sigma_v^{2(a)}) &= -\frac{1}{2} \sum_{i=1}^m \frac{1}{\psi_i + \sigma_v^{2(a)}} + \frac{1}{2} \sum_{i=1}^m \frac{(\hat{\theta}_i - \mathbf{x}_i^t \tilde{\boldsymbol{\beta}})^2}{(\psi_i + \sigma_v^{2(a)})^2} \\
 \boldsymbol{\beta}^{(a)}(\sigma_v^{2(a)}) &= \left[\sum_{i=1}^m \frac{\mathbf{x}_i^t \mathbf{x}_i}{\psi_i + b_i^2 \sigma_v^{2(a)}} \right]^{-1} \left[\sum_{i=1}^m \frac{\mathbf{x}_i^t y_i}{\psi_i + b_i^2 \sigma_v^{2(a)}} \right]
 \end{aligned} \right] \tag{7}$$

式中の (a) は、繰り返し計算における反復の回数を表している。なお、計算を簡略化するために、まず β を固定して σ_v^2 に対して対数尤度の最適化を行い、そうして求めた σ_v^2 から β の値を求めるとようなステップを踏んでいる。以下のようにして、逐次的に計算を行う。ここでは、 $\sigma_v^{2(a)}$ の初期値を0としている。

スコア法のアルゴリズム



σ_v^2 に対する尤度関数の概形を示したものが、図2である。観測値には平成14年9月の都道府県別完全失業率を、回帰モデルの説明変数には同じ時期の都道府県別有効求人倍率を、それぞれ用いている。推定には式(1)で $b_i = 1$ としたモデルを用いた。対数尤度関数は非常に滑らかな曲線であり、 σ_v^2 は1.4付近で最大値を取っていることが確認できる。上記の対数尤度関数について数値計算を行い、最適値を求めてみる。計算において、 $\sigma_v^{2(a)}$ と $\sigma_v^{2(a+1)}$ との差の絶対値が $1e-6$ (0.000001)より小さくなったときに、計算を停止することにした。図3は、その数値計算の出力結果である。初期値を0として数値計算を行った結果、 σ_v^2 は6回の繰り返し計算の後に、1.41779という値に収束した。対数尤度関数が滑らかなので、初期値を10、10000に変更した場合でも、ほぼ同じ反復回数で同じ値に速やかに収束している。なお尤度関数の描画及びスコア法による数値計算に用いたプログラムについては、それぞれプログラム①、プログラム②としてこの章の最後に示してある。

σ_v^2 に特に条件を課さない場合には、数値計算の過程において、分散の推定値が負の値を取る可能性があり、最終的な推定値も負の値に収束する可能性がある。そこで、 $\hat{\sigma}_v^2 = \exp(\tau)$ と変換して、 τ を推定することが考えられる。ただしこの場合には式(5)の対数尤度関数は更に複雑になり、スコア法における微分を明示的に求めることは面倒になる。そのような場合には数値微分などを用いて推定を行うことで、負担が軽減される。Rの非線形最適化パッケージ `optim` では数値微分による準ニュートン法を含む、様々な最適化の手法を利用できる。

先ほどのデータに対して $\hat{\sigma}_v^2 = \exp(\tau)$ と変換して準ニュートン法による最適化を行ったところ、2 回という少ない反復回数で、 $\hat{\sigma}_v^2$ はスコア法と同じ値の 1.41779 に収束した。この方法による推定のプログラムについては、プログラム③として本章の最後に示してある。

4 MSE の推定量

EBLUP 推定量を θ_i^E とする。MSE(平均二乗誤差)の推定量には多くの種類があり、 $\hat{\sigma}_v^2$ をどのような方法で推定したかによって、MSE の推定方法が変わってくる。各推定量の間の比較等の研究も数多く行われている(Rao(2003))。式 (1) で $b_i = 1$ として、 $\hat{\sigma}_v^2$ を最尤法で推定した場合には、その MSE の推定量を以下のようにして求めることができる(Rao(2003))。

$$MSE(\theta_i^{BLUP}) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) \quad (8)$$

$$\left[\begin{array}{l} g_{1i}(\hat{\sigma}_v^2) = \frac{\hat{\sigma}_v^2 \psi_i}{\psi_i + \hat{\sigma}_v^2} (= \gamma_i \psi_i) \\ g_{2i}(\hat{\sigma}_v^2) = (1 - \gamma_i)^2 \mathbf{x}_i^t \left[\sum_{i=1}^m \frac{\mathbf{x}_i^t \mathbf{x}_i}{\psi_i + \hat{\sigma}_v^2} \right]^{-1} \mathbf{x}_i \\ \bar{V}(\hat{\sigma}_v^2) = [I(\hat{\sigma}_v^2)]^{-1} = 2 \left[\sum_i \frac{1}{(\psi_i + \hat{\sigma}_v^2)^2} \right]^{-1} \end{array} \right]$$

式(8)の主要な項 $g_{1i}(\hat{\sigma}_v^2)$ の、標本数 m に対するオーダーは、 $O(1)$ である。式(8)の第 2 項のオーダーは回帰係数 β の推定値に関係している。以下の正則条件を仮定すると、式(8)の第 2 項のオーダーは $O(m)$ となる(Rao(2003))。

・ ψ_i 及び b_i が一様有界である。

$$\cdot \sup \left(\mathbf{x}_i^t \left[\sum_{i=1}^m \mathbf{x}_i^t \mathbf{x}_i \right]^{-1} \mathbf{x}_i \right) = O(m^{-1})$$

式 (8) の中の $\bar{V}(\hat{\sigma}_v^2)$ は、推定量 $\hat{\sigma}_v^2$ の漸近分散である。②の仮定は、標準的な回帰分析ならば通常満たされていると考えてよい。MSE の推定値の平方根を EBLUP 推定値で割った $\sqrt{MSE(\theta_i^{BLUP})} / \theta_i^{BLUP}$ を EBLUP 推定値の誤差率とし、推定量の比較の際の指標として用いる。

II 労働力調査への適用

1 推定方法及び結果の比較方法

I 節で紹介した EBLUP 推定量 θ_i^E の労働力調査への適用を考える。推定に用いる観測値データは、平成 3 年 1 月から平成 15 年 12 月までの 13 年間に於ける、都道府県別月次の完全失業率である。推定の枠組みは以下のとおりである。

- ・式(1)の Fay-Herriot モデルで $b_i = 1$ としたものを用いる。
- ・都道府県別の標本誤差 ψ_i には、第 4 章の方法で個票データから推定したものを用いる。
- ・最尤法で推定する際、非線形最適化では、 $\sigma_v^2 = \exp(\tau)$ と変換して、BFGS 公式による準ニュートン法を用いた推定を行う（I 節 3 参照）。
- ・回帰の部分の補助情報については、職業安定業務統計に於ける都道府県別月次の有効求人倍率を用いる。

比推定値と EBLUP 推定量による推定値に対して以下のような方法により比較を行い、結果を考察する。

- ・ベンチマークとの比較
- ・誤差率の比較
- ・時系列のグラフによる比較

ベンチマークとして用いるのは、以下の調査からそれぞれから計算した都道府県別の完全失業率である。

- ・平成 7 年国勢調査
- ・平成 12 年国勢調査
- ・平成 14 年就業構造基本調査

これらのベンチマークが真の値に近いものと想定し、ベンチマークとの距離を ARE (Absolute Relative Error) で測ることにより、比較を行った。ARE は以下の式で定義される。

$$ARE = \frac{|\theta_i^E - \theta_i^T|}{\theta_i^T} \quad (9)$$

$$\left[\theta_i^T : \text{ベンチマークの値} \right]$$

ベンチマークとなる調査の行われた年における、9 月の EBLUP 推定量と観測値で、どちらがベンチマークに近いかを比較する。なおベンチマークについては、小泉（2004）の方法により、全国の結果同士が合うように調整を施している。

2 完全失業率と補助情報との相関関係

完全失業率と関係が深いと思われる補助情報で、都道府県別に月次で入手できる統計のうち、主なものとして有効求人倍率が考えられる。有効求人倍率は、完全失業率と負の相関があると考えられる。有効求人倍率の詳細については、第 1 章 IV 節 2 を参照のこと。

まず小地域の完全失業率の分析に先立って、各説明変数と完全失業率との相関についてまとめる。平成 7 年、12 年、14 年の結果をグラフにまとめたものが、図 4 である。これらのグラフから完全失業率と有効求人倍率との相関は比較的高くなっていることが分かる。単純な回帰直線を当てはめた場合には、定数と係数のどちらの p 値も低い値となっている。しかし時間がたつにつれて、これらの変数間の相関が低くなる傾向が見られる。よってデータの前半の方が、後半に比べて当てはまりが良いと言える。

3 結果の考察

(1) 各種推定値の比較

平成 7 年 9 月、平成 12 年 9 月、平成 14 年 9 月における各種推定値を表の形にまとめたものが表 1-1~1-3 である。また、それらをグラフにしたものが図 5~7 である。これらの結果から、以下のことが分かる。

- ・観測値と EBLUP 推定量の値を直接比較する。両方の推定値にあまり違いはなく、良く似た結果となっている。しかし山梨県（都道府県番号 19 番）、鳥取県（31 番）、島根県（32 番）、佐賀県（41 番）といったところでは、両者の違いが大きくなっている。これらの都道府県は標本数が少なく、標本誤差が大きいと考えられる。よって EBLUP 推定量で加重平均の際に、回帰予測にかかるウエイトが大きくなっていると考えられる。EBLUP 推定量が比推定値を空間的に平滑化している様子については、グラフによく表れている。
- ・EBLUP 推定量で加重平均の際に観測値にかかるウエイト γ_i を見てみると、比推定の誤差がもともと低い北海道（1 番）、東京都（13 番）、神奈川県（14 番）、愛知県（23 番）、大阪府（27 番）、沖縄県（47 番）などについては、観測値にかかるウエイトが大きくなっている。これらの県は人口規模が大きく標本数も多いので、ある程度正確な推定値を得られることから、観測値にかかるウエイトが大きくなったものと思われる。標本数が少なく比推定の誤差の大きい山梨県、島根県、徳島県（36 番）、佐賀県、宮崎県（45 番）などについては、逆に観測値にかかるウエイトが小さくなっている。もとの比推定値の精度が低いため、モデルによる予測値の影響が大きい結果となっている。平成 7 年では大半の小地域で観測値のウエイトが 0.1~0.4 の範囲であった。しかし時間とともにウエイトの値は増加し、平成 14 年では全都道府県の 8 割以上が 0.6 以上のウエイトとなっている。補助情報と完全失業率との相関関係は年々低くなってきているためにモデルによる推定値の変動が増し、モデルの信頼度が低下したために、観測値にかかるウエイトが時間とともに大きくなったと考えることができる。

- ・ ARE を比較する。比推定値よりも EBLUP 推定量の方が ARE の値が低いところには、セルに色をつけて表示している。EBLUP 推定量が比推定値よりもベンチマークに近いところの方が、どの時期においても多くなっている。ただし平成 7 年の新潟県（15 番）、三重県（24 番）、滋賀県（25 番）、香川県（37 番）、沖縄県、では EBLUP 推定量の ARE の方が比推定値の ARE と比較して、かなり大きくなってしまっている。新潟県、三重県、滋賀県では、比推定にかかるウエイトがかなり小さくなっており、回帰予測への過剰な縮小が起きている可能性がある。
- ・ 誤差率を比較する。すべての都道府県において EBLUP 推定量の方が、比推定に比べて全体的に低い誤差率となっている。推定値同士の差はそれ程無いものの、誤差率で見た場合にはかなりの改善が見られる。特に鳥取県や島根県を始めとした標本数の少ない県では、いずれの時期においても、かなり改善している。ただし両者の差は後の時期にいくほど縮まってきており、改善の程度は時間とともに縮小してきている。補助情報と完全失業率との相関関係の低下は、誤差率の改善の度合いに、強く影響を与えていると考えられる。

(2) 時系列のグラフによる比較

時系列のグラフをまとめたものが図 8-1~8-6 である。これにより、以下のことが分かる。

- ・ 大阪府や香川県、沖縄県などでは、EBLUP 推定量がすべて、比推定値の下方に来るという現象が起きてしまっている。このような地域については、回帰直線からのばらつきが一定であるという仮定が妥当ではないと考えられる。EBLUP 推定量は地域の情報で回帰を行うので、全国の平均的な値よりも常に大きい値が出やすい小地域については、常に下方に引っ張られる可能性が高いといえる。またその様な高い値の周辺の値を取るような小地域についても、これらの高い値の影響を受けて結果がゆがんでいる可能性がある。各都道府県に関する先見的な情報を上手く取り込むことにより、Fay-Herriot モデルの b_i の部分を外部から与えて調整するといった対処が考えられる。
- ・ 宮城県、群馬県、京都府、徳島県では、平成 3 年から平成 12 年ごろまでの当てはまりは良いものの、それ以降の期間、特に平成 15 年以降にかけて当てはまりが悪くなっている。これらの県では、この時期に相関関係の変化が起きている可能性がある。この時点は、完全失業率の全国結果が大きく上昇した時期でもある。
- ・ 佐賀県の前系列において、平成 10 年 9 月に大幅な落ち込みがある。直後に元の水準に戻しており、この月の完全失業率は外れ値であると考えられる。このはずれ値の影響を強く受けて、この時期の EBLUP 推定量は、強く下に引っ張られている。

(3) まとめ

推定結果の比較についてまとめる。有効求人倍率を用いたときのモデルの当てはまりは、全体的に見て、良いように思える。誤差率はずの観測値と比較して低くなっており、また比推定よりもベンチマークに近い都道府県の割合も高くなっている。ほとんどの小地域については、EBLUP 推定量の方が良い結果であるといえる。

ただし観測値との相関関係は年々悪化してきており、その影響が推定値にも表れている。EBLUP 推定量は補助情報と観測値との関係に大きく影響するので、相関関係の悪化により、今後モデルの当てはまりが悪くなる可能性が高い。精度の高い小地域推定を行うためには、完全失業率との相関の高い、良質な補助情報を利用することが重要である。

大阪府や沖縄県など一部の県については、観測値よりも常に低めに評価されるという状況が確認された。完全失業率の小地域ごとのばらつきがほぼ一定であるという仮定 ($b_i = 1$) は妥当でない可能性が高い。このような点を考慮し、より良い説明変数を用いたり、外れ値と思われる県を除いたりするなど、適当な修正を施すことで、推定精度が向上すると考えられる。

III 今後の課題

1 Compromise EBLUP 推定量の適用

Fay and Ferriot(1979)は、Compromise EBLUP 推定量 ${}^*\theta_i^{BLUP}$ の利用を推奨している。Compromise EBLUP 推定量 ${}^*\theta_i^{BLUP}$ は、EBLUP 推定量 θ_i^{BLUP} と観測値の標本誤差 ψ_i を基にして、次のようにして定義される。

STEP 1 : 通常の EBLUP 推定量 θ_i^{BLUP} を求める。

STEP 2 : 定数 C を適当に定める。

STEP 3 : C 、 θ_i^{BLUP} 、 ψ_i から ${}^*\theta_i^{BLUP}$ を以下のように定義する。

$$\left\{ \begin{array}{ll} \text{if } \theta_i^{BLUP} \in [y_i - c\sqrt{\psi_i}, \hat{\theta}_i + c\sqrt{\psi_i}] & \text{then } {}^*\theta_i^{BLUP} = \theta_i^{BLUP} \\ \text{if } \theta_i^{BLUP} < y_i - c\sqrt{\psi_i} & \text{then } {}^*\theta_i^{BLUP} = y_i - c\sqrt{\psi_i} \\ \text{if } \theta_i^{BLUP} > y_i + c\sqrt{\psi_i} & \text{then } {}^*\theta_i^{BLUP} = y_i + c\sqrt{\psi_i} \end{array} \right.$$

${}^*\theta_i^{BLUP}$ は、観測値の標準誤差 ψ_i を基にして、ある一定のところで回帰への縮小を打ち切る推定量である。回帰予測への縮小が大きい場合にはそれを制限して、各地域の特徴を取り出すことができると考えられる。 C の値を小さくすれば、それだけ縮小の度合いが制限されることになる。この推定量では、 C の値の決め方に客観的な方法が無く、その選択が恣意的になるという問題点もある。

2 Unit Level の補助情報による解析

本稿では Area Level のモデルによる分析を行ってきた。しかしより細かい Unit Level の補助情報が得られる場合(都道府県別で、さらに男女や年齢別のデータが手に入る場合など)には、Unit Level の線形混合効果モデルを用いることにより、更に精密な推定を行うことができると期待さ

れる。また都道府県のデータを先見的な情報によって特徴あるグループに分けておき、第1段階をグループ、第2段階を都道府県ととらえることで、Unit Level の解析を行うことも可能である。このようなモデルを用いて、全国消費実態調査を補助情報とした場合の家計調査の教育費年平均を解析した結果が、篠崎（1991）にまとめられている。

3 多変量 Fay-Herriot モデルの適用

様々な就業状態の推定値などをベクトルとしてまとめて、多変数の Fay-Herriot モデルによる推定を行うことが考えられる。変数間に相関がある場合には、多変量 Fay-Herriot モデルを用いることで、そのような相関を上手く組み込んだ推定を行うことができると期待される。Rao(2003)では、多変量 Fay-Herriot モデルを用いた1世帯当たりの収入のメディアンを推定する研究について紹介されている。その研究における分析では、本来推定したい変量とは別に、直接的には関係の無い変量も同時に推定して、その変数間の相関を利用して精度を高めている。

変数を時系列方向に拡張して時間領域と空間領域の両方の情報を取り込んだ推定を行う方法も考えられる。これは Cross Sectional and Time Series モデルと呼ばれており、本稿の第3章では、その一種である Rao-Yu モデルを適用した場合の結果についてまとめている。

IV 図表

図 1 BLUP 推定量の意味

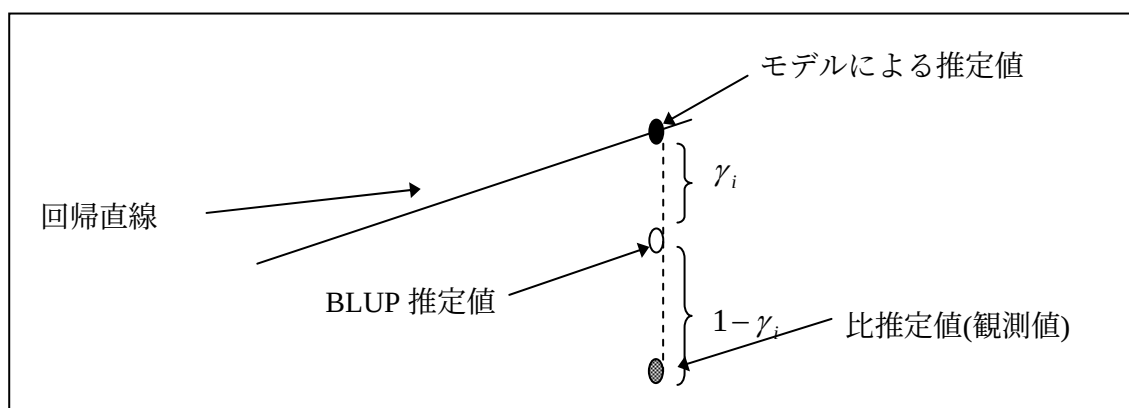


図 2 対数尤度関数のグラフ

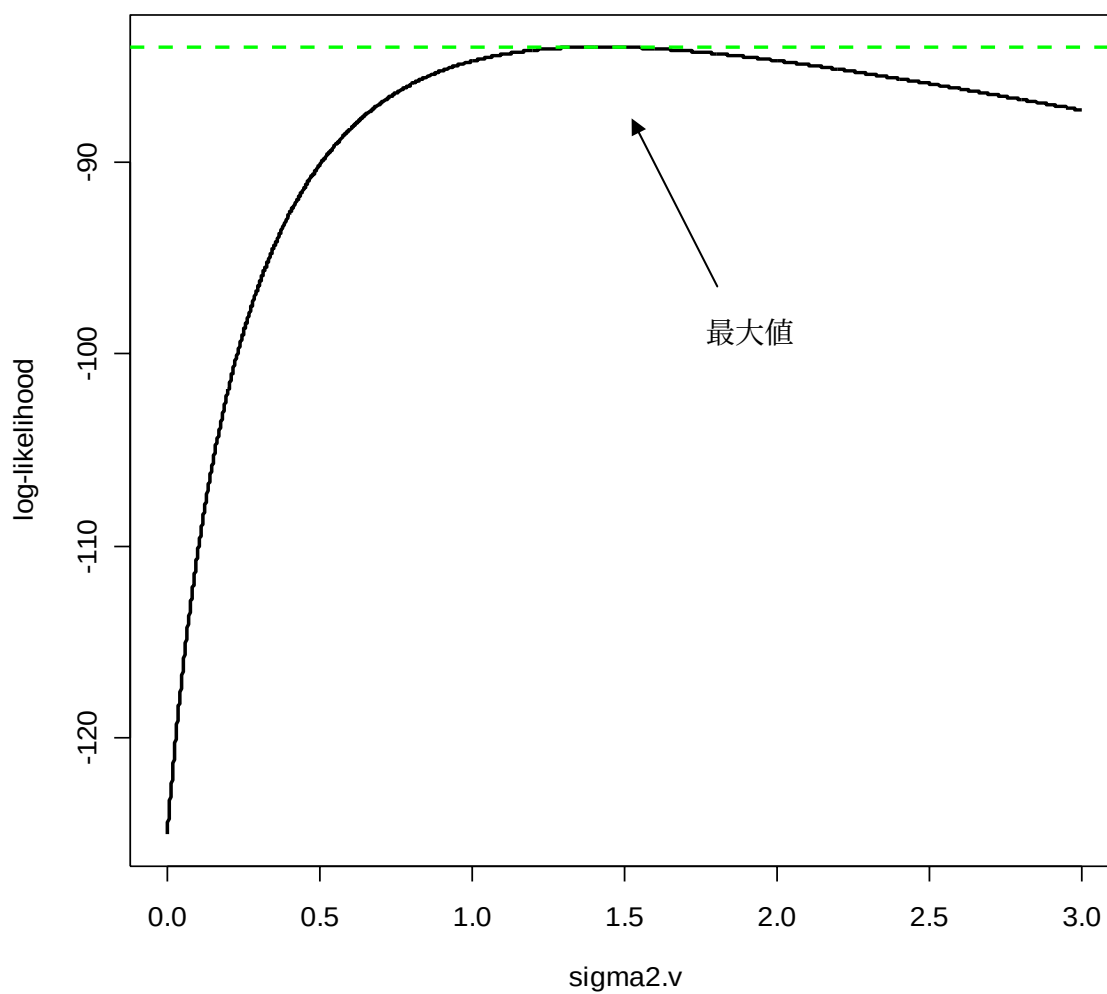


図 3 数値計算の出力結果

初期値：0

0	:	0.000000	-125.023261
1	:	1.327561	-84.042268
2	:	1.422290	-84.016196
3	:	1.417573	-84.016136
4	:	1.417801	-84.016136
5	:	1.417790	-84.016136
6	:	1.417790	-84.016136

初期値：10

0	:	10.000000	-103.708720
1	:	1.293726	-84.066807
2	:	1.424051	-84.016253
3	:	1.417488	-84.016136
4	:	1.417805	-84.016136
5	:	1.417790	-84.016136
6	:	1.417790	-84.016136

初期値：10000

0	:	10000.000000	-259.640211
1	:	1.181591	-84.216567
2	:	1.430198	-84.016595
3	:	1.417193	-84.016137
4	:	1.417819	-84.016136
5	:	1.417789	-84.016136
6	:	1.417790	-84.016136
7	:	1.417790	-84.016136

※ 各表とも左から順に、反復回数、 σ_v^2 の値、対数尤度

図 4 都道府県別 各種推定値の比較

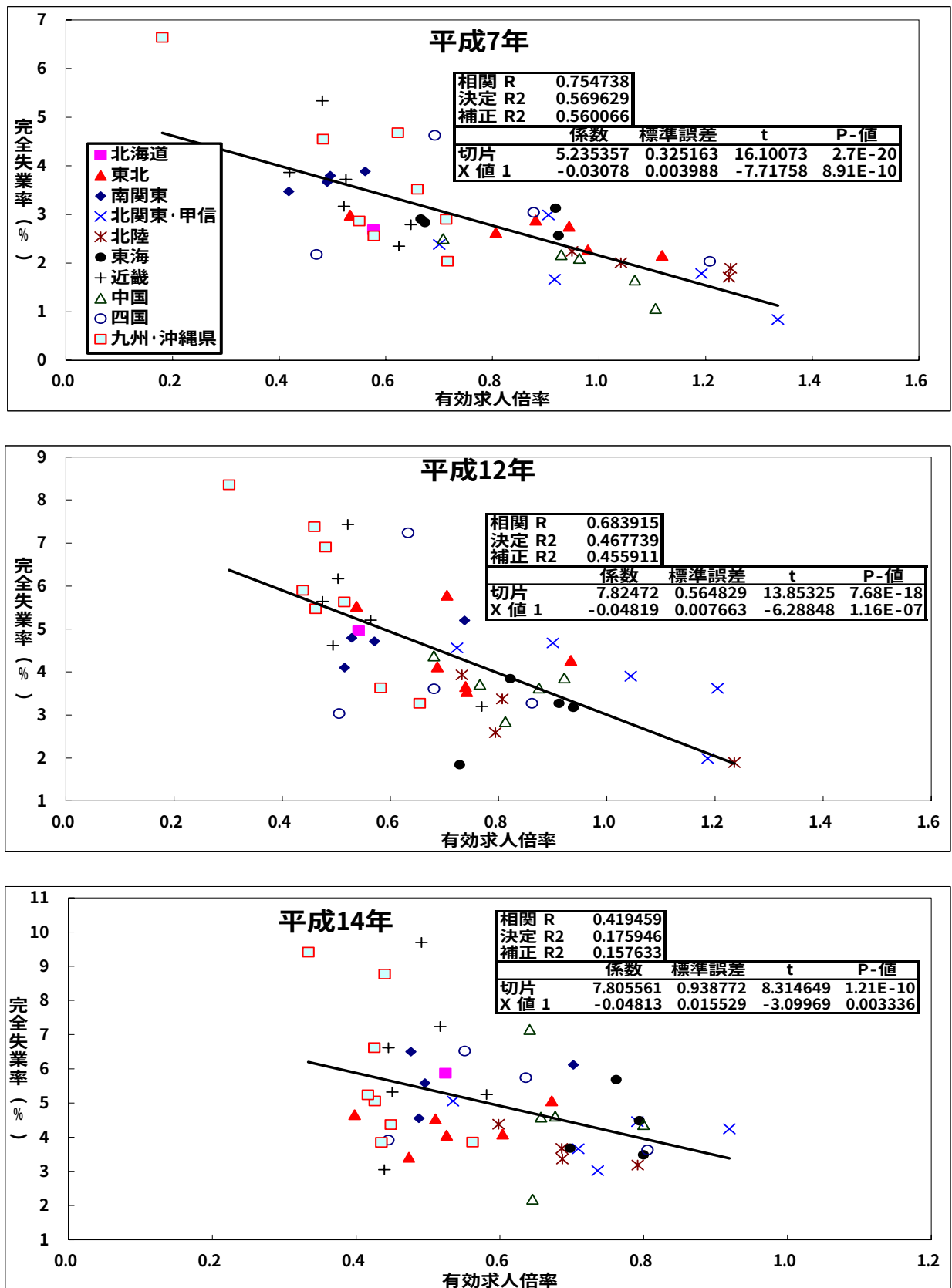


図 5 都道府県別 各種推定値の比較

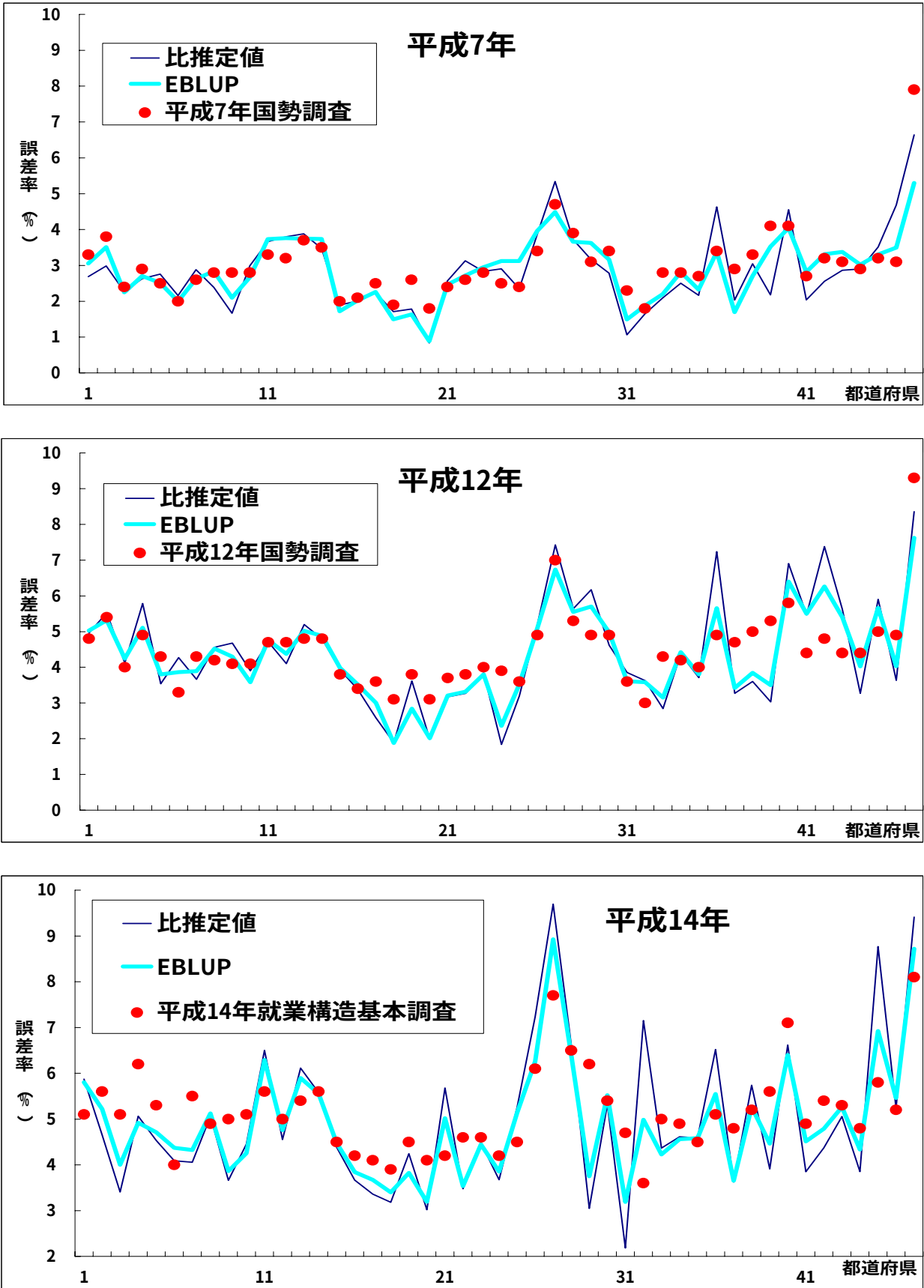


図 6 都道府県別 ARE の比較

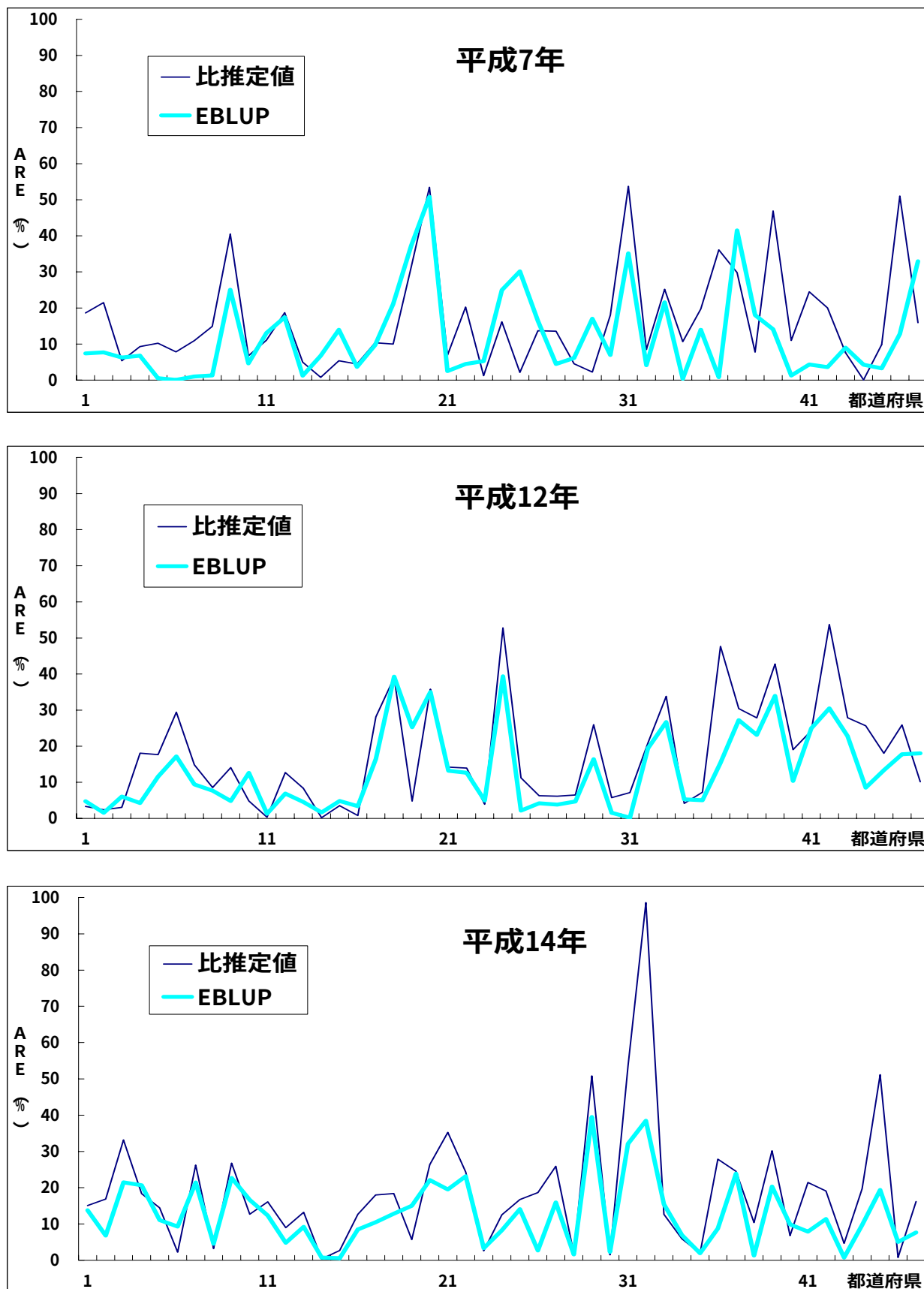


図 7 都道府県別 誤差率の比較

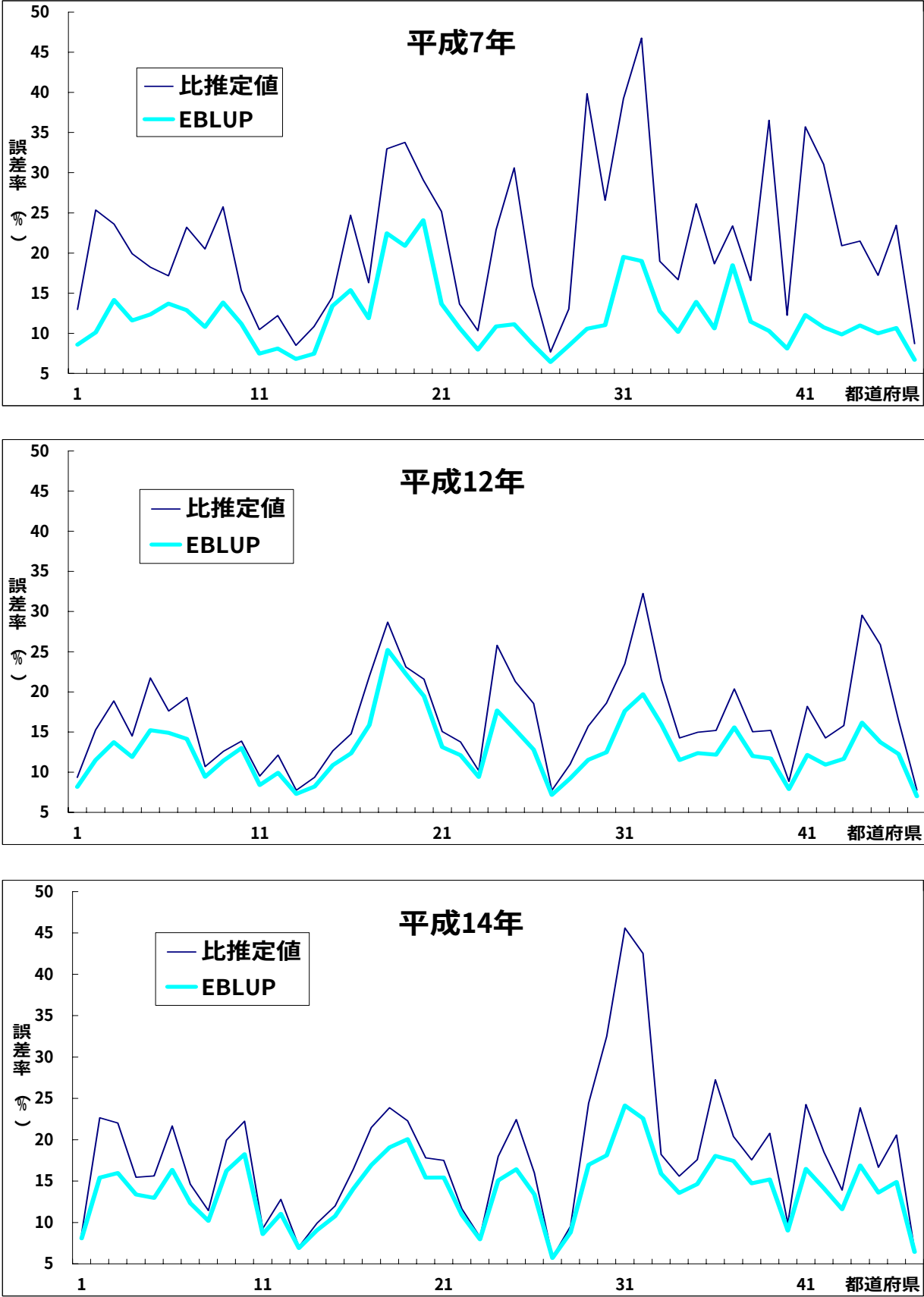


図 8-1 時系列のグラフ（北海道～茨城県）（単位：％）

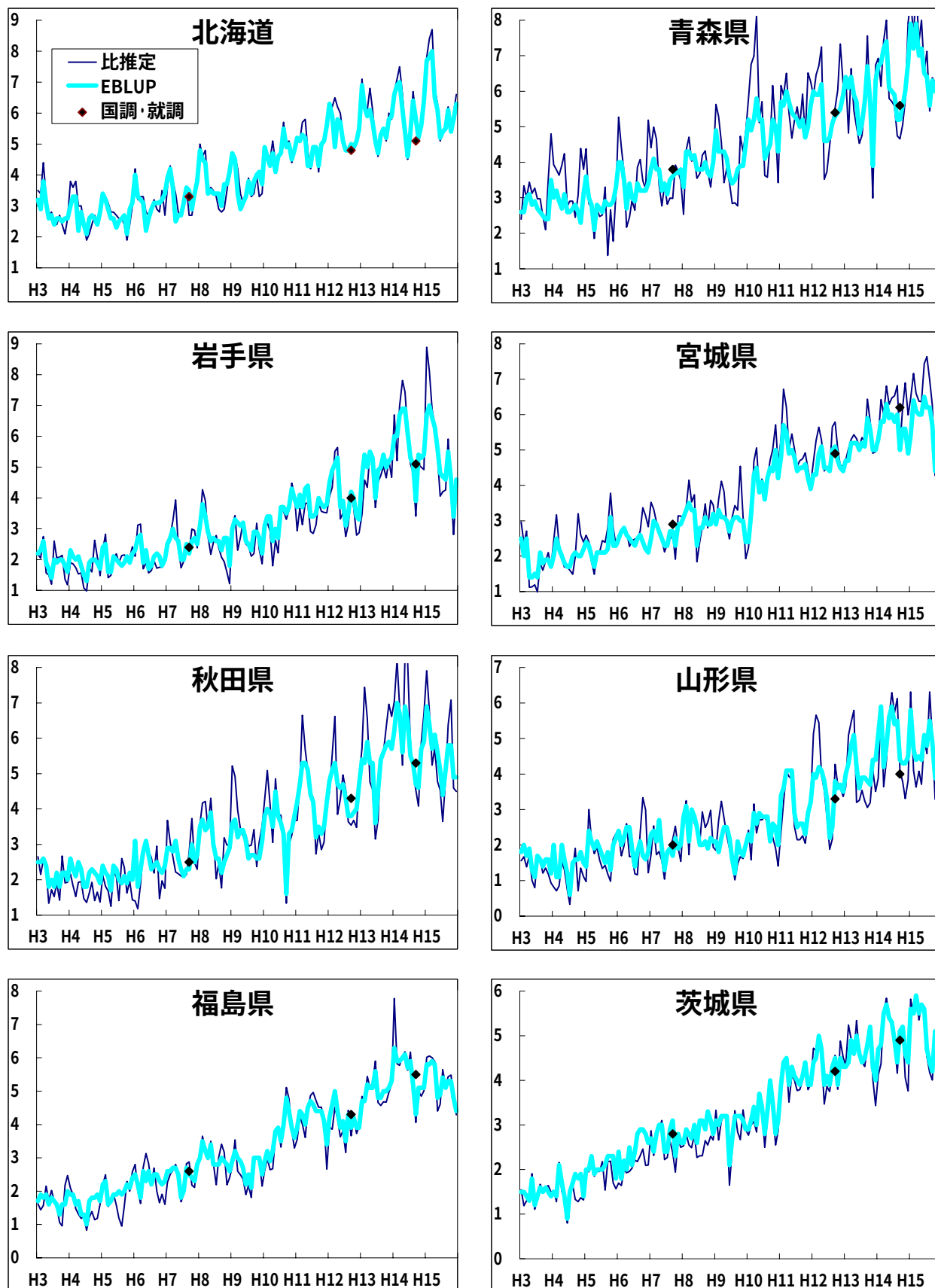


図 8-2 時系列のグラフ（栃木県～富山県）（単位：％）

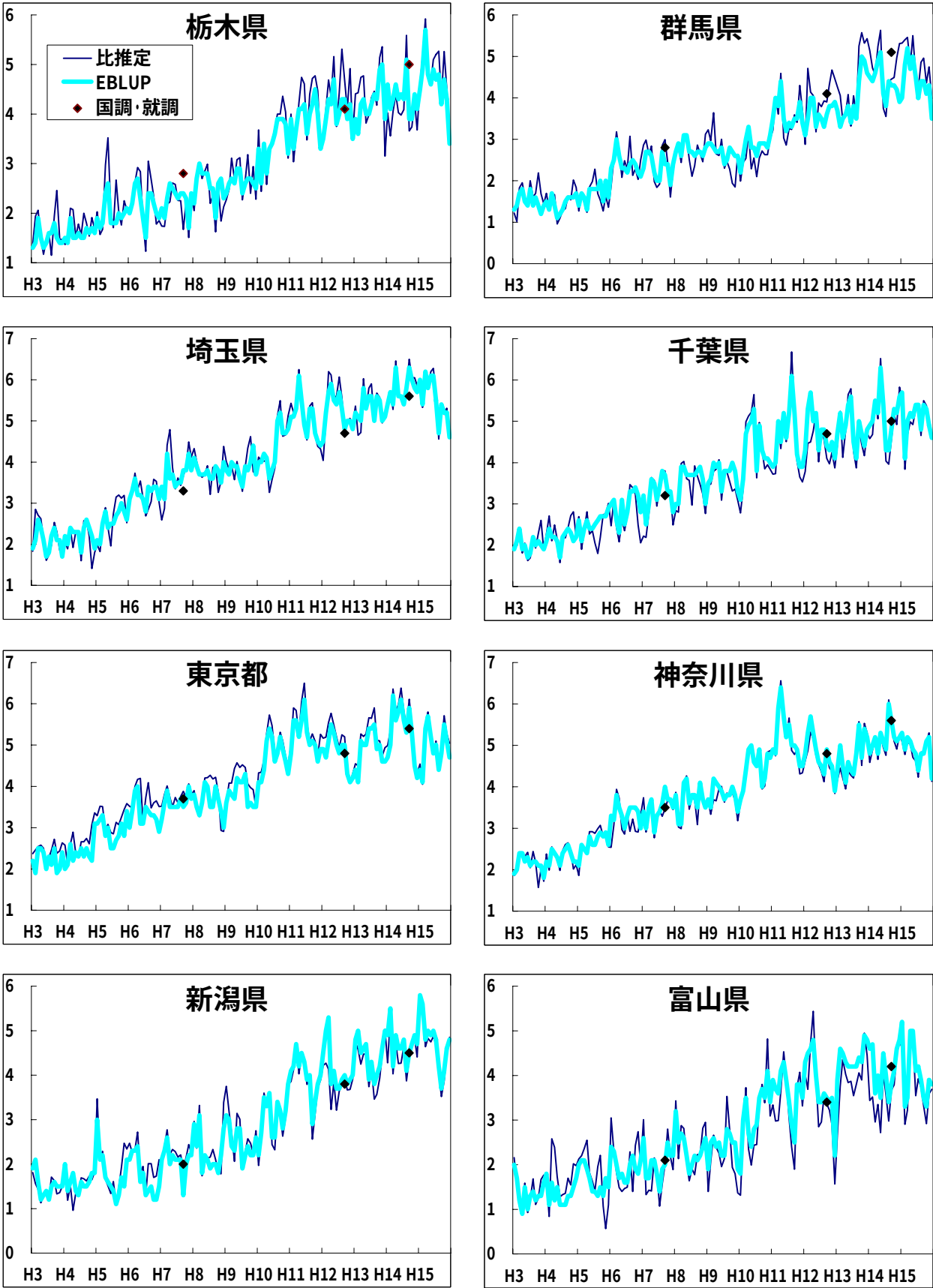


図 8-3 時系列のグラフ (石川県～三重県) (単位：%)

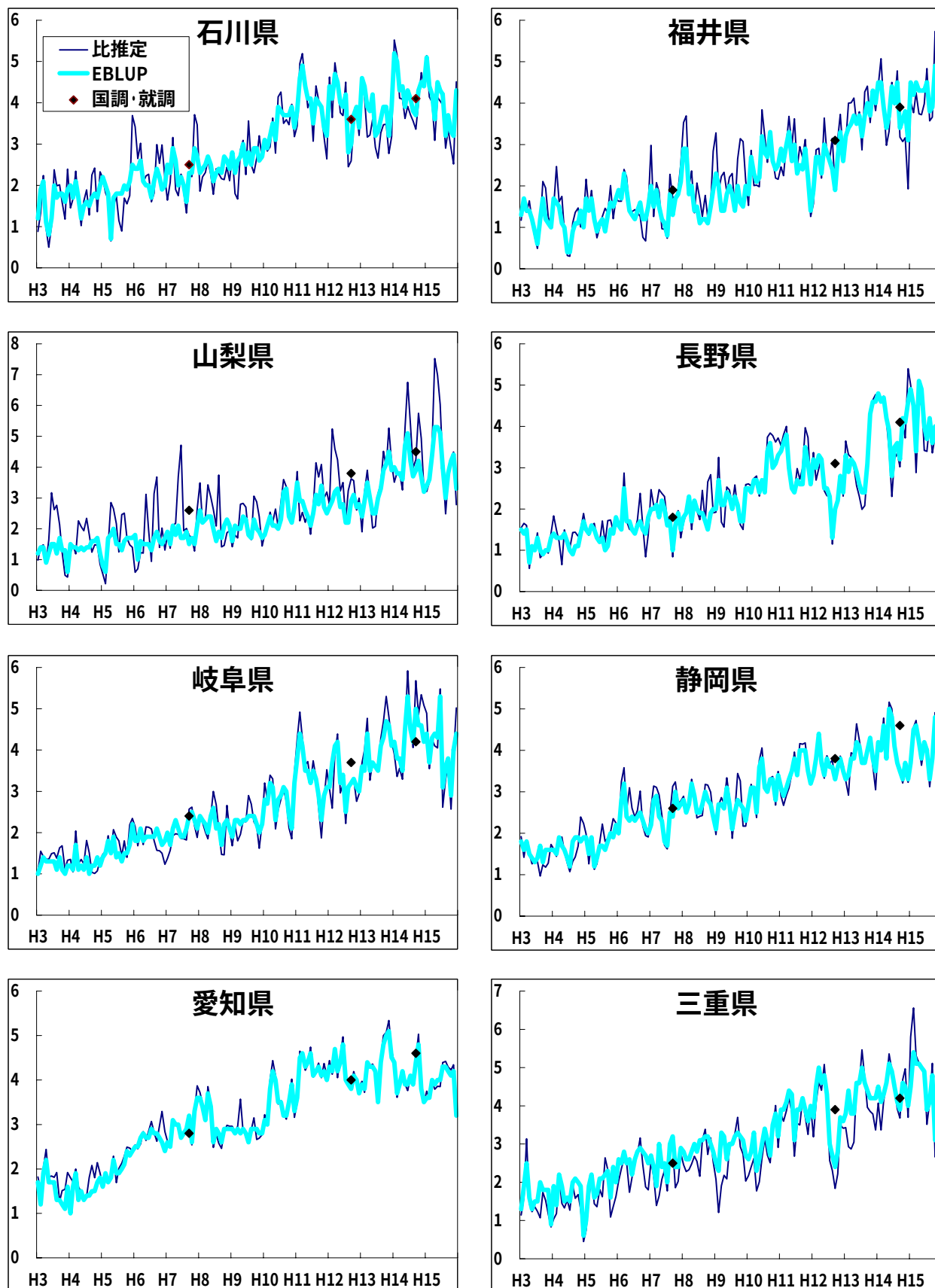


図 8-4 時系列のグラフ（静岡県～島根県）（単位：％）

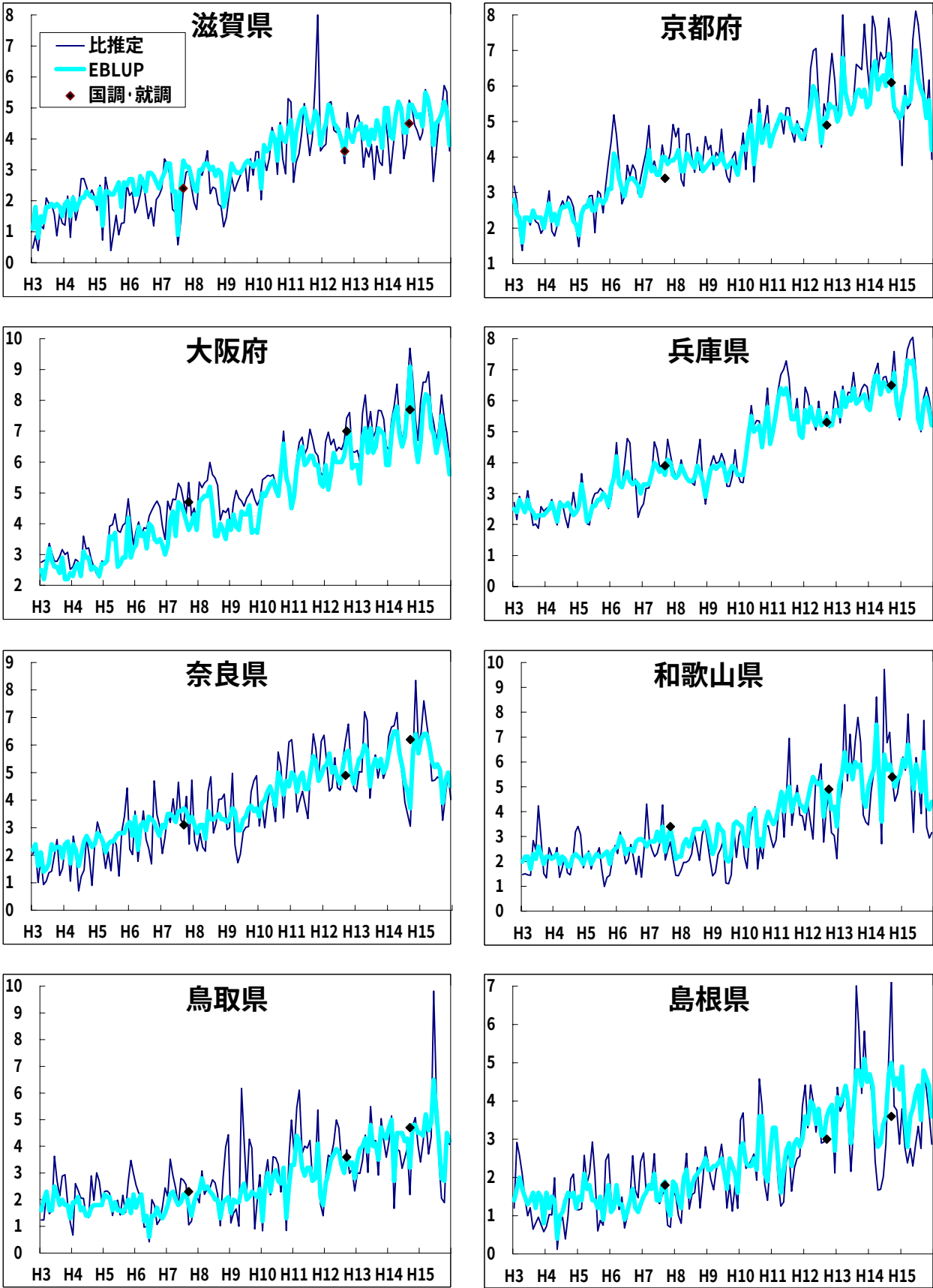


図 8-5 時系列のグラフ（岡山県～福岡県）（単位：％）

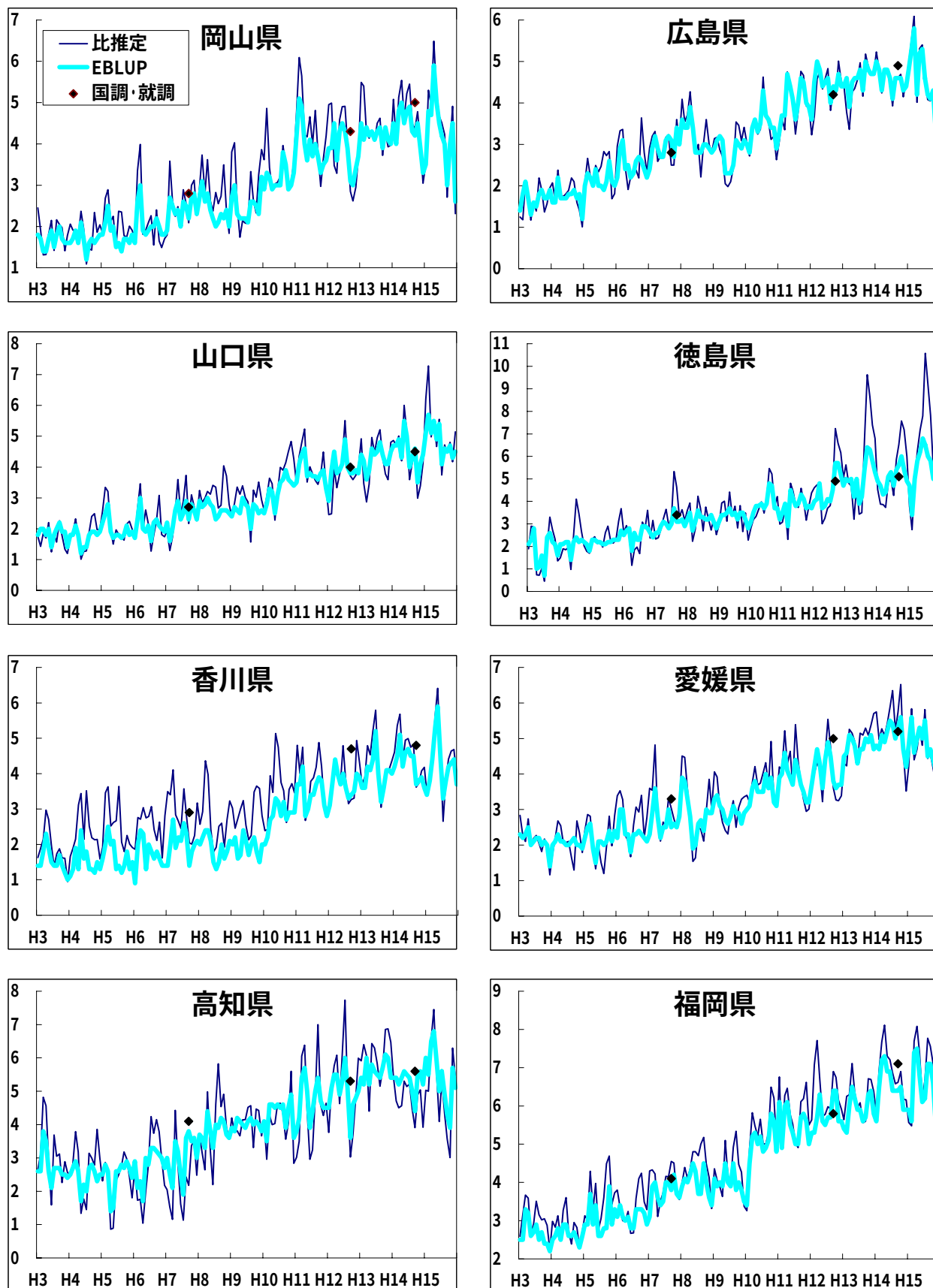


図 8-6 時系列のグラフ（佐賀県～沖縄県）（単位：％）

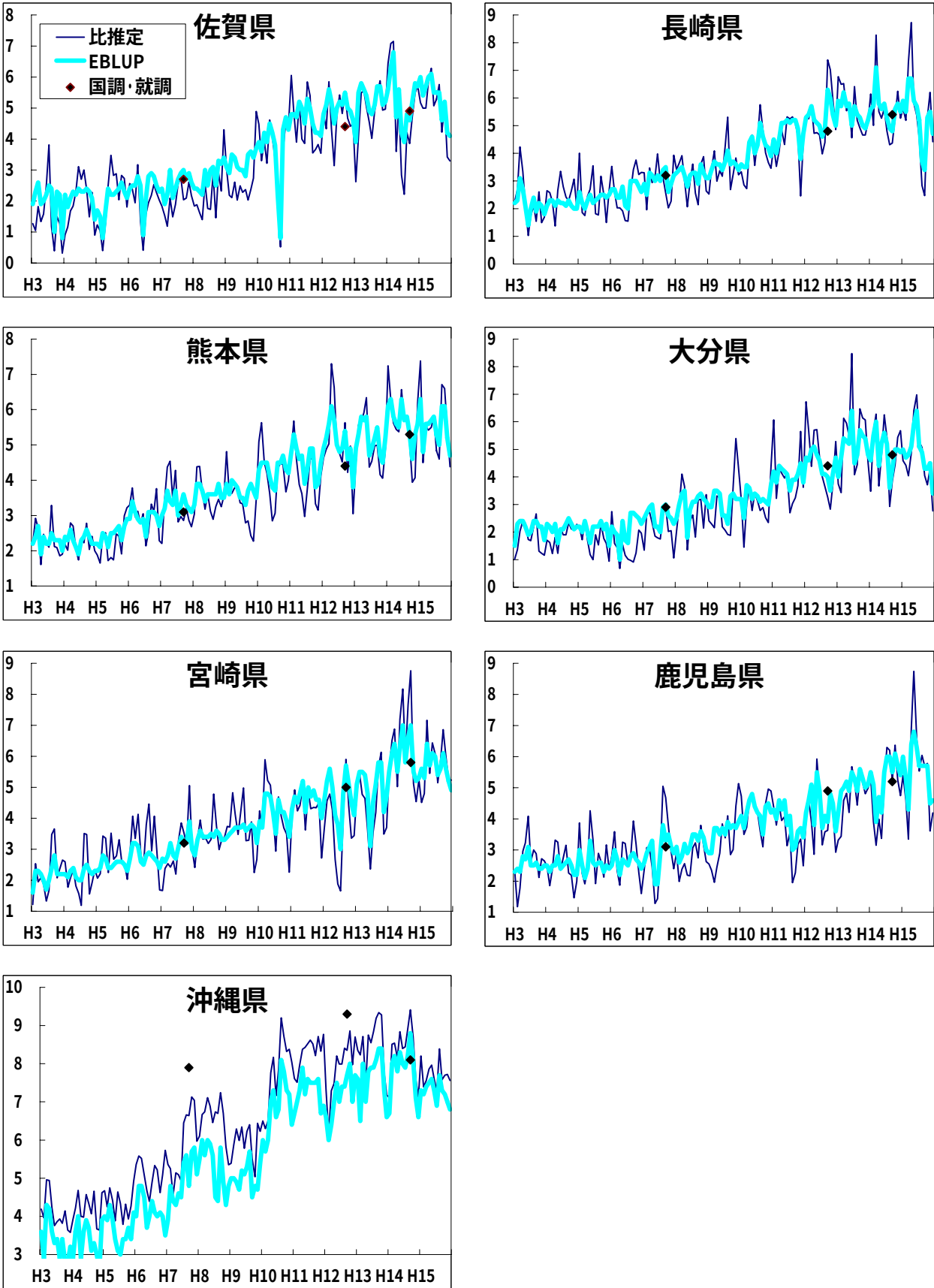


表 1-1 平成 7 年結果 (単位: %)

		ベンチ マーク (%)	推定値 (%)		ARE (%)		誤差率 (%)		比推定 値にかか るウエイ
		H7国調	比推定	EBLUP	比推定	EBLUP	比推定	EBLUP	
1	北海道	3.3	2.7	3.1	18.6	7.4	13.0	8.6	0.5
2	青森県	3.8	3.0	3.5	21.5	7.7	25.4	10.1	0.2
3	岩手県	2.4	2.3	2.2	5.4	6.3	23.6	14.2	0.3
4	宮城県	2.9	2.6	2.7	9.3	6.8	19.9	11.6	0.3
5	秋田県	2.5	2.8	2.5	10.2	0.5	18.2	12.3	0.4
6	山形県	2.0	2.2	2.0	7.8	0.1	17.2	13.7	0.5
7	福島県	2.6	2.9	2.6	10.9	1.0	23.2	12.9	0.2
8	茨城県	2.8	2.4	2.8	14.9	1.3	20.5	10.8	0.4
9	栃木県	2.8	1.7	2.1	40.5	25.1	25.7	13.8	0.4
10	群馬県	2.8	3.0	2.7	6.7	4.6	15.4	11.2	0.4
11	埼玉県	3.3	3.7	3.7	11.0	13.0	10.5	7.5	0.5
12	千葉県	3.2	3.8	3.8	18.7	17.4	12.2	8.1	0.4
13	東京都	3.7	3.9	3.7	5.0	1.2	8.5	6.8	0.6
14	神奈川県	3.5	3.5	3.7	0.8	6.8	10.8	7.5	0.5
15	新潟県	2.0	1.9	1.7	5.4	14.0	14.5	13.4	0.7
16	富山県	2.1	2.0	2.0	4.4	3.7	24.7	15.4	0.4
17	石川県	2.5	2.2	2.3	10.4	9.7	16.3	11.9	0.5
18	福井県	1.9	1.7	1.5	10.0	21.1	33.0	22.5	0.3
19	山梨県	2.6	1.8	1.6	31.3	37.2	33.8	20.9	0.3
20	長野県	1.8	0.8	0.9	53.4	50.8	29.1	24.1	0.7
21	岐阜県	2.4	2.6	2.5	6.9	2.5	25.2	13.7	0.3
22	静岡県	2.6	3.1	2.7	20.3	4.5	13.6	10.7	0.4
23	愛知県	2.8	2.8	2.9	1.2	5.3	10.3	8.0	0.6
24	三重県	2.5	2.9	3.1	16.2	24.9	22.9	10.9	0.2
25	滋賀県	2.4	2.3	3.1	2.1	30.1	30.6	11.1	0.2
26	京都府	3.4	3.9	4.0	13.7	16.4	15.9	8.7	0.3
27	大阪府	4.7	5.3	4.5	13.6	4.5	7.7	6.4	0.5
28	兵庫県	3.9	3.7	3.7	4.6	6.2	13.0	8.5	0.4
29	奈良県	3.1	3.2	3.6	2.2	17.0	39.8	10.6	0.1
30	和歌山県	3.4	2.8	3.2	18.0	7.0	26.6	11.0	0.2
31	鳥取県	2.3	1.1	1.5	53.7	35.2	39.2	19.5	0.5
32	島根県	1.8	1.6	1.9	8.4	4.1	46.8	19.0	0.2
33	岡山県	2.8	2.1	2.2	25.2	21.5	19.0	12.8	0.5
34	広島県	2.8	2.5	2.8	10.7	0.4	16.7	10.2	0.5
35	山口県	2.7	2.2	2.3	19.7	13.9	26.1	13.9	0.3
36	徳島県	3.4	4.6	3.4	36.1	0.8	18.7	10.6	0.2
37	香川県	2.9	2.0	1.7	29.8	41.5	23.4	18.5	0.4
38	愛媛県	3.3	3.0	2.7	7.8	18.0	16.6	11.5	0.4
39	高知県	4.1	2.2	3.5	46.9	14.1	36.5	10.3	0.2
40	福岡県	4.1	4.6	4.0	11.0	1.3	12.2	8.1	0.3
41	佐賀県	2.7	2.0	2.8	24.5	4.3	35.7	12.3	0.2
42	長崎県	3.2	2.6	3.3	20.0	3.6	31.0	10.8	0.2
43	熊本県	3.1	2.9	3.4	7.6	9.0	20.9	9.9	0.3
44	大分県	2.9	2.9	3.0	0.1	4.3	21.5	11.0	0.3
45	宮崎県	3.2	3.5	3.3	9.9	3.3	17.2	10.0	0.3
46	鹿児島県	3.1	4.7	3.5	51.0	12.8	23.4	10.7	0.1
47	沖縄県	7.9	6.6	5.3	15.9	32.9	8.7	6.7	0.3

表 1-2 平成 12 年結果 (単位：%)

		ベンチ マーク (%)	推定値 (%)		ARE (%)		誤差率 (%)		比推定 値にかか るウエイ ト
		H12国調	比推定	EBLUP	比推定	EBLUP	比推定	EBLUP	
1	北海道	4.8	5.0	5.0	3.3	4.7	9.3	8.2	0.8
2	青森県	5.4	5.5	5.3	2.4	1.6	15.3	11.5	0.5
3	岩手県	4.0	4.1	4.2	3.0	6.1	18.9	13.7	0.6
4	宮城県	4.9	5.8	5.1	18.1	4.2	14.5	11.9	0.5
5	秋田県	4.3	3.5	3.8	17.6	11.6	21.7	15.2	0.6
6	山形県	3.3	4.3	3.9	29.4	17.1	17.6	14.9	0.6
7	福島県	4.3	3.7	3.9	14.8	9.4	19.3	14.1	0.6
8	茨城県	4.2	4.6	4.5	8.6	7.7	10.7	9.4	0.8
9	栃木県	4.1	4.7	4.3	14.1	4.8	12.6	11.4	0.7
10	群馬県	4.1	3.9	3.6	4.8	12.6	13.9	13.0	0.7
11	埼玉県	4.7	4.7	4.8	0.3	1.3	9.5	8.4	0.8
12	千葉県	4.7	4.1	4.4	12.7	6.9	12.1	9.9	0.8
13	東京都	4.8	5.2	5.0	8.3	4.6	7.7	7.3	0.8
14	神奈川県	4.8	4.8	4.9	0.2	1.6	9.4	8.2	0.8
15	新潟県	3.8	3.9	4.0	3.5	4.8	12.7	10.9	0.8
16	富山県	3.4	3.4	3.5	0.8	3.3	14.8	12.3	0.8
17	石川県	3.6	2.6	3.0	28.1	16.5	21.9	15.9	0.7
18	福井県	3.1	1.9	1.9	39.0	39.3	28.7	25.2	0.7
19	山梨県	3.8	3.6	2.8	4.8	25.3	23.1	22.2	0.5
20	長野県	3.1	2.0	2.0	35.8	35.0	21.6	19.5	0.8
21	岐阜県	3.7	3.2	3.2	14.2	13.2	15.0	13.1	0.8
22	静岡県	3.8	3.3	3.3	13.9	12.7	13.8	12.1	0.8
23	愛知県	4.0	3.8	3.8	3.9	5.0	10.2	9.4	0.8
24	三重県	3.9	1.8	2.4	52.8	39.4	25.8	17.7	0.8
25	滋賀県	3.6	3.2	3.5	11.2	2.1	21.3	15.3	0.6
26	京都府	4.9	5.2	5.1	6.3	4.2	18.6	12.8	0.4
27	大阪府	7.0	7.4	6.7	6.1	3.8	7.7	7.2	0.7
28	兵庫県	5.3	5.6	5.5	6.5	4.7	11.0	9.2	0.7
29	奈良県	4.9	6.2	5.7	26.0	16.4	15.7	11.5	0.4
30	和歌山県	4.9	4.6	5.0	5.8	1.6	18.6	12.5	0.5
31	鳥取県	3.6	3.9	3.6	7.2	0.2	23.5	17.6	0.5
32	島根県	3.0	3.6	3.6	20.9	19.5	32.2	19.7	0.4
33	岡山県	4.3	2.8	3.2	33.8	26.7	21.5	16.0	0.7
34	広島県	4.2	4.4	4.4	4.1	5.3	14.3	11.5	0.7
35	山口県	4.0	3.7	3.8	7.2	5.0	15.0	12.4	0.7
36	徳島県	4.9	7.2	5.7	47.7	15.4	15.2	12.2	0.4
37	香川県	4.7	3.3	3.4	30.4	27.2	20.4	15.6	0.6
38	愛媛県	5.0	3.6	3.8	27.9	23.1	15.0	12.0	0.7
39	高知県	5.3	3.0	3.5	42.8	33.9	15.2	11.7	0.8
40	福岡県	5.8	6.9	6.4	19.0	10.4	8.8	7.9	0.7
41	佐賀県	4.4	5.5	5.5	24.3	25.0	18.2	12.1	0.4
42	長崎県	4.8	7.4	6.3	53.7	30.5	14.3	10.9	0.4
43	熊本県	4.4	5.6	5.4	27.9	22.8	15.8	11.7	0.5
44	大分県	4.4	3.3	4.0	25.7	8.5	29.5	16.2	0.4
45	宮崎県	5.0	5.9	5.7	18.0	13.3	25.9	13.7	0.2
46	鹿児島県	4.9	3.6	4.0	25.9	17.7	16.4	12.3	0.7
47	沖縄県	9.3	8.4	7.6	10.2	18.0	7.8	7.0	0.6

表 1-3 平成 14 年結果 (単位: %)

		ベンチ マーク (%)	推定値 (%)		ARE (%)		誤差率 (%)		比推定 値にかか るウエイ
		H14就調	比推定	EBLUP	比推定	EBLUP	比推定	EBLUP	
1	北海道	5.1	5.9	5.8	15.1	13.8	8.7	8.1	0.8
2	青森県	5.6	4.7	5.2	16.8	6.8	22.7	15.4	0.6
3	岩手県	5.1	3.4	4.0	33.2	21.5	22.0	16.0	0.7
4	宮城県	6.2	5.1	4.9	18.3	20.6	15.5	13.4	0.7
5	秋田県	5.3	4.5	4.7	14.5	11.0	15.6	13.0	0.7
6	山形県	4.0	4.1	4.4	2.2	9.3	21.7	16.4	0.6
7	福島県	5.5	4.1	4.3	26.2	21.4	14.6	12.3	0.8
8	茨城県	4.9	5.1	5.1	3.2	4.5	11.4	10.2	0.8
9	栃木県	5.0	3.7	3.9	26.8	22.7	20.0	16.2	0.7
10	群馬県	5.1	4.5	4.3	12.7	16.6	22.2	18.2	0.6
11	埼玉県	5.6	6.5	6.3	16.1	12.3	9.3	8.6	0.8
12	千葉県	5.0	4.6	4.8	9.0	4.8	12.8	11.0	0.8
13	東京都	5.4	6.1	5.9	13.2	9.2	7.1	6.9	0.9
14	神奈川県	5.6	5.6	5.6	0.3	0.7	9.9	9.1	0.8
15	新潟県	4.5	4.4	4.5	2.7	0.5	12.0	10.8	0.8
16	富山県	4.2	3.7	3.8	12.6	8.5	16.4	14.1	0.8
17	石川県	4.1	3.4	3.7	18.0	10.5	21.5	16.9	0.7
18	福井県	3.9	3.2	3.4	18.4	12.8	23.9	19.1	0.7
19	山梨県	4.5	4.2	3.8	5.7	15.0	22.3	20.1	0.6
20	長野県	4.1	3.0	3.2	26.3	22.1	17.8	15.4	0.8
21	岐阜県	4.2	5.7	5.0	35.2	19.5	17.5	15.4	0.6
22	静岡県	4.6	3.5	3.5	24.3	23.1	11.6	10.9	0.9
23	愛知県	4.6	4.5	4.4	2.6	3.3	8.2	8.0	0.9
24	三重県	4.2	3.7	3.9	12.5	8.2	18.0	15.1	0.8
25	滋賀県	4.5	5.3	5.1	16.7	14.1	22.4	16.4	0.5
26	京都府	6.1	7.2	6.3	18.6	2.7	16.0	13.4	0.5
27	大阪府	7.7	9.7	8.9	25.9	15.9	5.8	5.7	0.8
28	兵庫県	6.5	6.6	6.4	1.8	1.6	9.6	8.8	0.8
29	奈良県	6.2	3.0	3.8	50.8	39.5	24.4	17.0	0.7
30	和歌山県	5.4	5.3	5.5	1.5	2.3	32.5	18.1	0.3
31	鳥取県	4.7	2.2	3.2	53.5	32.1	45.6	24.1	0.6
32	島根県	3.6	7.1	5.0	98.6	38.4	42.5	22.6	0.1
33	岡山県	5.0	4.4	4.2	12.6	15.4	18.2	15.9	0.7
34	広島県	4.9	4.6	4.6	5.9	6.9	15.6	13.6	0.7
35	山口県	4.5	4.6	4.6	1.9	1.9	17.6	14.6	0.7
36	徳島県	5.1	6.5	5.5	27.9	8.7	27.3	18.0	0.3
37	香川県	4.8	3.6	3.7	24.5	23.9	20.4	17.4	0.7
38	愛媛県	5.2	5.7	5.3	10.4	1.3	17.6	14.7	0.6
39	高知県	5.6	3.9	4.5	30.2	20.3	20.8	15.2	0.7
40	福岡県	7.1	6.6	6.4	6.8	9.9	9.9	9.0	0.8
41	佐賀県	4.9	3.8	4.5	21.4	7.9	24.2	16.5	0.6
42	長崎県	5.4	4.4	4.8	19.1	11.4	18.5	14.1	0.7
43	熊本県	5.3	5.1	5.3	4.6	0.7	13.9	11.6	0.7
44	大分県	4.8	3.9	4.3	19.7	9.7	23.8	16.9	0.6
45	宮崎県	5.8	8.8	6.9	51.1	19.4	16.7	13.6	0.4
46	鹿児島県	5.2	5.2	5.5	0.8	5.0	20.6	14.9	0.5
47	沖縄県	8.1	9.4	8.7	16.2	7.6	6.6	6.4	0.8

IV 推定に用いたプログラム

EBLUP 推定量の計算に用いたプログラムを、以下に示す。

- ・プログラム①は、図 2 の対数尤度関数のグラフ描画に用いたプログラムである。定義域を細かく分割し、各分点における尤度の値を結んで、尤度関数の曲線を描いている。グラフには、最大値における接線が自動的に引かれるようになっている。曲線を描く関数 `Lik.Fnc` の引数は、以下のようにになっている。

`time`：尤度関数を求める時点定める。データ開始時点からの月数で指定する。

`grid`：定義域を分割する際の、分店の幅を定める。小さく設定すれば、より細かいグラフが描けるものの、計算にかかる時間は増える。

`range`：グラフの定義域を定める。定義域は閉区間 $[0, \text{range}]$ となる。小さく設定すれば、より広い範囲のグラフが描けるものの、計算にかかる時間は増える。

- ・プログラム②は、スコア法を用いて EBLUP 推定量を求めるプログラムである。推定値を求める関数 `Est.Score` の引数は、以下のようにになっている。

`time`：プログラム①と同じ。

`init`：数値計算の初期値を定める。

`tol`：数値計算の収束判定に用いる。1STEP 前との差が `tol` 未満になったときに、計算を打ち切る。

`maxiter`：数値計算における最大反復回数を定める。

- ・プログラム③は、BFGS 公式による準ニュートン法を用いて、EBLUP 推定量を求めるプログラムである。推定値を求める関数 `Est.BFGS` の引数は、プログラム①やプログラム②に出てきたものと同じである。

プログラム②、③を適用して計算すると、以下のような結果が出力される。

`original`：観測値。

`EBLUP`：EBLUP 推定量による推定結果。

`MSE`：EBLUP 推定量の平均二乗誤差。

`sigma2.v`：分散パラメータの推定値。

`beta`：EBLUP 推定量における回帰係数。

プログラム中のデータ `Y.dt`、`X.dt`、`E.dt` はそれぞれ、観測値、補助情報、標準誤差を表している。これらは全て、`T` 行 $\times$ 47 列の行列である。ここで `T` は、時系列の長さを表している。

```
#####
# Program ①
# Software : 「R」
# Title : Log-Likelihood Function
#####

Y.dt <- matrix(scan("D:/uemp.r.txt"), ncol = 47, byrow = T)
X.dt <- matrix(scan("D:/job.r.txt"), ncol = 47, byrow = T)
E.dt <- matrix(scan("D:/uemp.r.err.txt"), ncol = 47, byrow = T)
I <- rep(1, 47)

Dt.Cut <- function (time)
{
  Y <- Y.dt[time,]
  X <- X.dt[time,]
  Z <- cbind(I, X)
  E <- E.dt[time,]
  phi <- E ^ 2
}

Lik <- function(sig2.v)
{
  work <- sig2.v + phi
  Z.w <- Z / work
  beta <- solve(crossprod(Z.w, Z)) %*% crossprod(Z.w, Y)
  term1 <- crossprod(I, log(work))
  term2 <- m * log(2 * pi)
  for (i in 1:m) {
    temp <- (Y[i] - t(Z[i,]) %*% beta) ^ 2 / work[i]
    term2 <- term2 + temp
  }
  res <- - 0.5 * (term1 + term2)
  as.vector(res)
}

Lik.Fnc <- function (time, grid, range)
{
  Dt.Cut(time)
  range <- floor(range / grid)
  wrk1 <- cumsum(rep(grid, range + 1))
  wrk2 <- rep(NA, range + 1)
  for (j in 1:(range + 1)) {
    wrk2[j] <- Lik(grid * (j - 1))
  }
}
```

```
max.val <- max(wrk2)
res <- cbind(wrk1, wrk2)
colnames(res) <- c("scale", "log-likelihood")
plot(res, type = "l", lwd = 2, ylab = "log-likelihood",
      xlab = "sigma2.v")
abline(h = max.val, col = "green", lty = 2, lwd = 2)
res
}

result <- Lik.Fnc(141, 0.001, 3)

#####
# Program ②
# Software:「R」
# Title:EBLUP Estimsation By The Scoring Method
#####

Score <- function(init, tol, maxiter)
{
  work1 <- init
  iter <- 0
  repeat {
    s2 <- format(round(work1, 6), nsmall = 6)
    L <- format(round(Lik(work1), 6), nsmall = 6)
    cat(iter, ":", s2, L, "¥n")
    work2 <- 1 / (work1 + phi)
    Z.w <- Z * work2
    beta <- solve(crossprod(Z.w, Z)) %*% crossprod(Z.w, Y)
    work3 <- crossprod(work2, I)
    work4 <- crossprod(((Y - Z %*% beta) * work2) ^ 2, I)
    work5 <- work1 + as.vector((work4 - work3) /
      crossprod(work2 ^ 2, I))
    if (abs(work1 - work5) < tol) {
      s2 <- format(round(work5, 6), nsmall = 6)
      cat(iter + 1, ":", s2, L, "¥n")
      break
    } else if (iter > maxiter) {
      cat("Does not converges.", "¥n")
      break
    }
  }
  work1 <- work5
  iter <- iter + 1
}
```

```

    }
    work5
}

EBLUP <- function(sig2.v)
{
  work <- phi + sig2.v
  Z.w <- Z / work
  beta <- solve(crossprod(Z.w, Z)) %*% crossprod(Z.w, Y)
  gamma <- sig2.v / work
  EBLUP <- as.vector(gamma * Y + (I - gamma) * Z %*% beta)
  g1 <- gamma * phi
  g2 <- (I - gamma) ^ 2 *
    diag(Z %*% solve(crossprod(Z, Z / work)) %*% t(Z))
  MSE <- g1 + g2
  list(original = Y, EBLUP = EBLUP, MSE = MSE, gamma = gamma,
    sigma2.v = sig2.v, beta = beta)
}

Est.Score <- function (time, init, tol, maxiter)
{
  Dt.Cut(time)
  sig2.v <- Score(init, tol, maxiter)
  EBLUP(sig2.v)
}

result <- Est.Score(141, 0, 1e-6, 100)

#####
# Program ③
# Software:「R」
# Title:EBLUP Estimation By The Quasi-Newton Method
#       Using Broyden-Fletcher-Goldfarb-Shanno Formula
#####

BFGS <- function (init)
{
  fnc <- function(tau) -1 * Lik(exp(tau))
  opt <- optim(log(init), fnc, method = "BFGS",
    control = list(trace = 10))
  exp(opt$par)
}

```

高部 勲：EBLUP 推定量を用いた小地域推定

```
Est.BFGS <- function (time, init)

{
  Dt.Cut(time)
  sig2.v <- BFGS(init)
  EBLUP(sig2.v)
}

result <- Est.BFGS(141, 0.5)
```

第 3 章 Time Series & Cross Sectional Model を用いた小地域推定 ～階層的ベイズモデル及びマルコフ連鎖モンテカルロ法の利用～

本章では、Rao-Yu モデルを用いた都道府県別完全失業率の推定方法について考察する。Rao-Yu モデルは、Time Series and Cross Sectional モデルの一種である。Time Series and Cross Sectional モデルとは、第 2 章で解説した Fay-Herriot モデルを時系列方向への多変量モデルに拡張したものであり、空間的な情報と時系列の情報の両方を推定に利用することができる。

推定における計算の効率を高めるために、モデルを階層的ベイズモデルの形に表現して、ギブスサンプラーによるマルコフ連鎖モンテカルロ法を適用している。(Rao(2003))。

誤差率やベンチマークとの比較では、Rao-Yu モデルによる推定の方が良いという結果が得られた。しかし事後予測分布を用いてモデルの当てはまりを判定したところ、当てはまりはあまり良くなかった。

以下 I 節と II 節ではそれぞれ、今回の分析に必要なマルコフ連鎖モンテカルロ法および階層的ベイズモデルに関する理論を簡単に述べる。III 節では Rao-Yu モデルに関して述べる。IV 節では、III 節で述べたモデルを労働力調査へ適用した結果についてまとめている。V 節では今後の課題について述べている。

I ベイズ統計学および階層的ベイズモデル

以下では主に、中妻 (2004) を引用しつつ、本稿で用いるベイズ統計の手法及び階層的ベイズモデルについて説明する。古典的な統計分析と異なり、ベイズ統計学ではパラメータの真の値が、確率変数であると仮定する。そして観測値のデータが与えられたときに、各種パラメータの条件付確率分布を用いて推定を行う。事後分布は次の「ベイズの定理」を用いて求める。

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \quad (p(\theta|y) \propto p(y|\theta)p(\theta)) \quad (1)$$

$p(y|\theta)$ を尤度 (likelihood) と呼ぶ。尤度はパラメータ θ の関数であり、 θ のもつもらしさを表していると解釈できる。

$p(\theta)$ を事前分布と呼ぶ。観測値 y が観測される前に分析者が持っているパラメータに関する情報を表していると考えられる。事前分布のパラメータを、ハイパーパラメータと呼ぶ。

$p(\theta|y)$ を事後分布と呼ぶ。観測値のデータが与えられた場合の、パラメータ θ に関する確率分布である。ベイズ統計分析ではこの事後分布を基に推定を行う。

ベイズ統計学では、適当に定めた損失関数 $L(\theta, \bar{\theta})$ を用いて、パラメータの推定値と真の値との距離を測る。損失関数としては、二乗誤差損失 $L(\theta, \bar{\theta}) = (\theta - \bar{\theta})^2$ が多く用いられる。以下のような、 θ の事後分布による損失関数 $L(\theta, \bar{\theta})$ の平均を、リスク関数と呼ぶ。

$$R(\theta|y) = \int L(\theta, \bar{\theta})P(\theta|y)d\theta \quad (2)$$

ベイズ統計分析ではリスク関数を最小化する $\hat{\theta}$ を点推定に用いる。損失関数が二乗誤差損失の場合には、事後分布の平均 (事後平均) が、リスク関数を最小にする推定値になる (中妻 (2004))。

ベイズ統計分析を行うためには、事前分布を設定する必要がある。ベイズ統計分析では共役事前分布が用いられることが多い。共役事前分布とは、事前分布と事後分布がともに同じ分布に属するような分布の属である。共役事前分布を用いることで解析的に事後分布を導出するのが容易になり、計算の速度及び精度が向上する。事前の情報が少ない状態を表すために、無情報事前分布が用いられることも多い。無情報事前分布としては、Jeffreys の提唱したものがよく用いられる。

パラメータに対して階層的な構造を仮定したモデルを、階層的ベイズモデルと呼ぶ。階層的な構造を仮定した方が、モデルを簡潔に記述できる場合が多い。階層的ベイズモデルでは、事前分布が以下のように分解されると仮定する。

$$P(\theta) = \int \cdots \int P(\theta | \theta_1) P(\theta_1 | \theta_2) \cdots P(\theta_{k-1} | \theta_k) P(\theta_k) d\theta_1 \cdots d\theta_k \quad (3)$$

$P(\theta_{k-1} | \theta_k)$ は、第 k 段階の事前分布である。階層的ベイズモデルにより、大量のパラメータを持つようなモデルにもある程度対応できる。しかし後に述べる Rao-Yu モデルでは事後分布の形状が複雑になり、平均値を求める際の積分の次元が高くなる。そのような複雑な多重積分の評価を解析的に行うことはほぼ不可能であり、数値計算による近似を行う場合でも計算機にかかる負荷が大きい。このような場合に事後分布から互いに独立な大量の乱数 $\theta = (\theta_1, \theta_2, \dots, \theta_b)$ を発生させることができれば、その標本平均や標本分散を計算することで、大数の法則から事後分布の平均及び分散を推定することができる。次節ではそのような乱数発生の方法を紹介する。

II マルコフ連鎖モンテカルロ法

確率変数の時系列 $(x_1, x_2, \dots) = \{x_t\}_{t=1}^{\infty}$ を考える。次の式(6)のように、一期前の確率変数の実現値のみが $\{x_t\}_{t=1}^{\infty}$ の各時点の条件付分布に影響しているとき、 $\{x_t\}_{t=1}^{\infty}$ をマルコフ連鎖と呼ぶ。

$$P(x_t | x_{t-1}, \dots, x_1) = P(x_t | x_{t-1}) \quad (4)$$

各時点における条件付分布及び初期分布が分かっている場合には、マルコフ連鎖からの乱数発生を、以下のようにして容易に行うことができる。

$$\text{STEP. 1 : } P(x_1) \rightarrow \tilde{x}_1 \quad (t = 0)$$

$$\text{STEP. 2 : } P(x_t | \tilde{x}_{t-1}) \rightarrow \tilde{x}_t$$

t を $t+1$ として、STEP. 2 にもどる。

$t \rightarrow \infty$ のときに x_t の分布が特定の確率密度に収束する場合、マルコフ連鎖は不変密度を持つ

と表現し、不変密度を確率密度関数に持つような分布を不変分布と呼ぶ。 $\{x_t\}_{t=1}^{\infty}$ に不変密度が存在する場合には、上記のアルゴリズムを繰り返すことにより、不変密度に近い分布からの乱数を発生させることができる。実際はエルゴード性に関する条件も必要であるものの、ここでは省略する。

もし確率変数間にマルコフ連鎖の関係を作ることができて、かつ事後分布を不変密度として持つようにできれば、上記のアルゴリズムを十分多く繰り返すことで、事後分布に近い分布からの乱数を発生させることができる。そのような乱数発生の方法をマルコフ連鎖モンテカルロ法と呼ぶ。発生させた乱数は、モンテカルロ標本とも呼ばれる。マルコフ連鎖モンテカルロ法には様々な種類がある。その中でも比較的簡単にプログラムを組むことができるギブスサンプラーについて、以下に述べる。

k 変数の確率変数に対する同時確率密度関数 $P(\boldsymbol{\theta}) = P(\theta_1, \theta_2, \dots, \theta_k)$ を考える。特定の確率変数以外の残りすべての確率変数の値が与えられた下での条件付確率密度のことを、その確率変数に対する全条件付確率分布と呼ぶ。ギブスサンプラーは、全条件付の事後分布によって以下のよう構成されるアルゴリズムである。 $\theta_n^{(t)}$ は、 t 回目に発生させた θ_n の乱数を表す。

適当な初期値 $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_k^{(0)})$ を与える。

STEP. 1 : $P(\theta_1^{(t)} | \theta_2^{(t)}, \dots, \theta_k^{(t)}) \rightarrow \theta_1^{(t+1)}$

STEP. 2 : $P(\theta_2^{(t)} | \theta_1^{(t+1)}, \dots, \theta_k^{(t)}) \rightarrow \theta_2^{(t+1)}$

$\vdots$

STEP. $k-1$: $P(\theta_{k-1}^{(t)} | \theta_1^{(t+1)}, \dots, \theta_{k-2}^{(t+1)}, \theta_k^{(t)}) \rightarrow \theta_{k-1}^{(t+1)}$

STEP. k : $P(\theta_k^{(t)} | \theta_1^{(t+1)}, \dots, \theta_{k-1}^{(t+1)}) \rightarrow \theta_k^{(t+1)}$

t を $t+1$ として、STEP. 1 にもどる。

ギブスサンプラーのアルゴリズムで発生させた乱数は事後分布を不変密度として持つマルコフ連鎖になることが知られている（中妻（2004））。同時確率密度 $P(\boldsymbol{\theta}) = P(\theta_1, \theta_2, \dots, \theta_k)$ の形状が複雑な場合でも、その全条件付確率分布は低次元の単純な分布（正規分布、ガンマ分布など）になる場合が多く、そのような分布からの乱数は、通常の統計解析パッケージ等で簡単に発生させることができる。

マルコフ連鎖が定常分布に収束したと判断されるまで乱数を発生し続けることを、“burn in” と呼ぶ。burn in までに発生させた乱数は初期値の影響を強く受けていると考えられるので、計算には用いない。burn in 以降の乱数は、“burn out” と呼ばれる。単一のマルコフ連鎖から乱数を発生させることを、“single run” と呼ぶ。これに対して、複数のマルコフ連鎖から乱数を発生させ、その平均により推定を行う方法を “multi run” と呼ぶ。multi run を適用した方がパラメータの初期

状態の影響を少なくできるので、不変分布への収束が **single run** の場合に比べて速くなる（中妻（2004））。また **multi run** を適用した場合には、マルコフ連鎖が定常分布に収束しているかどうかの診断をすることができる。これについてはIV節で述べる。

III Time Series and Cross Sectional モデル

1 Rao-Yu モデル

第2章で紹介した Fay-Herriot モデルは、多変量の場合に拡張することができる。これにより、複数の属性の変数を同時に扱うことができる。また変数間の相関を推定に利用することもできる。Fay-Herriot モデルを時系列方向への多変量モデルへと拡張したものが、Time Series and Cross Sectional モデルであり、時間領域と空間領域の両方の情報を推定に用いることができる。Time Series and Cross Sectional モデルの一種である Rao-Yu モデルは、地域の効果 v_i の他に、ランダムな時系列の効果を表す確率変数 u_{it} を含んでおり、次のような形で表される。

$$y_{it} = \theta_{it} + e_{it} \quad (5)$$

$$\theta_{it} = \mathbf{x}_{it} \boldsymbol{\beta} + v_i + u_{it}, \quad i=1, \dots, 47, \quad t=1, \dots, 14$$

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$$

$$\left[\begin{array}{l} y_{it} : \text{第 } i \text{ 小地域、時点 } t \text{ における完全失業率の観測値データ} \\ \mathbf{x}_{it} : \text{第 } i \text{ 小地域、時点 } t \text{ における補助情報（定数項を含む）} \\ \theta_{it} : \text{第 } i \text{ 小地域、時点 } t \text{ における真の完全失業率} \\ e_{it} : \text{第 } i \text{ 小地域、時点 } t \text{ における標本誤差} \\ v_i \sim N(0, \sigma_v^2) : \text{第 } i \text{ 小地域における地域の効果} \\ u_{it} : \text{第 } i \text{ 小地域、時点 } t \text{ における時系列の効果} \\ \varepsilon_{it} \sim N(0, \sigma_e^2) \end{array} \right]$$

y_{it} は、各 i に関しては独立であると仮定される。しかし標本誤差に系列相関がある（第4章参照）と考えられることから、 t に関しては独立ではない。第4章によると、標本誤差の系列相関は、理論上13期分ある。このことを考慮して、14期分の変数をまとめて扱うことにする。よって変数ベクトルの次元は14になる。

2 階層的ベイズモデルを用いた Rao-Yu モデルの表現

式 (5) の Rao-Yu モデルで、 y_{it} に関する式を以下のような分布の形に書き換えておく。

$$\mathbf{y}_i \sim N_{14}(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i) \quad (6)$$

$$\left[\begin{array}{l} \boldsymbol{\Sigma}_i : \text{第 } i \text{ 県における } \mathbf{e}_i' \text{ の分散共分散行列} \\ \mathbf{y}_i = (y_{i1}, \dots, y_{i,14}) \text{ , } \boldsymbol{\theta}_i = (\theta_{i1}, \dots, \theta_{i,14}) \text{ , } \mathbf{e}_i = (e_{i1}, \dots, e_{i,14}) \end{array} \right]$$

標本誤差の分散共分散行列 $\boldsymbol{\Sigma}_i$ は、推定値 $\hat{\boldsymbol{\Sigma}}_i$ で置き換える。 $\hat{\boldsymbol{\Sigma}}_i$ は以下のようにして推定する。

なお標本誤差の自己相関については、第 4 章の方法で推定したものを用いている。

$$\hat{\boldsymbol{\Sigma}}_i = (\hat{\sigma}_{its}^2)_{14 \times 14} \text{ , } i=1, \dots, 47 \text{ , } t, s=1, \dots, 14 \quad (7)$$

$$\left[\begin{array}{l} \hat{\sigma}_{itt}^2 = \overline{C}_i y_{it}^2 \text{ , } \hat{\sigma}_{its} = \bar{\rho}_i \hat{\sigma}_{itt} \hat{\sigma}_{iss} \\ \overline{C}_i = \frac{1}{14} \sum_{t=1}^{14} C_{it} \text{ , } \bar{\rho}_i = \frac{1}{47} \sum_{k=1}^{47} \rho_{ik} \\ C_{it} : \text{第 } i \text{ 小地域の時点 } t \text{ における完全失業率の誤差率} \\ \rho_{ik} : \text{第 } i \text{ 小地域における標本誤差の } k \text{ 次の自己相関} \end{array} \right]$$

誤差率及び誤差の自己相関をそのまま用いるよりも、上記のように時間又は空間に関して平均してから用いる方が、誤差率の改善の点から見て良い結果になるという報告がある (Rao & Gambino (2003))。

式 (5) ~ (7) により、Rao-Yu モデルを以下のような階層的ベイズモデルの形に表現できる。

$$\mathbf{y}_i | \boldsymbol{\theta}_i \sim N_{14}(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i) \quad (8)$$

$$\theta_{it} | \boldsymbol{\beta}, u_{it}, \sigma_v^2 \sim N(\mathbf{x}_{it} \boldsymbol{\beta} + \rho u_{it}, \sigma_v^2)$$

$$u_{it} | u_{i,t-1}, \sigma_\varepsilon^2 \sim N(\rho u_{i,t-1}, \sigma_\varepsilon^2)$$

$$\sigma_v^2 \sim IG(a_1, b_1) \text{ , } \sigma_\varepsilon^2 \sim IG(a_2, b_2)$$

$$f(\boldsymbol{\beta}) \propto 1 \text{ (無情報事前分布)}$$

$$[IG : \text{逆ガンマ分布}]$$

回帰係数 β に関する情報がほとんどないという状態を推定に反映させるために、 β の事前分布として無情報事前分布 (flat prior) を採用している。また逆ガンマ分布は σ_v^2 の共約事前分布である (Rao(2003))、中妻 (2004))。 σ_v^2 についても事前の情報がないことを反映させるために、事前分布のハイパーパラメータを十分小さい正の値にとる。 Rao & Gambino (2003) では $a_1 = a_2 = b_1 = b_2 = 0.001$ に設定しており、本稿の分析でもこの値に設定することとする。0.001 という数値はかなり小さく、他の数値にはほとんど影響を与えない。

未知のパラメータは $(\beta, \sigma_v^2, \sigma_e^2, \mathbf{u}, \theta, \theta_1, \dots, \theta_{47})$ であり、特に $\theta = (\theta_1, \dots, \theta_{47})$ が最も知りたいパラメータである。各パラメータに対する全条件付事後分布は、それぞれ以下のように示される (Rao & Gambino (2003))。

$$\beta | \mathbf{y}, \sigma_v^2, \sigma_e^2, \mathbf{u}, \theta \sim N((X'X)^{-1}X'(\theta - \mathbf{u}), \sigma_v^2(X'X)^{-1}) \quad (9)$$

$$\sigma_v^2 | \mathbf{y}, \sigma_e^2, \mathbf{u}, \theta \sim IG\left(a_1 + \frac{47 \times 14}{2}, b_1 + \sum_{i=1}^m \sum_{t=1}^T \frac{(\theta_{it} - \mathbf{x}_{it}\beta - u_{it})^2}{2}\right)$$

$$\sigma_e^2 | \mathbf{y}, \sigma_v^2, \mathbf{u} \sim IG\left(a_2 + \frac{47 \times (14-1)}{2}, b_2 + \sum_{i=1}^m \sum_{t=1}^T \frac{(u_{it} - \rho u_{i,t-1})^2}{2}\right)$$

$$u_{i1} | \mathbf{y}, \beta, \sigma_v^2, \sigma_e^2, u_{i2}, \theta \quad [t=1]$$

$$\sim N\left(\left(\frac{1}{\sigma_v^2} + \frac{\rho^2}{\sigma_e^2}\right)^{-1} \left(\frac{\theta_{i1} - \mathbf{x}_{i1}\beta}{\sigma_v^2} + \frac{\rho u_{i2}}{\sigma_e^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{\rho^2}{\sigma_e^2}\right)^{-1}\right)$$

$$u_{it} | \mathbf{y}, \beta, \sigma_v^2, \sigma_e^2, u_{i,t-1}, u_{i,t+1}, \theta \quad [2 \leq t \leq 13]$$

$$\sim N\left(\left(\frac{1}{\sigma_v^2} + \frac{1+\rho}{\sigma_e^2}\right)^{-1} \left(\frac{\theta_{it} - \mathbf{x}_{it}\beta}{\sigma_v^2} + \frac{\rho u_{i,t-1} + \rho u_{i,t+1}}{\sigma_e^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{1+\rho}{\sigma_e^2}\right)^{-1}\right)$$

$$u_{i,14} | \mathbf{y}, \beta, \sigma_v^2, \sigma_e^2, u_{i,13}, \theta \quad [t=T]$$

$$\sim N\left(\left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_e^2}\right)^{-1} \left(\frac{\theta_{i,13} - \mathbf{x}_{i,13}\beta}{\sigma_v^2} + \frac{\rho u_{i,13}}{\sigma_e^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_e^2}\right)^{-1}\right)$$

$$\theta_i | \mathbf{y}, \beta, \sigma_v^2, \sigma_e^2, \mathbf{u} \sim N\left((\sigma_v^{-2}\mathbf{I}_{14} + \Sigma_i^{-1})^{-1}(\Sigma_i^{-1}\mathbf{y}_i + \sigma_v^{-2}(X_i\beta + \mathbf{u}_i)), (\sigma_v^{-2}\mathbf{I}_{14} + \Sigma_i^{-1})^{-1}\right)$$

$$\left[\begin{array}{l} \mathbf{y} = (\mathbf{y}_1^t, \dots, \mathbf{y}_{47}^t) : Tm \times 1 \text{ ベクトル} \quad , \quad \mathbf{X} = (X_1^t, \dots, X_{47}^t) : Tm \times 2 \text{ ベクトル} \\ \boldsymbol{\theta} = (\boldsymbol{\theta}_1^t, \dots, \boldsymbol{\theta}_{47}^t) : Tm \times 1 \text{ ベクトル} \quad , \quad \mathbf{u} = (\mathbf{u}_1^t, \dots, \mathbf{u}_{47}^t) : Tm \times 1 \text{ ベクトル} \\ \mathbf{I}_{14} : 14 \times 14 \text{ 単位行列} \end{array} \right]$$

上記の全条件付事後分布を基にして、マルコフ連鎖モンテカルロ法を適用し、都道府県別の完全失業率の推定を行う。Rao-Yu モデルによる推定を行うと、14 か月分の推定値 $\boldsymbol{\theta}$ が算出される。このうち最新月 ($t=14$) の数値だけを、推定値として用いる。

各確率変数の全条件付事後分布の平均が解析的に計算できる場合には、それらの平均や分散を計算することで、より効率的な推定を行うことができ、モンテカルロ法によるシミュレーションの誤差を抑えることができる。このような推定量を、ラオ・ブラックウェル推定量と呼ぶ。ラオ・ブラックウェル推定量の計算に必要な全条件付事後分布の平均は、ギブスサンプラーを計算している過程で、副産物として得られる場合が多い。よって計算に伴う新たな付加は、それ程多くない。multi run による推定を行った場合には、Rao-Yu モデルにおけるラオ・ブラックウェル推定量は以下のようなになる Rao & Gambino (2003)。

$$E[\theta_i | y_i] = \frac{1}{LD} \sum_{l=1}^L \sum_{k=d+1}^{d+D} [(\sigma_v^{-2[lk]} \mathbf{I}_{14} + \boldsymbol{\Sigma}_i^{-1}) \times (\boldsymbol{\Sigma}_i^{-1} \mathbf{y}_i + \sigma_v^{-2[lk]} (X_i \boldsymbol{\beta}^{[lk]} + \mathbf{u}_i^{[lk]}))] \quad (10)$$

$$V[\theta_i | y_i] = \frac{1}{LD} \sum_{l=1}^L \sum_{k=d+1}^{d+D} (\sigma_v^{-2[lk]} \mathbf{I}_{14} + \boldsymbol{\Sigma}_i^{-1}) \quad (11)$$

$$+ \frac{1}{(LD)^2} \sum_{l=1}^L \sum_{k=d+1}^{d+D} [(\sigma_v^{-2[lk]} \mathbf{I}_{14} + \boldsymbol{\Sigma}_i^{-1}) \times (\boldsymbol{\Sigma}_i^{-1} \mathbf{y}_i + \sigma_v^{-2[lk]} (X_i \boldsymbol{\beta}^{[lk]} + \mathbf{u}_i^{[lk]}))] \\ \times [(\sigma_v^{-2[lk]} \mathbf{I}_{14} + \boldsymbol{\Sigma}_i^{-1}) \times (\boldsymbol{\Sigma}_i^{-1} \mathbf{y}_i + \sigma_v^{-2[lk]} (X_i \boldsymbol{\beta}^{[lk]} + \mathbf{u}_i^{[lk]}))]^t \\ - \frac{1}{(LD)^2} [\sum_{l=1}^L \sum_{k=d+1}^{d+D} [(\sigma_v^{-2[lk]} \mathbf{I}_{14} + \boldsymbol{\Sigma}_i^{-1}) \times (\boldsymbol{\Sigma}_i^{-1} \mathbf{y}_i + \sigma_v^{-2[lk]} (X_i \boldsymbol{\beta}^{[lk]} + \mathbf{u}_i^{[lk]}))]]$$

$$\times [\sum_{l=1}^L \sum_{k=d+1}^{d+D} [(\sigma_v^{-2[lk]} \mathbf{I}_{14} + \boldsymbol{\Sigma}_i^{-1}) \times (\boldsymbol{\Sigma}_i^{-1} \mathbf{y}_i + \sigma_v^{-2[lk]} (X_i \boldsymbol{\beta}^{[lk]} + \mathbf{u}_i^{[lk]}))]]^t$$

$$\left[\begin{array}{l} L : \text{同時に発生させるマルコフ連鎖の数} \\ d : \text{burn in までに各マルコフ連鎖内で発生させた乱数の総数} \\ D : \text{burn out 後に発生させた乱数の総数} \end{array} \right]$$

今回は 14 か月分の推定値 $\boldsymbol{\theta}$ のうち、 $t=14$ の数値だけを、推定値として用いるので、実際は式(10)では 14 行目の成分のみに、式(11)では 14 行 14 列目の成分のみに関して計算を行えばよい。

ここで、上記の全条件付事後分布にマルコフ連鎖モンテカルロ法を適用した場合のモンテカルロ標本のサンプルパスを観察してみる。観測値のデータには、平成 12 年 1 月から平成 13 年 2 月までの都道府県別月次の完全失業率を用いた。補助情報としては、同じ時期の有効求人倍率を用

いた。また時系列の効果 u_{it} の係数を 0.5 とした。multi run によりモンテカルロ標本を発生させることとし、 $L=3$ 、 $D=2000$ 、 $d=1000$ と設定した。サンプルパスの生成に用いたプログラムについては、この章の最後に示してある。次の各パラメータについてサンプルパスを描き、その様子を調べることにする。マルコフ連鎖モンテカルロ法における未知パラメータの初期値は、V 節の方法で設定した。

- ・ 回帰係数 β
- ・ 分散パラメータ σ_v^2 及び σ_e^2
- ・ 東京都の $t=14$ における u_{it} 及び θ_{it} ($u_{13,14}$ 、 $\theta_{13,14}$)

これらのパラメータのサンプルパスを示したものが、図 1 である。またそれらをヒストグラムに表したものが図 2 である。なお各グラフの表題とパラメータの関係は、以下のとおりである。

Beta Intercept：回帰モデルの定数、Beta Coefficient：回帰モデルの係数

Sigma2.v： σ_v^2 、Sigma2.e： σ_e^2 、U.it： $u_{13,14}$ 、Theta.it： $\theta_{13,14}$

以下にそれらの特徴を示す。

- ・ 回帰係数 β のサンプルパスは、係数及び定数ともに一定の平均の周りに散らばっているように見える。三つの連鎖は適度に混じり合っており、それぞれのマルコフ連鎖が共通の平均に近づいていることが分かる。各々のサンプルパスの自己相関は高いように感じられる。ヒストグラムで見ると、定数の分布は若干右にゆがんでいるように見える。また係数の分布はやや丸みを帯びているように見える。しかしほぼ正規分布のような形をしている。
- ・ 分散パラメータ σ_v^2 及び σ_e^2 については、最初にかなり大きい値を取っているものの、その後急速に安定した状態へ推移している。ヒストグラムには逆ガンマ分布の形が比較的良く出ている。各々のサンプルパスには一部周期的と思われる動きが見られ、自己相関が若干高いように思われる。
- ・ u_{it} 及び θ_{it} については、三つの連鎖がかなりよく混ざっており、初期値の影響はほとんど受けていない。自己相関もかなり少ないように見える。安定した定常分布に、ほぼ収束していると考えられる。

IV 不変分布への収束判断及びモデル選択

1 不変分布への収束判断

II 章でも触れたように、マルコフ連鎖モンテカルロ法で **multi run** を適用した場合には、マルコフ連鎖が定常分布に収束しているかどうかを診断することができる。診断には次の統計量を用いる (Rao(2003))。

$$\hat{R}_{i,14} = \hat{V}_{i,14} / \hat{W}_{i,14} \quad (12)$$

$$\left[\begin{array}{l} \hat{V}_{i,14} = \frac{D-1}{D} \hat{W}_{i,14} + \frac{1}{D} \hat{B}_{i,14} \\ \frac{1}{D} \hat{B}_{i,14} = \frac{1}{L-1} \sum_{l=1}^L \left(\bar{\theta}_{i,14}^{[l]} - \bar{\bar{\theta}}_{i,14} \right)^2 : \text{between run variance} \\ \hat{W}_{i,14} = \frac{1}{L(D-1)} \sum_{l=1}^L \sum_{k=d+1}^{d+D} \left(\theta_{i,14}^{[lk]} - \bar{\theta}_{i,14}^{[l]} \right)^2 : \text{within run variance} \\ \bar{\theta}_{i,14}^{[l]} = \frac{1}{D} \sum_{k=d+1}^{d+D} \theta_{i,14}^{[lk]}, \quad \bar{\bar{\theta}}_{i,14} = \frac{1}{L} \sum_{l=1}^L \bar{\theta}_{i,14}^{[l]} \\ \theta_{i,14}^{[lk]} : \text{第 } i \text{ 県の時点 } t=14 \text{ における、} l \text{ 番目の連鎖内の } k \text{ 番目の乱数} \end{array} \right]$$

上記の統計量では、最新月 ($t=14$) の結果に注目して、統計量を計算している。各小地域について $\hat{R}_{i,14}$ が 1 に近ければ、マルコフ連鎖が定常分布へ収束していると判断される。どの程度 1 に近ければ良いのかという判断の問題がある。今回の分析では、1.5 以上の値を取るような小地域が無ければ、マルコフ連鎖が定常分布へほぼ収束していると判断することにした。

2 モデルの選択

階層的ベイズモデルによる分析を行う場合、回帰の説明変数やモデル内の定数の選択によって、複数のモデルが考えられることがある。このような場合、分析に当たってはモデル間の優劣を比較し、最適なモデルを選ぶ必要がある。最適なモデルの選択は、予測分布からサンプリングしたモンテカルロ標本を用いて行うことが出来る。予測分布とは、事後平均により推定したパラメータ $\boldsymbol{\theta}$ を尤度 $f(\mathbf{y}|\boldsymbol{\theta})$ に代入したものである。予測分布からのモンテカルロ標本 $\mathbf{y}^{*[lk]}$ のサンプリングは、尤度 $f(\mathbf{y}|\boldsymbol{\theta})$ に $\boldsymbol{\theta}^{[lk]}$ を代入した分布 $f(\mathbf{y}|\boldsymbol{\theta}^{[lk]})$ ($=N(\mathbf{y}|\boldsymbol{\theta}^{[lk]})$) から行う。ギブスサンプラーでは式 (6) を利用して、比較的簡単に $\mathbf{y}^{*[lk]}$ を発生させることが可能なので、計算の際に新たにかかる負荷は少ない。

$\mathbf{y}_{obs}$ を観測値のデータとした場合、設定したモデルが適切に選択されていて、かつパラメータ

$\theta^{[lk]}$ が精度良く推定されていれば、 $\mathbf{y}^{*[lk]}$ は事後予測分布 $f(\mathbf{y}^* | \mathbf{y}_{obs})$ に近い分布からのサンプルとみなすことができる。この場合には $\mathbf{y}^{*[lk]}$ と $\mathbf{y}_{obs}$ は比較的近い値を取ることが予想される。もしモデルが適切でなければ、これらは離れた値を取る。そこで何らかの方法で観測値と $\mathbf{y}^{*[lk]}$ との距離を測ることができれば、それをモデル選択の際の基準として用いることができる。そのような距離を測る基準として、以下の二つの統計量が考えられる。

(1) d 値：

モデル同士の優劣の比較には、事後予測分布を用いた次の統計量を用いる。

$$\begin{aligned} d(\mathbf{y}^*, \mathbf{y}_{obs}) &= E(\|\mathbf{y}^{*[ld]} - \mathbf{y}_{obs}\| / (47 \times 14) | \mathbf{y}_{obs}) \\ &= \sum_{l=1}^L \sum_{k=d+1}^{d+D} \|\mathbf{y}^{*[lk]} - \mathbf{y}_{obs}\| / (47 \times 14) \end{aligned} \quad (13)$$

$\|\cdot\|$ は通常のユークリッドノルム（二乗和の平方根）を表す。期待値は LD 個のサンプルについて取っている。統計量 d の値が小さいモデルの方が、当てはまりが良いと言える。モデル間の相対的な比較を行う統計量である。

(2) 事後予測 p 値：

モデルの当てはまりについての診断は、事後予測 p 値と呼ばれる以下の推定量を用いて行う。このとき、次の統計量を用いて診断を行う。

$$\hat{p} = \frac{1}{L} \sum_{l=1}^L \hat{p}_i^{[l]} \quad (14)$$

$$\left[\begin{array}{l} \hat{p}^{[l]} = \frac{1}{D} \sum_{k=d+1}^{d+D} I\{T(\mathbf{y}_i^{*[lk]}, \boldsymbol{\theta}_i^{[lk]}) \geq T(\mathbf{y}_i^{obs}, \boldsymbol{\theta}_i^{[lk]})\} \\ T(\mathbf{y}^*, \boldsymbol{\theta}) = \sum_{i=1}^{47} (\mathbf{y}_i^* - \boldsymbol{\theta}_i) \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i^* - \boldsymbol{\theta}_i)^T \\ \mathbf{y}^* = (\mathbf{y}_1^*, \dots, \mathbf{y}_{47}^*) \quad , \quad \mathbf{y}_i^* = (\mathbf{y}_{i,1}^*, \dots, \mathbf{y}_{i,14}^*) \end{array} \right]$$

式 (9) の中の $I\{\bullet\}$ は定義関数であり、 $\{\bullet\}$ の中の命題が真のときに 1 をとり、偽のときに 0 をとる。またモデルがデータに良く当てはまっている場合には、 $T(\mathbf{y}_i^{*[lk]}, \boldsymbol{\theta}_i^{[lk]})$ と $T(\mathbf{y}_i^{obs}, \boldsymbol{\theta}_i^{[lk]})$ が近い値をとることが予想される。そのときには統計量 $\hat{p}$ が 0.5 付近の値をとると考えられる。

$\hat{p}$ が極端に 0 や 1 に近い値をとる場合には、モデルの当てはまりが悪いと考えられる。モデル間の相対的な比較とともに、 $\hat{p}$ が 0.5 付近にあるかどうかという絶対的な当てはまりの良さを確認できる統計量である。

V 労働力調査への適用

1 推定方法及び結果の比較方法

これまでに紹介した Rao-Yu モデル及びマルコフ連鎖モンテカルロ法を適用して、都道府県別月次の完全失業率を推定する。推定の枠組みは以下のとおりである。

- ・式 (14) の全条件付き事後分布からギブスサンプラーによりモンテカルロ標本を発生させる。その際に $L = 10$ 、 $D = 2000$ と設定し、各マルコフ連鎖内で発生させた 2000 個の乱数のうち最初の 1000 個は Burn in の状態であるとして、推定値の計算には用いないことにする。
- ・都道府県別の標本誤差 ψ_i 及び標本誤差の自己相関 ρ_{ik} には、第 4 章の方法で個票データから推定したものをを用いる。
- ・回帰の部分の補助情報については、職業安定業務統計における都道府県別月次の有効求人倍率を用いる。
- ・マルコフ連鎖モンテカルロ法における未知パラメータの初期値は、次のようにして定めた。なお推定値が初期値に依存しないように、初期値をかなり広範囲にばらまいている。

β 、 σ_v^2 、 σ_e^2 、 $\dots$ $U(0,5)$ (一様分布) からそれぞれの成分の初期値を発生させる

$\theta = (\theta_1, \dots, \theta_{47}) \dots U(0,10)$ (一様分布) からそれぞれの成分の初期値を発生させる

$u = (u_1, \dots, u_{47}) \dots N(0,5)$ (正規分布) からそれぞれの成分の初期値を発生させる

AR (1) 過程 u_{it} の係数について、 $\rho = 1, 0.75, 0.5$ の場合を考えて、それら III つのモデルの比較を行う。 $\rho = 1$ とした場合には、 u_{it} がランダムウォークに従うと仮定していることになる。すべてのモデルで推定を行うと、分析量が多くなる。そこで以下のような手順により、推定に用いるモデルを一つに絞ることとする。

STEP.1: 式 (6) の統計量 $\hat{R}_{i,14}$ により、マルコフ連鎖の定常分布への収束の具合を調べる。今

回の分析では、 $\hat{R}_{i,14}$ を小数第 1 位で四捨五入して、すべての都道府県で 1 になった場合に、定常分布へ収束していると判断することにした。

STEP.2：STEP.1 で残ったモデルに対して式 (7) の統計量 $d(\mathbf{y}^*, \mathbf{y}_{obs})$ 及び統計量 $\hat{p}$ を用いた優劣の比較を行う。

こうして選定した Rao-Yu モデルと比推定値について、ベンチマークとの比較及び誤差率の比較を行い、結果を考察する。ベンチマークとして用いるのは、以下の調査それぞれから計算した都道府県別の完全失業率である。

- ・平成 7 年国勢調査
- ・平成 12 年国勢調査
- ・平成 14 年就業構造基本調査

これらのベンチマークが真の値に近いものと想定し、ベンチマークとの距離を ARE (Absolute Relative Error) で測ることにより、比較を行った。ARE は以下の式で定義される。

$$ARE = \frac{|\theta_i^E - \theta_i^T|}{\theta_i^T} \quad (15)$$

$$\left[\theta_i^T : \text{ベンチマークの値} \right]$$

ベンチマークとなる調査の行われた年における、9 月の EBLUP 推定量と観測値で、どちらがベンチマークに近いかを比較する。なおベンチマークについては、小泉 (2004) の方法により、全国の結果同士が合うように調整を施している。

マルコフ連鎖モンテカルロ法のプログラム作成には、R を用いた。しかし R はインタープリタ言語であり、マルコフ連鎖モンテカルロ法のように、ループを多用する計算が遅いという特性がある。Rao-Yu モデルのような複雑なモデルを用いて推定を行う場合には、計算にかかる時間がさらに長くなる。今回の分析では、 $L = 10$ 、 $D = 2000$ と設定したときに、1 回の推定で約 50 分程度の時間がかかった。このことからすべての月について推定を行うことは困難であるので、分析はベンチマークとなる調査の行われた年に限っている。なお今回の計算に用いたプログラムについては、この章の最後に示している。

2 結果の考察

(1) 定常分布への収束及びモデルの選択

- ・まず、マルコフ連鎖が収束しているかどうかを見る。表 1 の R 値を見ると、 $\rho = 1$ とした場合にはどの時点でも R 値が大きくなっている。特に平成 7 年の熊本県 (都道府県番号 43 番) 及び鹿児島県 (46 番) では、R 値が 1.6 を超えてしまっている。よって $\rho = 1$ のときには初期値の影響が強く残っており、マルコフ連鎖が定常分布へ収束しているとはいえない状況となっている。一方 $\rho = 0.75, 0.5$ とした場合には、どの時点のどの都道府県でも、R 値がほぼ 1 に近くなっており、マルコフ連鎖は比較的良く定常分布へ収束しているといえる。よって、 $\rho = 0.75, 0.5$ とした場合について、更に分析を行うこととする。

- ・ $\rho=0.75, 0.5$ とした場合の d 値を比較すると、どの月においても $\rho=0.5$ としたときが、最も低くなっている。よって三つのモデルを相対的に比較した場合には、 $\rho=0.5$ としたモデルが、最も良く当てはまっているといえる。
- ・ ただし P 値を見ると、どの年のどのモデルにおいても 0.9 にかなり近い値となっており、絶対的に見た場合のモデルの当てはまりは、それ程良くないともいえる。 P 値が 0.9 に近くなった原因としては、観測値と補助情報間の相関関係の低下や、大阪府 (27 番)、沖縄県 (47 番) などのはずれ値の存在が挙げられる。

以上より、今回の分析では $\rho=0.5$ とした場合のモデルを後の分析で用いることにする。

(2) 各種推定値の比較

平成 7 年 9 月、平成 12 年 9 月、平成 14 年 9 月における各種推定値を表の形にまとめたものが表 3-1~3-3 である。またそれらをグラフにしたものが図 3~5 である。これらの結果から、以下のことが分かる。

- ・ 比推定値と Rao-Yu モデルの値を直接比較する。両方の推定値にあまり違いはなく、良く似た結果となっている。しかし平成 7 年と平成 12 年のグラフを見ると、もとの観測値よりも全体的に低めに推定されているように見える。
- ・ ARE を比較する。比推定値よりも Rao-Yu モデルの方が ARE の値が低いところについては、セルに色をつけて表示している。Rao-Yu モデルが比推定値よりもベンチマークに近いところの方が、どの時期においても多くなっている。グラフで見ると、平成 12 年の山梨県 (19 番) については、Rao-Yu モデルの方がかなり悪い結果になっている。しかしこれ以外のほぼすべての都道府県について、ARE は改善されている。特に島根県 (32 番)、徳島県 (36 番)、長崎県 (42 番)、鹿児島県では、時期によっては大幅な改善が見られる。
- ・ 誤差率を比較する。どちらの説明変数を用いた場合でも、すべての都道府県において Rao-Yu モデルの方が、比推定に比べて全体的に低い誤差率となっている。鳥取県、島根県、佐賀県 (41 番) などのサンプル数の少ない小地域では、誤差率が大幅に改善しているのが見て取れる。ただし山形県、栃木県、群馬県、愛知県 (23 番) と大阪府などの小地域では、一部の時期で観測値よりも誤差が高くなっている。Rao-Yu モデルの誤差率は MSE をもとにして計算しており、観測値の誤差率と比較してバイアスの分だけ高く出ていることが原因であると考えられる。またシミュレーションの誤差が考えられる。マルコフ連鎖のサンプルパスやヒストグラムを見ると、若干正規分布や逆ガンマ分布からのずれや、自己相関が見られるところがある。これについては、乱数を発生させる回数や一度に発生させるマルコフ連鎖の数を増やすなどの改良を加えれば、改善される可能性もある。ただし 1 回の推定にかかる時間は更に多くなると考えられるので、あまり実用的な方法ではない。

(3) まとめ

推定結果の比較についてまとめる。マルコフ連鎖モンテカルロ法について、 R 値による分析では、 $\rho=1$ の場合を除いて定常分布にほぼ収束しているとの結果になった。 d 値及び P 値による分析では、 $\rho=0.5$ のモデルが選択された。ただしどのモデルにおいても P 値は高い値を示し、モデルの当てはまりはそれ程良くないことが示唆された。マルコフ連鎖が定常

分布に収束比推定よりもベンチマークに近い都道府県の割合も高くなっている。ほとんどの小地域については、Rao-Yu モデルによる推定値の方が、良い結果であるといえる。誤差率はもとの観測値と比較して低くなっているものの、観測値よりも誤差率が悪化しているところ一部あった。有効求人倍率を用いたときのモデルの当てはまりは、全体的に見て、良いように思える。ただし観測値との相関関係は年々悪化してきており、その影響が推定値にも現れている。

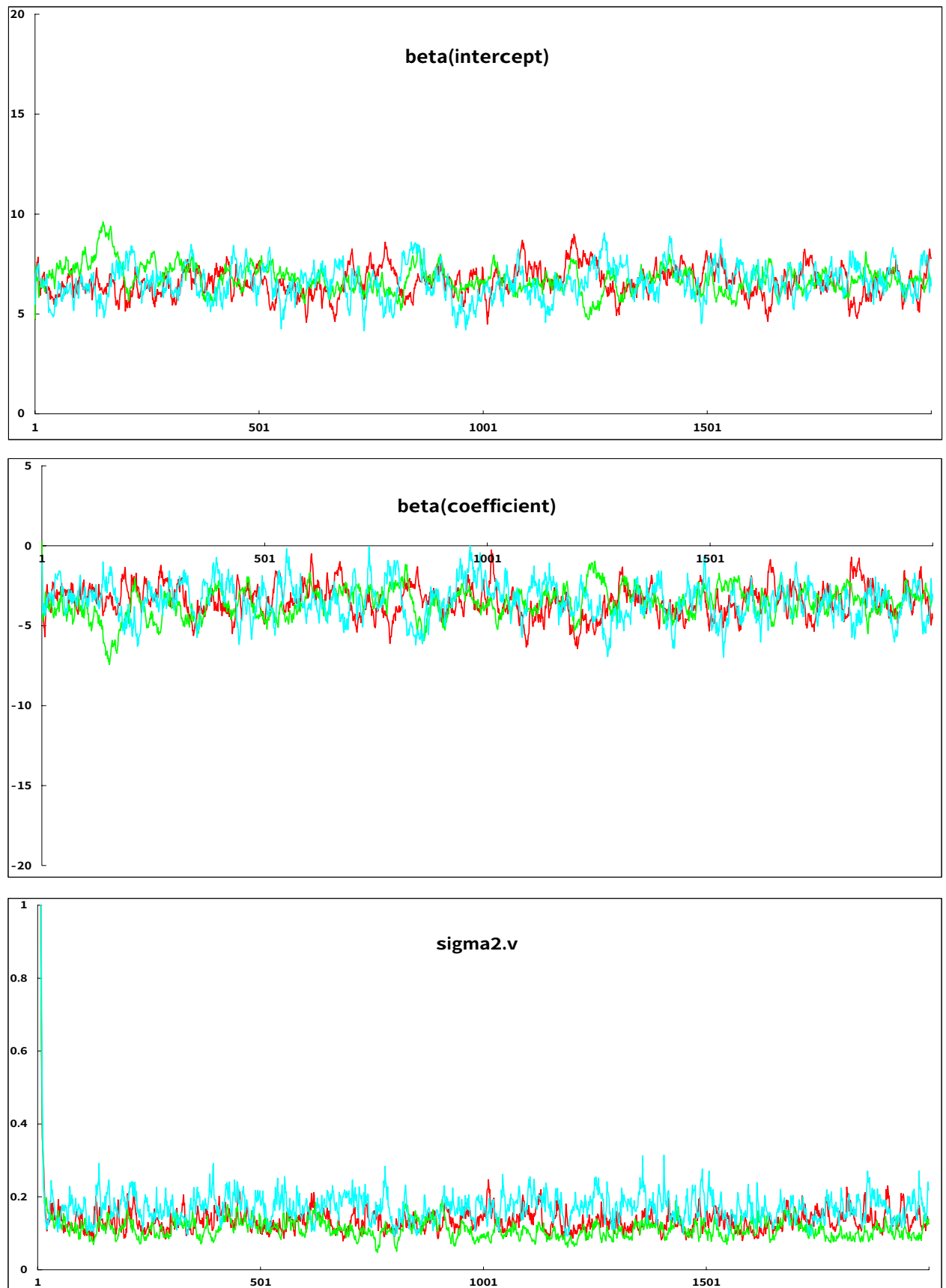
推定には R によるプログラムを用いた。しかし計算にかなりの時間がかかるので、この点についても何らかの改善が必要である。

VI 今後の課題

Time Series and Cross Sectional モデルを用いた小地域の完全失業率の分析には、Rao-Yu モデル以外にも類似の研究がある。Datta, Lahili, Maiti & Lu (1999)では、一度に 48 か月分のデータを用いたモデルを扱っている。推定に用いるデータが 48 か月分と長いので、この場合は季節性も考慮する必要がある。そこでモデルには月の効果、年の効果を扱うダミー変数を取り入れられている。この研究の労働力調査への適用を考える。ただしモデルはかなり複雑である。

VII 図表

図 1 各種パラメータのモンテカルロ標本のサンプルパス



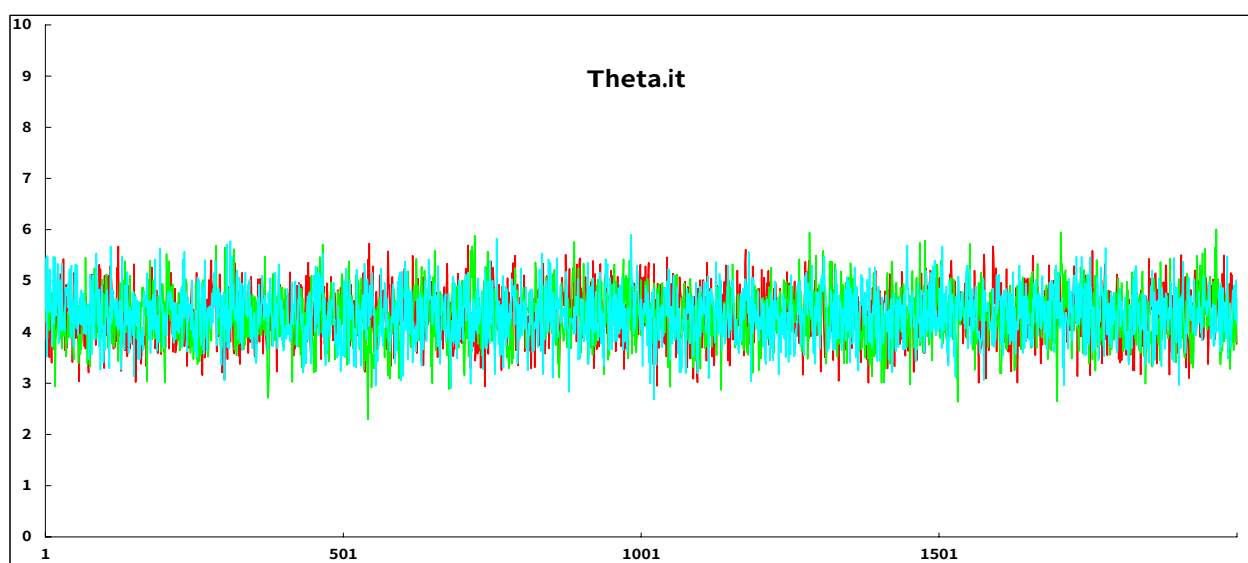
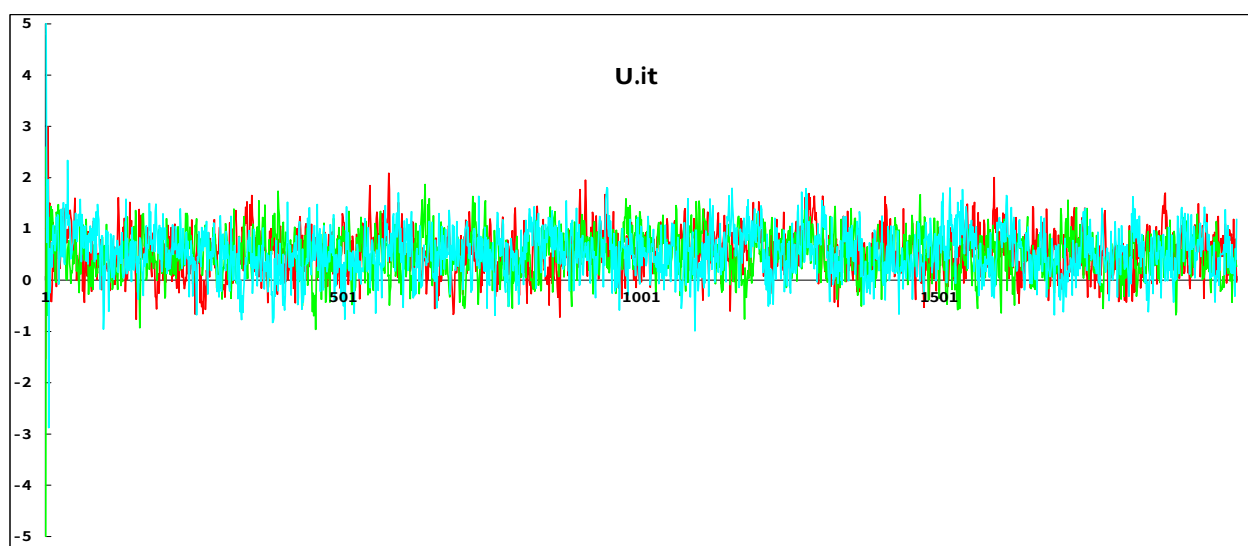
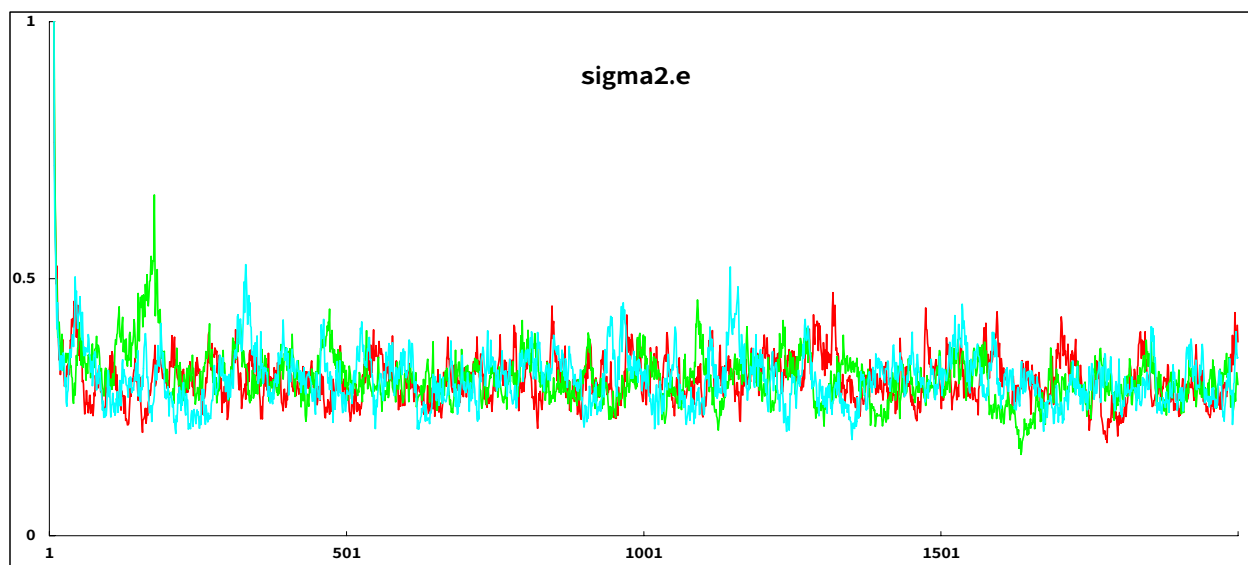
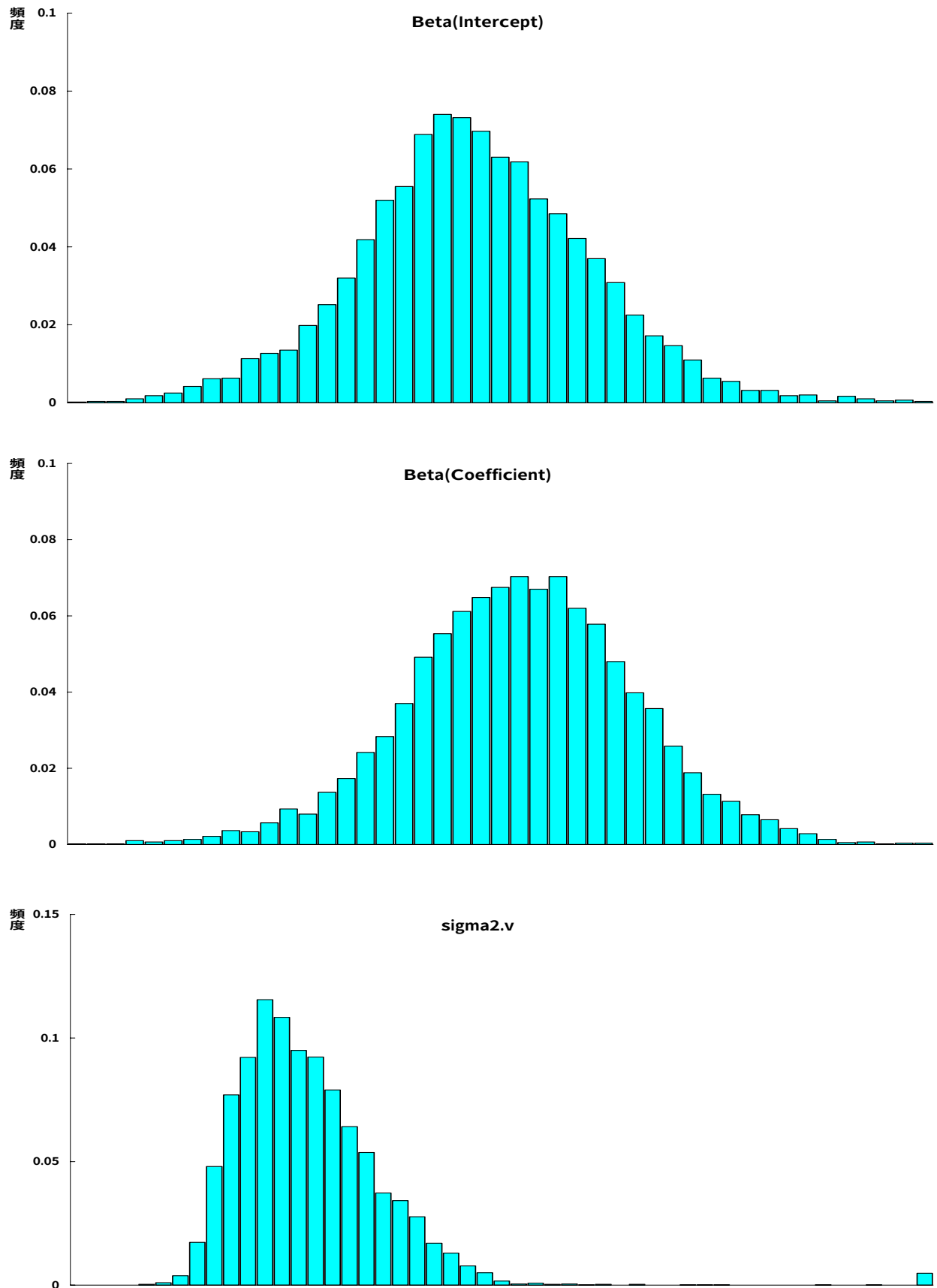


図 2 各種パラメータのモンテカルロ標本のヒストグラム



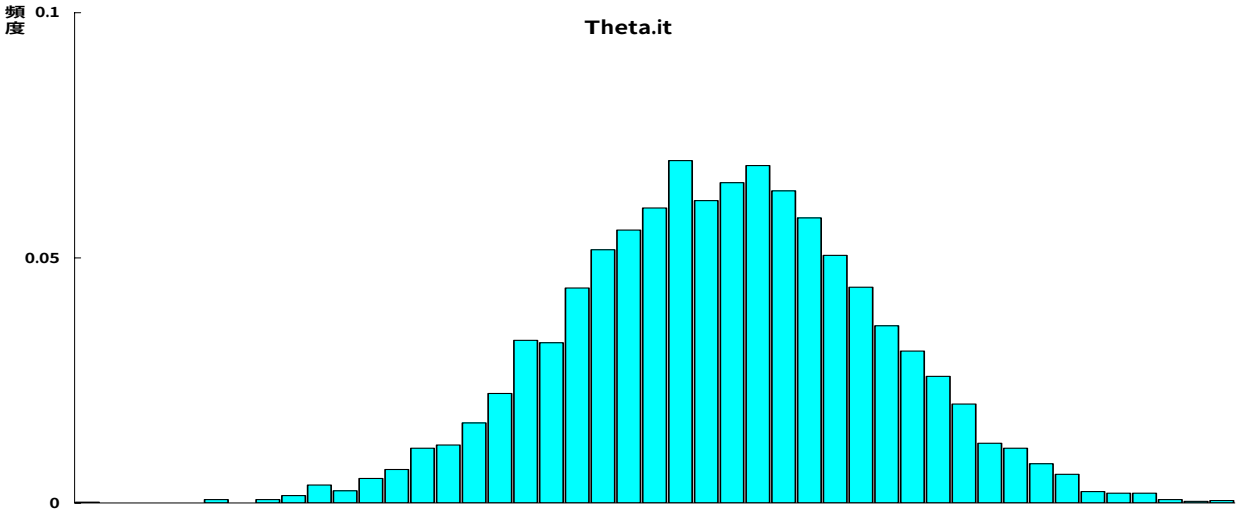
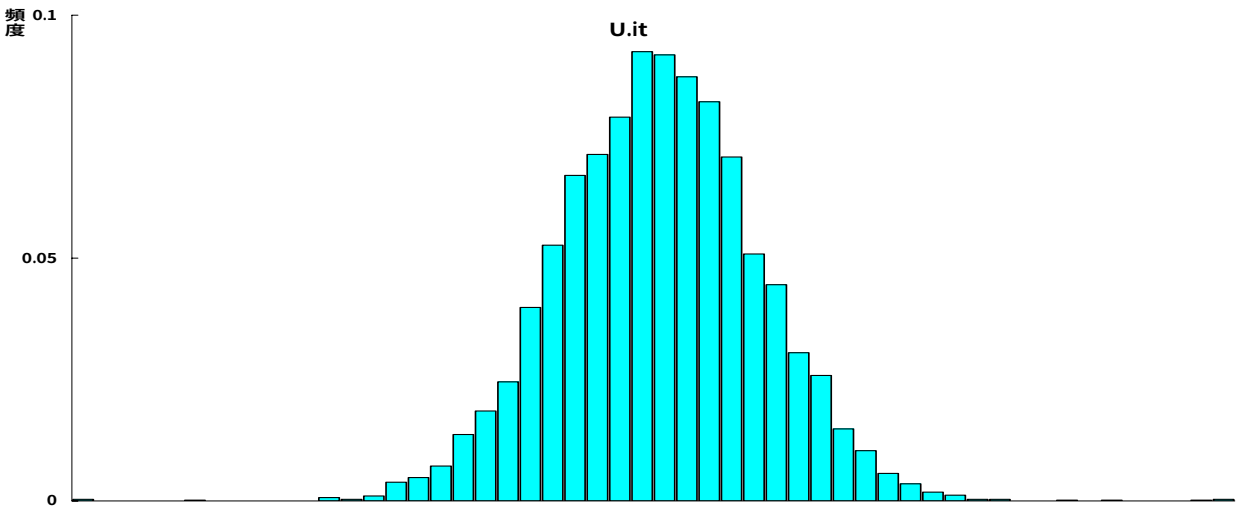
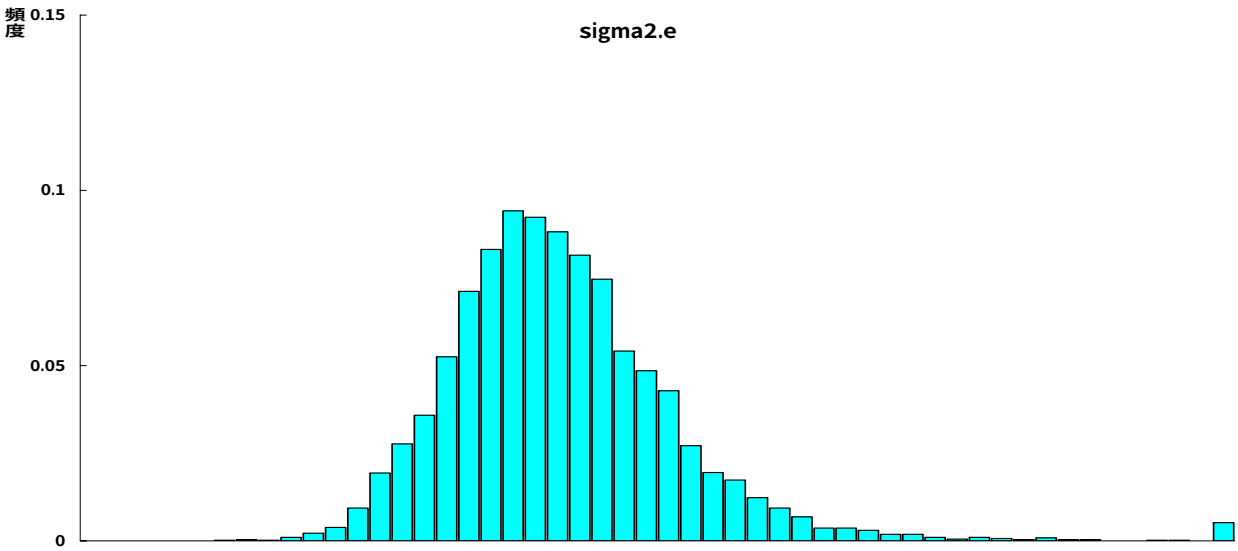


図3 都道府県別 各種推定値の比較

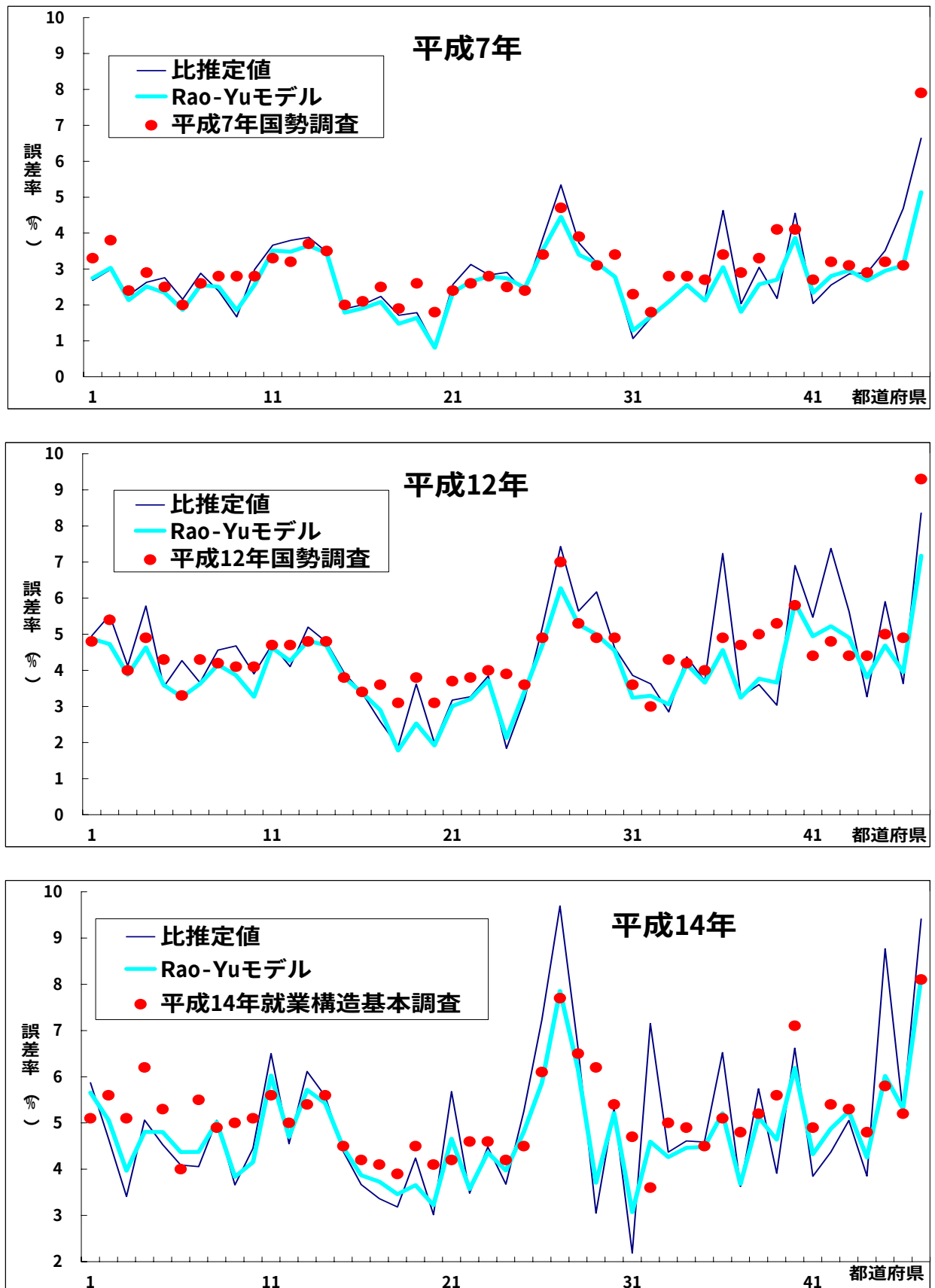


図 4 都道府県別 ARE の比較

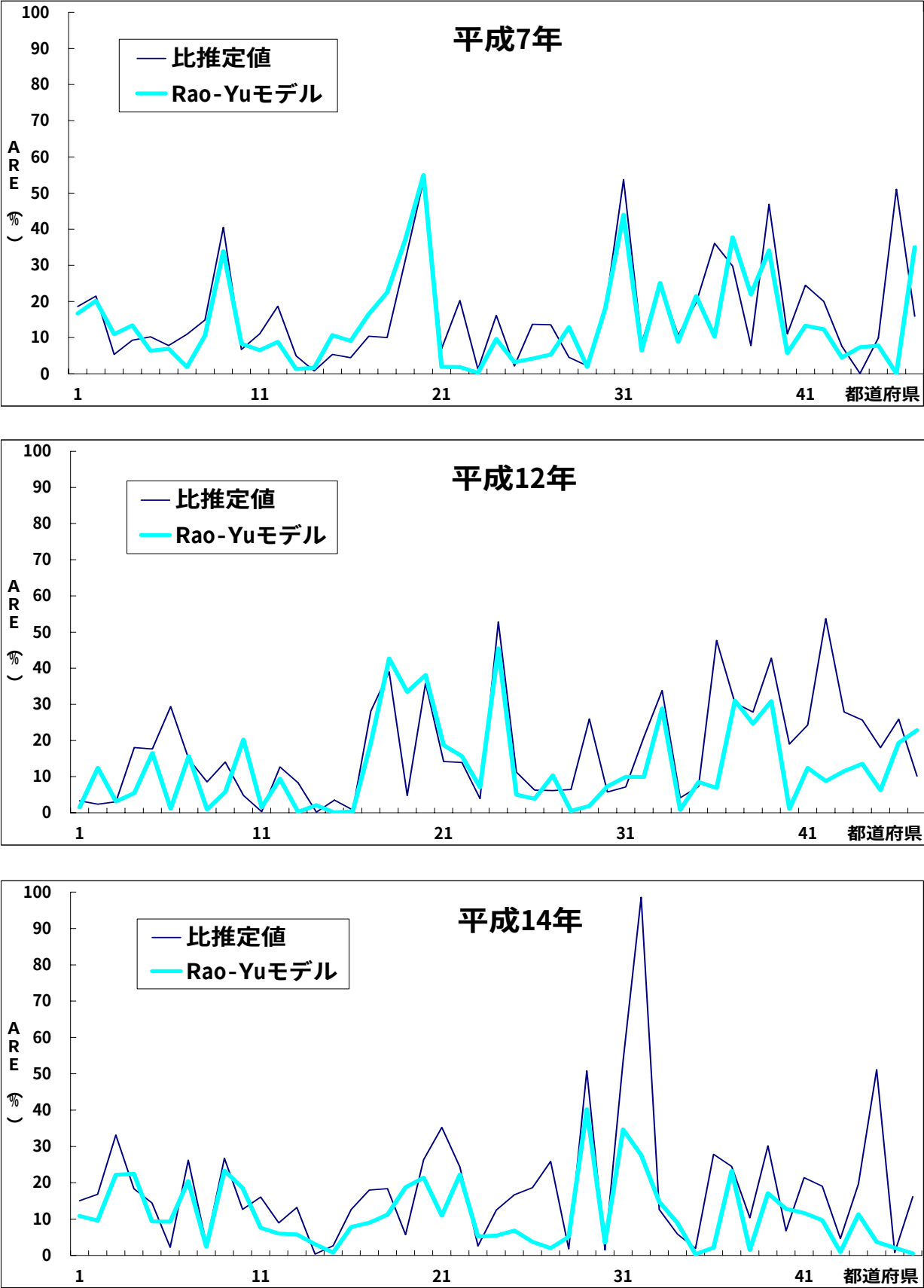


図 5 都道府県別 誤差率の比較

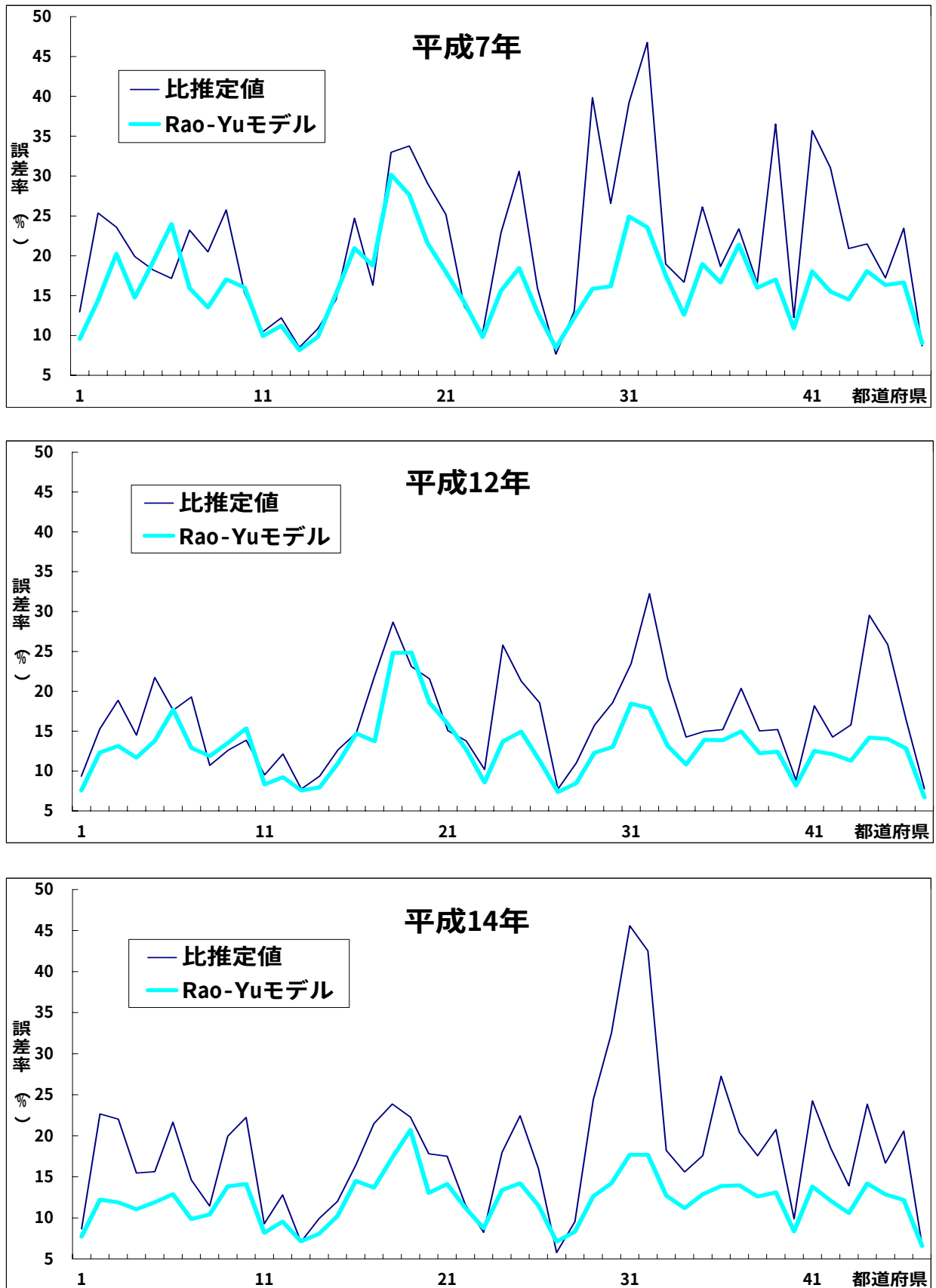


表 1 R 値の比較

		H7			H12			H14		
		$\rho = 0.5$	$\rho = 0.75$	$\rho = 1$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 1$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 1$
1	北海道	1.01	1.01	1.11	1.02	1.01	1.07	1.00	1.01	1.13
2	青森県	1.00	1.01	1.58	1.02	1.03	1.40	1.01	1.04	1.23
3	岩手県	1.01	1.01	1.25	1.01	1.02	1.11	1.02	1.01	1.11
4	宮城県	1.02	1.02	1.24	1.01	1.02	1.09	1.01	1.02	1.31
5	秋田県	1.02	1.02	1.43	1.02	1.02	1.12	1.01	1.01	1.27
6	山形県	1.02	1.02	1.36	1.01	1.05	1.11	1.03	1.03	1.33
7	福島県	1.01	1.03	1.46	1.00	1.02	1.10	1.00	1.01	1.12
8	茨城県	1.01	1.01	1.37	1.02	1.03	1.14	1.01	1.03	1.04
9	栃木県	1.00	1.01	1.10	1.02	1.05	1.12	1.02	1.04	1.15
10	群馬県	1.01	1.03	1.30	1.02	1.04	1.25	1.01	1.04	1.18
11	埼玉県	1.00	1.02	1.15	1.02	1.02	1.29	1.00	1.03	1.06
12	千葉県	1.01	1.00	1.40	1.00	1.01	1.04	1.01	1.01	1.14
13	東京都	1.00	1.01	1.13	1.02	1.01	1.10	1.00	1.03	1.06
14	神奈川県	1.01	1.01	1.21	1.02	1.02	1.08	1.00	1.02	1.07
15	新潟県	1.02	1.03	1.17	1.01	1.01	1.14	1.01	1.02	1.29
16	富山県	1.01	1.02	1.16	1.01	1.03	1.12	1.01	1.03	1.15
17	石川県	1.01	1.03	1.18	1.01	1.01	1.28	1.02	1.01	1.25
18	福井県	1.02	1.04	1.14	1.00	1.03	1.03	1.03	1.05	1.12
19	山梨県	1.01	1.04	1.11	1.01	1.06	1.18	1.01	1.13	1.08
20	長野県	1.01	1.01	1.10	1.02	1.04	1.14	1.01	1.03	1.05
21	岐阜県	1.01	1.01	1.08	1.01	1.03	1.07	1.01	1.06	1.11
22	静岡県	1.03	1.02	1.10	1.01	1.01	1.06	1.00	1.03	1.07
23	愛知県	1.00	1.00	1.18	1.00	1.02	1.07	1.00	1.04	1.04
24	三重県	1.01	1.01	1.43	1.01	1.01	1.17	1.01	1.04	1.23
25	滋賀県	1.02	1.03	1.56	1.01	1.02	1.22	1.01	1.02	1.18
26	京都府	1.00	1.01	1.31	1.02	1.01	1.22	1.02	1.02	1.30
27	大阪府	1.02	1.02	1.36	1.04	1.03	1.16	1.02	1.04	1.07
28	兵庫県	1.00	1.00	1.40	1.01	1.01	1.06	1.01	1.02	1.20
29	奈良県	1.01	1.01	1.40	1.03	1.03	1.50	1.02	1.03	1.22
30	和歌山県	1.01	1.01	1.59	1.01	1.04	1.13	1.01	1.06	1.42
31	鳥取県	1.00	1.02	1.29	1.01	1.02	1.19	1.02	1.02	1.21
32	島根県	1.01	1.01	1.40	1.02	1.07	1.25	1.01	1.07	1.10
33	岡山県	1.02	1.02	1.21	1.02	1.02	1.08	1.02	1.06	1.10
34	広島県	1.01	1.01	1.21	1.02	1.02	1.15	1.02	1.03	1.07
35	山口県	1.01	1.01	1.05	1.00	1.02	1.09	1.00	1.02	1.16
36	徳島県	1.02	1.00	1.42	1.01	1.04	1.32	1.03	1.06	1.22
37	香川県	1.02	1.03	1.46	1.01	1.04	1.15	1.01	1.05	1.34
38	愛媛県	1.01	1.02	1.43	1.01	1.02	1.11	1.01	1.04	1.19
39	高知県	1.01	1.01	1.51	1.02	1.05	1.17	1.02	1.02	1.24
40	福岡県	1.01	1.01	1.27	1.02	1.02	1.09	1.00	1.02	1.10
41	佐賀県	1.01	1.01	1.25	1.02	1.03	1.15	1.02	1.02	1.34
42	長崎県	1.01	1.01	1.32	1.03	1.04	1.14	1.01	1.02	1.11
43	熊本県	1.01	1.01	1.65	1.01	1.03	1.12	1.02	1.01	1.14
44	大分県	1.00	1.01	1.26	1.00	1.07	1.15	1.01	1.03	1.49
45	宮崎県	1.01	1.02	1.24	1.02	1.05	1.53	1.01	1.03	1.30
46	鹿児島県	1.01	1.02	1.67	1.01	1.02	1.20	1.02	1.03	1.09
47	沖縄県	1.01	1.02	1.23	1.02	1.02	1.03	1.02	1.02	1.10

表 2 P 値および d 値の比較

	H7			H12			H14		
	$\rho = 0.5$	$\rho = 0.75$	$\rho = 1$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 1$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 1$
事後予測P値	0.93	0.97	0.90	0.95	0.98	0.86	0.95	0.98	0.81
d	0.76	0.78	0.87	1.19	1.22	1.38	1.56	1.61	1.90

表3-1 平成7年結果 ($p = 0.5$)

$p=0.5$		ベンチ マーク (%)	推定値 (%)		ARE (%)		誤差率 (%)		R値
		H7国調	比推定	Rao-Yu	比推定	Rao-Yu	比推定	Rao-Yu	
1	北海道	3.3	2.7	2.7	18.6	16.7	13.0	9.5	1.01
2	青森県	3.8	3.0	3.0	21.5	20.2	25.4	14.4	1.00
3	岩手県	2.4	2.3	2.1	5.4	11.0	23.6	20.2	1.01
4	宮城県	2.9	2.6	2.5	9.3	13.4	19.9	14.7	1.02
5	秋田県	2.5	2.8	2.3	10.2	6.4	18.2	19.3	1.02
6	山形県	2.0	2.2	1.9	7.8	6.9	17.2	24.0	1.02
7	福島県	2.6	2.9	2.6	10.9	1.8	23.2	15.9	1.01
8	茨城県	2.8	2.4	2.5	14.9	10.6	20.5	13.5	1.01
9	栃木県	2.8	1.7	1.9	40.5	33.9	25.7	17.1	1.00
10	群馬県	2.8	3.0	2.6	6.7	8.3	15.4	16.0	1.01
11	埼玉県	3.3	3.7	3.5	11.0	6.5	10.5	9.9	1.00
12	千葉県	3.2	3.8	3.5	18.7	8.8	12.2	11.2	1.01
13	東京都	3.7	3.9	3.7	5.0	1.3	8.5	8.2	1.00
14	神奈川県	3.5	3.5	3.4	0.8	1.6	10.8	9.8	1.01
15	新潟県	2.0	1.9	1.8	5.4	10.7	14.5	15.3	1.02
16	富山県	2.1	2.0	1.9	4.4	9.1	24.7	20.9	1.01
17	石川県	2.5	2.2	2.1	10.4	16.5	16.3	18.8	1.01
18	福井県	1.9	1.7	1.5	10.0	22.5	33.0	30.2	1.02
19	山梨県	2.6	1.8	1.6	31.3	37.0	33.8	27.6	1.01
20	長野県	1.8	0.8	0.8	53.4	54.9	29.1	21.6	1.01
21	岐阜県	2.4	2.6	2.4	6.9	1.9	25.2	17.9	1.01
22	静岡県	2.6	3.1	2.6	20.3	1.9	13.6	14.1	1.03
23	愛知県	2.8	2.8	2.8	1.2	0.3	10.3	9.8	1.00
24	三重県	2.5	2.9	2.7	16.2	9.6	22.9	15.6	1.01
25	滋賀県	2.4	2.3	2.5	2.1	3.2	30.6	18.5	1.02
26	京都府	3.4	3.9	3.5	13.7	4.2	15.9	12.9	1.00
27	大阪府	4.7	5.3	4.5	13.6	5.3	7.7	8.5	1.02
28	兵庫県	3.9	3.7	3.4	4.6	12.9	13.0	12.2	1.00
29	奈良県	3.1	3.2	3.2	2.2	1.9	39.8	15.9	1.01
30	和歌山県	3.4	2.8	2.8	18.0	18.2	26.6	16.2	1.01
31	鳥取県	2.3	1.1	1.3	53.7	44.0	39.2	24.9	1.00
32	島根県	1.8	1.6	1.7	8.4	6.4	46.8	23.6	1.01
33	岡山県	2.8	2.1	2.1	25.2	25.1	19.0	17.6	1.02
34	広島県	2.8	2.5	2.6	10.7	8.8	16.7	12.6	1.01
35	山口県	2.7	2.2	2.1	19.7	21.5	26.1	19.0	1.01
36	徳島県	3.4	4.6	3.1	36.1	10.3	18.7	16.7	1.02
37	香川県	2.9	2.0	1.8	29.8	37.8	23.4	21.4	1.02
38	愛媛県	3.3	3.0	2.6	7.8	22.0	16.6	16.0	1.01
39	高知県	4.1	2.2	2.7	46.9	34.1	36.5	17.0	1.01
40	福岡県	4.1	4.6	3.9	11.0	5.7	12.2	10.9	1.01
41	佐賀県	2.7	2.0	2.3	24.5	13.3	35.7	18.0	1.01
42	長崎県	3.2	2.6	2.8	20.0	12.3	31.0	15.5	1.01
43	熊本県	3.1	2.9	3.0	7.6	4.5	20.9	14.5	1.01
44	大分県	2.9	2.9	2.7	0.1	7.4	21.5	18.1	1.00
45	宮崎県	3.2	3.5	3.0	9.9	7.8	17.2	16.3	1.01
46	鹿児島県	3.1	4.7	3.1	51.0	0.1	23.4	16.7	1.01
47	沖縄県	7.9	6.6	5.1	15.9	35.0	8.7	9.0	1.01

表 3-2 平成 12 年結果 ($\rho = 0.5$)

$\rho=0.5$		ベンチ マーク (%)	推定値 (%)		ARE (%)		誤差率 (%)		R値
		H12国調	比推定	Rao-Yu	比推定	Rao-Yu	比推定	Rao-Yu	
1	北海道	4.8	5.0	4.9	3.3	1.5	9.3	7.6	1.02
2	青森県	5.4	5.5	4.7	2.4	12.4	15.3	12.3	1.02
3	岩手県	4.0	4.1	3.9	3.0	3.1	18.9	13.1	1.01
4	宮城県	4.9	5.8	4.6	18.1	5.5	14.5	11.7	1.01
5	秋田県	4.3	3.5	3.6	17.6	16.5	21.7	13.8	1.02
6	山形県	3.3	4.3	3.3	29.4	1.1	17.6	17.7	1.01
7	福島県	4.3	3.7	3.6	14.8	15.7	19.3	12.9	1.00
8	茨城県	4.2	4.6	4.2	8.6	1.0	10.7	11.9	1.02
9	栃木県	4.1	4.7	3.9	14.1	5.7	12.6	13.5	1.02
10	群馬県	4.1	3.9	3.3	4.8	20.3	13.9	15.3	1.02
11	埼玉県	4.7	4.7	4.6	0.3	1.6	9.5	8.3	1.02
12	千葉県	4.7	4.1	4.3	12.7	9.4	12.1	9.2	1.00
13	東京都	4.8	5.2	4.8	8.3	0.3	7.7	7.6	1.02
14	神奈川県	4.8	4.8	4.7	0.2	2.1	9.4	8.0	1.02
15	新潟県	3.8	3.9	3.8	3.5	0.1	12.7	11.0	1.01
16	富山県	3.4	3.4	3.4	0.8	0.5	14.8	14.7	1.01
17	石川県	3.6	2.6	2.9	28.1	19.4	21.9	13.7	1.01
18	福井県	3.1	1.9	1.8	39.0	42.6	28.7	24.8	1.00
19	山梨県	3.8	3.6	2.5	4.8	33.4	23.1	24.8	1.01
20	長野県	3.1	2.0	1.9	35.8	38.1	21.6	18.6	1.02
21	岐阜県	3.7	3.2	3.0	14.2	18.7	15.0	15.9	1.01
22	静岡県	3.8	3.3	3.2	13.9	15.6	13.8	12.7	1.01
23	愛知県	4.0	3.8	3.7	3.9	7.0	10.2	8.6	1.00
24	三重県	3.9	1.8	2.1	52.8	45.5	25.8	13.7	1.01
25	滋賀県	3.6	3.2	3.4	11.2	5.0	21.3	14.9	1.01
26	京都府	4.9	5.2	4.7	6.3	3.9	18.6	11.3	1.02
27	大阪府	7.0	7.4	6.3	6.1	10.3	7.7	7.4	1.04
28	兵庫県	5.3	5.6	5.3	6.5	0.5	11.0	8.5	1.01
29	奈良県	4.9	6.2	5.0	26.0	1.9	15.7	12.3	1.03
30	和歌山県	4.9	4.6	4.5	5.8	7.3	18.6	13.0	1.01
31	鳥取県	3.6	3.9	3.2	7.2	10.0	23.5	18.4	1.01
32	島根県	3.0	3.6	3.3	20.9	9.9	32.2	17.9	1.02
33	岡山県	4.3	2.8	3.1	33.8	28.9	21.5	13.2	1.02
34	広島県	4.2	4.4	4.2	4.1	0.8	14.3	10.8	1.02
35	山口県	4.0	3.7	3.7	7.2	8.5	15.0	13.9	1.00
36	徳島県	4.9	7.2	4.6	47.7	6.9	15.2	13.9	1.01
37	香川県	4.7	3.3	3.2	30.4	31.0	20.4	15.0	1.01
38	愛媛県	5.0	3.6	3.8	27.9	24.7	15.0	12.2	1.01
39	高知県	5.3	3.0	3.7	42.8	30.9	15.2	12.4	1.02
40	福岡県	5.8	6.9	5.9	19.0	1.1	8.8	8.2	1.02
41	佐賀県	4.4	5.5	4.9	24.3	12.4	18.2	12.5	1.02
42	長崎県	4.8	7.4	5.2	53.7	8.8	14.3	12.1	1.03
43	熊本県	4.4	5.6	4.9	27.9	11.5	15.8	11.3	1.01
44	大分県	4.4	3.3	3.8	25.7	13.6	29.5	14.2	1.00
45	宮崎県	5.0	5.9	4.7	18.0	6.2	25.9	14.0	1.02
46	鹿児島県	4.9	3.6	4.0	25.9	19.3	16.4	12.8	1.01
47	沖縄県	9.3	8.4	7.2	10.2	22.9	7.8	6.7	1.02

表 3-3 平成 14 年結果 ($\rho = 0.5$)

$\rho=0.5$		ベンチ マーク (%)	推定値 (%)		ARE (%)		誤差率 (%)		R値
		H14就調	比推定	Rao-Yu	比推定	Rao-Yu	比推定	Rao-Yu	
1	北海道	5.1	5.9	5.7	15.1	10.9	8.7	7.7	1.00
2	青森県	5.6	4.7	5.1	16.8	9.6	22.7	12.2	1.01
3	岩手県	5.1	3.4	4.0	33.2	22.2	22.0	11.9	1.02
4	宮城県	6.2	5.1	4.8	18.3	22.5	15.5	11.1	1.01
5	秋田県	5.3	4.5	4.8	14.5	9.4	15.6	11.9	1.01
6	山形県	4.0	4.1	4.4	2.2	9.3	21.7	12.9	1.03
7	福島県	5.5	4.1	4.4	26.2	20.5	14.6	9.8	1.00
8	茨城県	4.9	5.1	5.0	3.2	2.4	11.4	10.4	1.01
9	栃木県	5.0	3.7	3.8	26.8	23.3	20.0	13.9	1.02
10	群馬県	5.1	4.5	4.2	12.7	18.6	22.2	14.1	1.01
11	埼玉県	5.6	6.5	6.0	16.1	7.6	9.3	8.2	1.00
12	千葉県	5.0	4.6	4.7	9.0	6.0	12.8	9.6	1.01
13	東京都	5.4	6.1	5.7	13.2	5.8	7.1	7.2	1.00
14	神奈川県	5.6	5.6	5.4	0.3	3.1	9.9	8.1	1.00
15	新潟県	4.5	4.4	4.5	2.7	0.8	12.0	10.2	1.01
16	富山県	4.2	3.7	3.9	12.6	7.8	16.4	14.5	1.01
17	石川県	4.1	3.4	3.7	18.0	9.0	21.5	13.7	1.02
18	福井県	3.9	3.2	3.5	18.4	11.2	23.9	17.3	1.03
19	山梨県	4.5	4.2	3.7	5.7	18.8	22.3	20.7	1.01
20	長野県	4.1	3.0	3.2	26.3	21.4	17.8	13.0	1.01
21	岐阜県	4.2	5.7	4.7	35.2	10.9	17.5	14.1	1.01
22	静岡県	4.6	3.5	3.6	24.3	22.3	11.6	11.2	1.00
23	愛知県	4.6	4.5	4.4	2.6	5.2	8.2	8.7	1.00
24	三重県	4.2	3.7	4.0	12.5	5.5	18.0	13.4	1.01
25	滋賀県	4.5	5.3	4.8	16.7	6.9	22.4	14.2	1.01
26	京都府	6.1	7.2	5.9	18.6	3.8	16.0	11.5	1.02
27	大阪府	7.7	9.7	7.9	25.9	2.0	5.8	7.1	1.02
28	兵庫県	6.5	6.6	6.2	1.8	5.2	9.6	8.3	1.01
29	奈良県	6.2	3.0	3.7	50.8	40.2	24.4	12.6	1.02
30	和歌山県	5.4	5.3	5.2	1.5	3.5	32.5	14.2	1.01
31	鳥取県	4.7	2.2	3.1	53.5	34.7	45.6	17.7	1.02
32	島根県	3.6	7.1	4.6	98.6	27.6	42.5	17.7	1.01
33	岡山県	5.0	4.4	4.3	12.6	14.8	18.2	12.7	1.02
34	広島県	4.9	4.6	4.5	5.9	8.9	15.6	11.2	1.02
35	山口県	4.5	4.6	4.5	1.9	0.3	17.6	12.9	1.00
36	徳島県	5.1	6.5	5.2	27.9	2.2	27.3	13.9	1.03
37	香川県	4.8	3.6	3.7	24.5	23.3	20.4	14.0	1.01
38	愛媛県	5.2	5.7	5.1	10.4	1.5	17.6	12.6	1.01
39	高知県	5.6	3.9	4.6	30.2	17.1	20.8	13.1	1.02
40	福岡県	7.1	6.6	6.2	6.8	12.8	9.9	8.4	1.00
41	佐賀県	4.9	3.8	4.3	21.4	11.7	24.2	13.8	1.02
42	長崎県	5.4	4.4	4.9	19.1	9.7	18.5	12.1	1.01
43	熊本県	5.3	5.1	5.2	4.6	1.0	13.9	10.6	1.02
44	大分県	4.8	3.9	4.3	19.7	11.3	23.8	14.2	1.01
45	宮崎県	5.8	8.8	6.0	51.1	3.7	16.7	12.9	1.01
46	鹿児島県	5.2	5.2	5.3	0.8	2.1	20.6	12.2	1.02
47	沖縄県	8.1	9.4	8.1	16.2	0.6	6.6	6.6	1.02

VIII 推定に用いたプログラム

今回の推定に用いたプログラムを以下に示す。**HB** という関数により、分析を行うことが出来る。**HB** の引数は以下のとおりである。

end : 推定を行う期間の最後の月を定める。データ開始時点からの月数で指定する。
rho : 時系列の効果の AR 係数を定める。 $-1 \leq \text{rho} \leq 1$ の範囲で選択する。
max.c : 発生させるマルコフ連鎖の総数を定める。
max.i : 一つのマルコフ連鎖の中で発生させるモンテカルロ標本の総数を定める。

プログラム中のデータの内容は、以下のとおりである。

Y.dt : 観測値 T 行 × 47 列の行列
X.dt : 補助情報 T 行 × 47 列の行列
E.dt : 標準誤差 T 行 × 47 列の行列
SEA : 誤差の系列相関 13 行 × 47 列の行列

ここで T は時系列の長さを表す。計算結果の出力は、以下のとおりである。

Z : 真の値の推定値。
V : Z の平均二乗誤差。
p : 事後予測 p 値。
d : d 統計量。
R : R 統計量。
b1 : 回帰係数 (定数)。
b2 : 回帰係数 (係数)。
vs : 回帰係数 (定数)。
es : 回帰係数 (定数)。
b1lk : **b1** のモンテカルロ標本。
b2lk : **b2** のモンテカルロ標本。
vs1k : **vs** のモンテカルロ標本。
es1k : **es** のモンテカルロ標本。
ulk : 時系列の効果 **u** のモンテカルロ標本。
Th1k : 真の値のモンテカルロ標本。
Z1k : 真の値のラオ・ブラックウェル推定に用いるモンテカルロ標本。

```
#####
# Software : 「R」
# Title : Rao-Yu Model Estimation
#         Using Markov Chain Monte Carlo Method
#####

library(MASS)

m <- 47
T <- 14
Y.dt <- matrix(scan("D:/uemp.r.txt"), ncol = m, byrow = TRUE)
X.dt <- matrix(scan("D:/job.r.txt"), ncol = m, byrow = TRUE)
E.dt <- matrix(scan("D:/uemp.r.err.TXT"), ncol = m, byrow = TRUE)
CV.dt <- E.dt / Y.dt
SEA <- matrix(scan("D:/uemp.r.SEA.TXT"), ncol = m, byrow = TRUE)
SEA <- apply(SEA, 1, mean)
Im <- matrix(rep(1, m))
IT <- matrix(rep(1, T))
ImT <- matrix(rep(1, m * T))

sigma.culc <- function (CV)
{
  res <- array(NA, c(T, T, m))
  for (i in 1:m) {
    work <- diag(T)
    for (lag in 1:(T - 1)) {
      work[abs(row(work) - col(work)) == lag] <- SEA[lag]
    }
    res[ , , i] <- work * crossprod(t(CV[i] * Y[, i]))
  }
  res
}

gen.b <- function (u.vec, Z.vec, X.vec, iXtX, vs)
{
  mu <- iXtX %*% crossprod(X.vec, Z.vec - u.vec)
  v <- vs * iXtX
  mvrnorm(1, mu, v)
}

gen.ivs <- function (u.vec, Z.vec, X.vec, b)
{
  s <- 0.001 + m * T / 2
  wrk <- Z.vec - X.vec %*% b - u.vec
}
```

```

    r <- 0.001 + crossprod(wrk) / 2
    rgamma(1, s, r)
  }

gen.ies <- function (rho, u.mat)
{
  s <- 0.001 + m * (T - 1) / 2
  wrk1 <- u.mat[1:(T - 1), , drop = FALSE]
  wrk2 <- u.mat[2:T, , drop = FALSE]
  wrk3 <- as.vector(wrk2 - rho * wrk1)
  r <- 0.001 + crossprod(wrk3) / 2
  rgamma(1, s, r)
}

gen.u <- function(rho, Z.mat, u.mat, b, ivs, ies)
{
  res <- matrix(NA, T, m)
  for (t in 1:T) {
    if (t == 1) {
      wrk1 <- rho ^ 2
      wrk2 <- u.mat[2,]
    } else if (t == T) {
      wrk1 <- 1
      wrk2 <- res[T - 1,]
    } else {
      wrk1 <- 1 + rho ^ 2
      wrk2 <- res[t - 1,] + u.mat[t + 1,]
    }
    v <- 1 / (ivs + wrk1 * ies)
    s <- sqrt(v)
    wrk3 <- matrix(Z.mat[t,])
    wrk4 <- cbind(Im, X[t,]) %*% b
    wrk5 <- matrix(wrk2) * rho * ies
    wrk6 <- (wrk3 - wrk4) * ivs
    mu <- (wrk5 + wrk6) * v
    res[t,] <- t(mu + matrix(rnorm(m, 0, s)))
  }
  res
}

gen.Z <- function(u.mat, b, ivs, ies)
{

```

```

thT <- ZT <- V1 <- V2 <- Ti <- di <- t(rep(NA, m))
Z <- matrix(NA, T, m)
for (i in 1:m) {
  Xi <- cbind(IT, matrix(X[, i]))
  Yi <- matrix(Y[, i])
  ui <- matrix(u.mat[, i])
  Zs.mat <- sig.mat[, , i]
  iZs.mat <- solve(Zs.mat)
  iZs <- solve(diag(ivs, T) + iZs.mat)
  V1[i] <- iZs[T, T]
  wrk1 <- iZs.mat %*% Yi
  wrk2 <- (Xi %*% b + ui) * ivs
  Zm <- iZs %*% (wrk1 + wrk2)
  ZT[i] <- Zm[T,]
  V2[i] <- Zm[T,] ^ 2
  Z[, i] <- riZ <- mvrnorm(1, Zm, iZs)
  Yp <- mvrnorm(1, riZ, Zs.mat)
  thT[i] <- Yp[T]
  G1 <- Yp - riZ
  G2 <- Yi - riZ
  wrk3 <- t(G1) %*% iZs.mat %*% G1
  wrk4 <- t(G2) %*% iZs.mat %*% G2
  Ti[i] <- wrk3 - wrk4
  di[i] <- crossprod(Yp - Yi)
}
d <- di %*% Im / (T * m)
p <- 0
if (Ti %*% Im > 0) p <- 1
list(Z, ZT, V1, V2, p, d, thT)
}

Gibbs.run <- function(rho, max.iter, chain)
{
  vslk <- eslk <- plk <- dlk <- matrix(rep(NA, max.iter))
  ulk <- Zlk <- V1 <- V2 <- thlk <- matrix(NA, max.iter, m)
  blk <- matrix(NA, max.iter, 2)
  vs <- init.vs[chain]
  u.mat <- init.u[, , chain]
  Z.mat <- init.Z[, , chain]
  X.vec <- cbind(ImT, matrix(X))
  XtX <- crossprod(X.vec)
  iXtX <- solve(XtX)
  for (i in 1:max.iter) {
    if (i %% 100 == 0) cat(" iter:", i, "¥n")

```

```

    u.vec <- matrix(u.mat)
    Z.vec <- matrix(Z.mat)
    blk[i,] <- b <- gen.b(u.vec, Z.vec, X.vec, iXtX, vs)
    b <- matrix(b)
    ivs <- gen.ivs(u.vec, Z.vec, X.vec, b)
    vslk[i,] <- 1 / ivs
    ies <- gen.ies(rho, u.mat)
    eslk[i,] <- 1 / ies
    u.mat <- gen.u(rho, Z.mat, u.mat, b, ivs, ies)
    ulk[i,] <- u.mat[T,]
    Z <- gen.Z(u.mat, b, ivs, ies)
    Z.mat <- Z[[1]]
    Zlk[i,] <- Z[[2]]
    V1[i,] <- Z[[3]]
    V2[i,] <- Z[[4]]
    plk[i,] <- Z[[5]]
    dlk[i,] <- Z[[6]]
    thlk[i,] <- Z[[7]]
  }
  list(blk, vslk, eslk, ulk, Zlk, V1, V2, plk, dlk, thlk)
}

Gibbs.chain <- function(rho, max.c, max.i)
{
  b1lk <- b2lk <- vslk <- eslk <- plk <- dlk <-
    matrix(NA, max.i, max.c)
  V1lk <- V2lk <- ulk <- Zlk <- thlk <-
    array(NA, c(max.i, m, max.c))
  b.o <- max.i / 2 + 1
  for (ch in 1:max.c){
    cat(" chain:", ch, "¥n")
    Gbs <- Gibbs.run(rho, max.i, ch)
    b1lk[, ch] <- Gbs[[1]][,1]
    b2lk[, ch] <- Gbs[[1]][,2]
    vslk[, ch] <- Gbs[[2]]
    eslk[, ch] <- Gbs[[3]]
    ulk[, , ch] <- Gbs[[4]]
    Zlk[, , ch] <- Gbs[[5]]
    V1lk[, , ch] <- Gbs[[6]]
    V2lk[, , ch] <- Gbs[[7]]
    plk[, ch] <- Gbs[[8]]
    dlk[, ch] <- Gbs[[9]]
  }
}

```

```

    thlk[, , ch] <- Gbs[[10]]
  }
  Z <- apply(Zlk[b.o:max.i,,], 2, mean)
  wrk1 <- apply(V1lk[b.o:max.i,,], 2, mean)
  wrk2 <- apply(V2lk[b.o:max.i,,], 2, mean)
  wrk3 <- apply(Zlk[b.o:max.i,,], 2, mean) ^ 2
  V <- wrk1 + wrk2 - wrk3
  b1 <- mean(b1lk[b.o:max.i,])
  b2 <- mean(b2lk[b.o:max.i,])
  vs <- mean(vslk[b.o:max.i,])
  es <- mean(eslk[b.o:max.i,])
  p <- mean(plk[b.o:max.i,])
  d <- mean(dlk[b.o:max.i,])
  W <- apply(apply(Zlk[b.o:max.i,,], c(2,3), var), 1, mean)
  B <- apply(apply(Zlk[b.o:max.i,,], c(2,3), mean), 1, var)
  R <- ((1 - 1 / (b.o - 1)) * W + B) / W
  cat(" finished", "\n")
  list(Z = Z, V = V, p = p, d = d, R = R, vs = vs, es = es,
       b1 = b1, b2 = b2, b1lk = b1lk, b2lk = b2lk, vslk = vslk,
       eslk = eslk, ulk = ulk, thlk = thlk, Zlk = Zlk)
}

HB <- function (end, rho, max.c, max.i)
{
  if (abs(rho)> 1) {
    stop(" Rho is not appropriate")
    break
  }
  start <- end - T + 1
  Y <- Y.dt[start:end, , drop = FALSE]
  X <- X.dt[start:end, , drop = FALSE]
  CV <- CV.dt[start:end, , drop = FALSE]
  CV <- apply(CV, 2, mean)
  sig.mat <- sigma.culc(CV)
  init.es <- runif(max.c, 0, 5)
  init.vs <- runif(max.c, 0, 5)
  num <- T * m * max.c
  dm <- c(T, m, max.c)
  init.u <- array(rnorm(num, 0, 5), dm)
  init.Z <- array(runif(num, 0, 10), dm)
  Gibbs.chain(rho, max.c, max.i)
}

H <- HB(57, 0.5, 10, 2000)

```

第 4 章 完全失業率の都道府県別月次標本誤差と

その系列相関の推定方法

本稿では第 1 章から第 3 章にかけて、様々な小地域推計の手法を紹介している。そのいずれの手法においても、推定には都道府県別・月次の標本誤差を必要とする。また第 1 章の時系列モデルによる推定、及び第 3 章の階層的ベイズモデルによる推定では、標本誤差の系列相関の推定値も必要になってくる。そこでこの章では、都道府県別・月次の標本誤差の推定方法、及び標本誤差の系列相関の推定方法についてまとめている。

Taylor 展開による線形化を基本とした石井 (1988) の方法を参考にして、標本誤差を推定する。完全失業率の推定に関しては、Wolter(1985) を参考にしていく。

標本誤差の系列相関の推定には、Pfeffermann, D. Feder, M. Signorelli, D. (1998) による方法を用いる。推定には、標本誤差の特徴を表すと考えられる “Pseudo Sampling Error” を利用する。その際に標本のローテーション構造を考慮した、“Stream” という概念が重要な役割を果たす。

I 労働力調査の標本誤差の推定

1 線形推定値の標本誤差の推定

労働力調査は全国約 4 万世帯 (約 10 万人) を層化 2 段抽出方により抽出する標本調査である。調査結果は、推計人口をベンチマーク人口とする比推定により算出される。ベンチマークとなる推計人口は、男女・年齢階級・地域 (7 大都市) を組み合わせた 56 区分に対して、繰り返し比例補正を用いて推定される。区分ごとに比推定を計算し、その合計を最終的な推定値とする。今回の推定では計算の簡単化のために、56 区分ごとの比推定は考慮せず、全体の推定値に対する分散のみを考えることにする。また自衛官・受刑者及び新設集団住宅地域については、サンプルから除いている。

以下の記述では、ある小地域のある層を考えることにする。層内において、属性 X を有する人口の線形推定値を $\hat{x}$ とし、層内の母集団総数の線形推定値を $\hat{y}$ とする。このとき各線形推定値、 $\hat{x}$ の分散 $\text{var}(\hat{x})$ 及び $\hat{x}$ 、 $\hat{y}$ の共分散 $\text{cov}(\hat{x}, \hat{y})$ は、以下のようにして推定できる (石井(1988))。

$$\text{var}(\hat{x}) = \frac{1}{m-1} \sum_{i=1}^m \left(w \sum_{j=1}^{n_i} x_{ij} - \hat{x} \right)^2 \quad (1)$$

$$\text{cov}(\hat{x}, \hat{y}) = \frac{1}{m-1} \sum_{i=1}^m \left(w \sum_{j=1}^{n_i} x_{ij} - \hat{x} \right) \left(w \sum_{j=1}^{n_i} y_{ij} - \hat{y} \right) \quad (2)$$

$$\hat{x} = \frac{w}{m} \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij} \quad : \text{ある小地域のある層における完全失業者数の線形推定値} \quad (3)$$

$$\hat{y} = \frac{w}{m} \sum_{i=1}^m \sum_{j=1}^{n_i} y_{ij} \quad : \text{ある小地域のある層における母集団総数の線形推定値} \quad (4)$$

$$\left[\begin{array}{l} x_{ij} : \text{第 } i \text{ 調査区、第 } j \text{ 標本住戸内の、属性 } X \text{ を有する調査人口の推定値} \\ y_{ij} : \text{第 } i \text{ 調査区、第 } j \text{ 標本住戸内の調査人口} \\ n_i : \text{第 } i \text{ 調査区内の標本住戸数} \\ m : \text{層内の標本調査区数} \\ w : \text{層内の標本調査区のウェイト（抽出率の逆数）の合計} \end{array} \right]$$

労働力調査の場合、標本設計は地域別に行われており、都道府県別には行われていない。都道府県別の比推定用乗率は、線形推定乗率に全国のものを用いて、ベンチマークのみ都道府県の推計人口を用いて算出される。つまり都道府県という単位を、地域の中のクロス項目としてみていることになる。よって式 (1) ～ (4) で、該当する都道府県以外の x_{ij} 及び y_{ij} を 0 と置くことで、都道府県別の線形推定値を近似的に求めることができる。

2 比推定値の標本誤差の推定

これを用いて、都道府県別完全失業率の比推定値の標本誤差を推定する。まず 2 変数関数 $g(x, y)$ による曲面 Γ を考える。点 (x^*, y^*) の近傍における $g(x, y)$ の値を、 Γ 上の点 $(x^*, y^*, g(x^*, y^*))$ における接平面で近似すると、以下ようになる。

$$g(x, y) = g(x^*, y^*) + (x - x^*) \frac{\partial g}{\partial x}(x^*, y^*) + (y - y^*) \frac{\partial g}{\partial y}(x^*, y^*) \quad (5)$$

よって確率変数 $g(\hat{x}, \hat{y})$ を、 $\bar{x} = E(\hat{x})$ 、 $\bar{y} = E(\hat{y})$ の近傍で、以下のように近似することができる。

$$g(\hat{x}, \hat{y}) = g(\bar{x}, \bar{y}) + (\hat{x} - \bar{x}) \frac{\partial g}{\partial x}(\bar{x}, \bar{y}) + (\hat{y} - \bar{y}) \frac{\partial g}{\partial y}(\bar{x}, \bar{y}) \quad (6)$$

式 (6) によって、 $g(\hat{x}, \hat{y})$ の分散の近似値を次のように推定することができる。

$$\begin{aligned} & \text{var}(g(\hat{x}, \hat{y})) \\ &= \text{var}(\hat{x}) \left[\frac{\partial g}{\partial x}(\bar{x}, \bar{y}) \right]^2 + 2 \left[\frac{\partial g}{\partial x}(\bar{x}, \bar{y}) \right] \left[\frac{\partial g}{\partial y}(\bar{x}, \bar{y}) \right] \text{cov}(\hat{x}, \hat{y}) + \text{var}(\hat{y}) \left[\frac{\partial g}{\partial y}(\bar{x}, \bar{y}) \right]^2 \end{aligned} \quad (7)$$

層内のベンチマーク人口を Y とした場合には、その層の比推定値は $Y \frac{\hat{x}}{\hat{y}}$ と表される。よって式

(7) で $g(\hat{x}, \hat{y}) = Y \frac{\hat{x}}{\hat{y}}$ と置くことにより、次のようにして比推定値の分散を推定できる。

$$\begin{aligned}\text{var}\left(Y \frac{\hat{x}}{\hat{y}}\right) &= \left[\text{var}(\hat{x}) \left(\frac{Y}{\bar{y}}\right)^2 - 2 \left(\frac{Y}{\bar{y}}\right) \left(-\frac{\bar{x}Y}{\bar{y}^2}\right) \text{cov}(\hat{x}, \hat{y}) + \text{var}(\hat{y}) \left(-\frac{\bar{x}Y}{\bar{y}^2}\right)^2 \right] \\ &= \left(\frac{Y}{\bar{y}}\right)^2 [\text{var}(\hat{x}) - 2\bar{R} \text{cov}(\hat{x}, \hat{y}) + \bar{R}^2 \text{var}(\hat{y})] \quad (\bar{R} = \frac{\bar{x}}{\bar{y}})\end{aligned}\quad (8)$$

式 (8) において $\bar{X}$ を $\hat{X}$ で、 $\bar{Y}$ を $\hat{Y}$ で置き換えた以下の式により、比推定の標本誤差の分散を推定することができる。

$$\left(\frac{Y}{\hat{y}}\right)^2 [\text{var}(\hat{x}) - 2\hat{R} \text{cov}(\hat{x}, \hat{y}) + \hat{R}^2 \text{var}(\hat{y})] \quad (\hat{R} = \frac{\hat{x}}{\hat{y}}) \quad (9)$$

以上は、層内における比推定値の分散の推定方法であった。層間では抽出は独立に行われていると考えられる。よって全体の分散は、各層における比推定の分散の合計により、求めることができる。

3 完全失業率比推定値の標本誤差の推定

$\hat{Z}$ を完全失業者数の比推定値とし、 $\hat{W}$ を労働力人口の比推定値とする。 $\hat{Z}$ と $\hat{W}$ の分散は、それぞれ式 (9) により推定することが出来る。 $\hat{R} = \frac{\hat{Z}}{\hat{W}}$ としたときに、 $\hat{R}$ と $\hat{W}$ は無相関であると仮定する。このとき式 (9) で $g(\hat{Z}, \hat{W}) = \frac{\hat{Z}}{\hat{W}}$ と置くことで、以下のようにして完全失業率の比推定値の標本誤差を推定することができる (Wolter(1985))。

$$\begin{aligned}\text{var}\left(\frac{\hat{Z}}{\hat{W}}\right) &= \left(\frac{1}{\bar{w}}\right)^2 [\text{var}(\hat{Z}) - 2\bar{R} \text{cov}(\hat{Z}, \hat{W}) + \bar{R}^2 \text{var}(\hat{W})] \\ &= \left(\frac{1}{\bar{w}}\right)^2 [\text{var}(\hat{Z}) - 2\bar{R} \text{cov}(\hat{R}\hat{W}, \hat{W}) + \bar{R}^2 \text{var}(\hat{W})] \quad (\hat{Z} = \hat{R}\hat{W}) \\ &= \left(\frac{1}{\bar{w}}\right)^2 [\text{var}(\hat{Z}) - \bar{R}\hat{R} \text{var}(\hat{W})] \quad (\text{cov}(\hat{R}\hat{W}, \hat{W}) = \hat{R} \text{cov}(\hat{W}, \hat{W}))\end{aligned}\quad (10)$$

最後の等式は、 $\hat{R}$ と $\hat{W}$ が無相関であることによる。式(10)において $\bar{Z}$ を $\hat{Z}$ で、 $\bar{W}$ を $\hat{W}$ で置き換えた以下の式により、完全失業率の比推定値の標本誤差を推定することができる。

$$\left(\frac{1}{\hat{w}}\right)^2 [\text{var}(\hat{Z}) - \hat{R}^2 \text{var}(\hat{W})] \quad (11)$$

以上は、層内における比推定値の分散の推定方法であった。層間では抽出は独立に行われていると考えられる。よって全体の分散は、各層における比推定の分散の合計により推定できる。

II 標本誤差の系列相関の推定

1 Stream の概念

“真の標本誤差”を直接観測することができないので、標本誤差の自己相関の推定は困難である。そこで標本誤差の近似として Pseudo Sampling Error というものを考え、これをもとにして標本誤差の自己相関を推定する (Pfeffermann, D. Feder, M. Signorelli, D. (1998))。Pseudo Sampling Error の推定には、標本のローテーション構造を利用する。その際に “Stream” という概念が重要な役割を果たす。Stream は調査回数をもとに副標本を組み合わせたもので、同一の Stream に属する PSE は相関を持ち、異なる Stream 間の Pseudo Sampling Error は無相関であると仮定される。

この節では Stream の概念について述べる。各副標本はいくつかの調査区が集まってできている。我が国の労働力調査では、各調査区は 4 か月連続で調査された後、8 か月調査から外れ、その後再び 4 か月連続で調査される。また一つの調査区を前期と後期に分けて、そこから別々に標本住戸を抽出する。よって調査対象となった住戸は、1 年目に 2 か月連続して調査を受け、その後 10 か月間調査から外れ、2 年目の同じ時期に再び連続して 2 か月間、合計で 4 か月間調査を受けるという、2-10-2 のローテーション構造になる。以上のことから、ある時点を固定したときに、そこから 1 期前及び 12 期前とは半分ずつ、11 期前及び 13 期前とは 4 分の 1 ずつ、標本住戸の重なりがあることが分かる (図 1 参照)。標本住戸が重なっている部分は、正の相関があると考えられる。よって標本誤差の系列相関は、1 次・11 次・12 次・13 次で高い値を示すと予想される (図 2 参照)。

労働力調査の月ごとの標本住戸は、調査を開始した時期及び調査回数をもとにして、以下のような 8 組の副標本に区別される (総務省統計局 (2002)、総務省統計局 (2005))。

- ・ 1、5、9 月調査開始住戸 (1 年目調査) (組符号 A-1)
- ・ 1、5、9 月調査開始住戸 (2 年目調査) (組符号 A-2)
- ・ 2、6、10 月調査開始住戸 (1 年目調査) (組符号 B-1)
- ・ 2、6、10 月調査開始住戸 (2 年目調査) (組符号 B-2)
- ・ 3、7、11 月調査開始住戸 (1 年目調査) (組符号 C-1)
- ・ 3、7、11 月調査開始住戸 (2 年目調査) (組符号 C-2)
- ・ 4、8、12 月調査開始住戸 (1 年目調査) (組符号 D-1)
- ・ 4、8、12 月調査開始住戸 (2 年目調査) (組符号 D-2)

標本住戸について、副標本ごとの調査回数を示したものが図 2 である。セル内の数値は、調査を受けた回数を表している。時点を固定した場合の各副標本について、同じ数値を持つようなセルは無いことが分かる。よって調査を受けた回数で、各調査区がどの副標本に属しているかを区別することができる。

副標本それぞれから、全国の完全失業率を推定することができる。こうして完全失業率の推定値の時系列が 8 組できる。それぞれの時系列について、調査回数が $\cdots \rightarrow 1 \rightarrow 2 \rightarrow \cdots \rightarrow 7 \rightarrow 8 \rightarrow 1 \rightarrow \cdots$ の繰り返しになるように組み合わせるものを、Stream と呼ぶ (図 3 参照)。なお図 3 では、調査開始時点における調査回数を、その Stream の名前としている。同一世帯 (住戸) はすべて同じ Stream 内に属しており、その調査回数ごとの様子は次のようになっている (図 1 参照)。矢印は調査回数の変化を表す。

- ・ 1→2、3→4、5→6、7→8 では、同じ住戸が 2 か月連続しており、標本誤差には強い系列相関があると考えられる。
- ・ 2→3、6→7 では前期と後期の境目である。よって標本として抽出される住戸は異なるものの、同じ調査区から抽出されているので、多少の相関があると考えられる。
- ・ 4→5 は、1 年目と 2 年目の境目である。これらは同じ住戸なので、系列相関があると考えられる。
- ・ 8→1 では、標本調査区が完全に交代する。しかし新たな調査区は近くの地域から選ばれる場合が多いので、この時点でも若干の系列相関があると考えられる。

以上の性質により、同一の Stream 内における標本誤差の自己相関は高いと考えられる。一方、異なる Stream は地理的に離れているので、相関が無いと仮定される。

2 Pseudo Sampling Error を用いた系列相関の推定

前節で定義した Stream をもとに Pseudo Sampling Error と呼ばれる量を計算し、それをもとにして標本誤差の系列相関を推定する。以下、その方法について述べる。

副標本ごとに推定された 8 組の完全失業率の時系列は、それぞれ図 4-1 のように並んでいる。調査の回数によって標本誤差の水準は異なると考えられるので、 $e^j(t)$ を求める際には、図 4-1 を図 4-2 のように調査回数ごとに並べ替える。

このように並び替えた副標本に対して、Pseudo Sampling Error $e^j(t)$ は、以下のようにして定義される。

$$e^j(t) = y^j(t) - \bar{y}(t) - [\bar{y}^j - \bar{y}] \quad (12)$$

$$\left[\begin{array}{l} y^j(t) : \text{時点 } t \text{ で } j \text{ 回目の調査を受けた副標本による完全失業率の推定値} \\ \bar{y}(t) = \frac{1}{8} \sum_{j=1}^8 y^j(t) , \quad \bar{y}^j = \frac{1}{T} \sum_{t=1}^T y^j(t) , \quad \bar{y} = \frac{1}{8T} \sum_{t=1}^T \sum_{j=1}^8 y^j(t) \\ e^j(t) : \text{時点 } t \text{ で } j \text{ 回目の調査を受けた Stream の Pseudo Sampling Error} \end{array} \right]$$

$e^j(t)$ は、同一時点における副標本間の平均からの乖離として定義されている。これは標本誤差の代用とみなすことができる。

標本数の少ない小地域では、以下のような状況が発生する場合がある。

- ・ 小地域や時点によっては、副標本の数が 8 に満たなくなる場合がある。
- ・ 項目によっては、推定値が 0 になってしまう。

このような場合には、該当する副標本の値を、同一時点における副標本内の平均 $\bar{y}(t)$ で置き換えて対応している。

労働力調査の標本誤差は時間とともに変化すると考えられるので、 $e^j(t)$ の値が安定しなくなる可能性がある。そこで $e^j(t)$ を各時点における標準偏差で割った $\bar{e}^j(t)$ を考える。

$$\bar{e}^j(t) = \frac{e^j(t)}{\sqrt{\frac{1}{8-1} \sum_{s=1}^8 e^j(t)}} \quad (13)$$

$\bar{e}^j(t)$ の変動は安定していると考えられる。

$\bar{e}^j(t)$ をもとにして、標本誤差の系列相関を求めることができる。副標本を、図 4-3 のように、stream に組み直す。なお図 4-3 では各 stream について、時系列の開始時点における調査回数を、その stream の名前としている。stream ごとの自己相関係数 C_k^s は、 $\bar{e}^j(t)$ を用いて以下のように推定することができる。

$$C_k^s = \frac{\sum_{t=k+1}^T \bar{e}^j(t) \bar{e}^j(t-k)}{\sum_{t=k+1}^T \bar{e}^j(t)^2} \quad (14)$$

前節でも述べたように、別々の Stream 間では自己相関がないと考えられる。よって C_k^s をもとにして、次のように標本誤差の自己相関係数 $\hat{\rho}(k)$ を推定することができる (Pfeffermann, D. Feder, M. Signorelli, D. (1998))。

$$\hat{\rho}(k) = \frac{\sum_{s=1}^8 \bar{C}_k^s}{\sum_{s=1}^8 \bar{C}_0^s} \quad (15)$$

こうして求めた自己相関 $\hat{\rho}(k)$ のうち、1 次から 13 次までの値を、各手法の推定に用いることとする。

Pseudo Sampling Error を用いた標本誤差の系列相関の推定では、標本住戸を八つの Stream に分けてしまう。都道府県別の標本数はもともと少ないので、Stream ごとの標本数はかなり少なくなる場合がある。しかし十分多い年数のデータを用いることで Stream ごとの標本数が増し、安定し

た推定量が得られると考えられる。今回は平成 3 年 1 月から平成 15 年 12 月までの 13 年分の観測値を用いて、標本誤差の系列相関を推定した。13 年分のデータを用いれば、推定値は十分安定していると考えられる。

都道府県別の誤差の自己相関をグラフにしたものが図 5 である。ほぼすべての地域で、図 2 で予想したような系列相関が得られている。各都道府県の 1 次から 14 次までの系列相関をまとめたものが表 1 である。

組符号	ある月 (現在)														ある月 (1年前)			
	1か月後		1か月前	2か月前	3か月前	4か月前	5か月前	6か月前	7か月前	8か月前	9か月前	10か月前	11か月前	12か月前	13か月前	14か月前		
A-1 (1年目)																		
A-2 (2年目)																		
B-1 (1年目)																		
B-1 (2年目)																		
C-1 (1年目)																		
C-1 (2年目)																		
D-1 (1年目)																		
D-1 (2年目)																		

 は、同じ住戸を表す

後期 前期
 2ヶ月目 1ヶ月目
 同一調査区

自己相関係数

1

0.5

0

0

0

0

0

0

0

0

0

0

0.25

0.5

0.25

0

次数

次数	自己相関係数
0	1
1	0.5
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0.25
12	0.5
13	0.25
14	0

図 3 “Stream”

組符号	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月	4月
A-1 (1年目)	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
A-2 (2年目)	5	6	7	8	5	6	7	8	5	6	7	8	5	6	7	8
B-1 (1年目)	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3
B-1 (2年目)	8	5	6	7	8	5	6	7	8	5	6	7	8	5	6	7
C-1 (1年目)	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2
C-1 (2年目)	7	8	5	6	7	8	5	6	7	8	5	6	7	8	5	6
D-1 (1年目)	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1
D-1 (2年目)	6	7	8	5	6	7	8	5	6	7	8	5	6	7	8	5

※ セル内の数値は調査回数を表す。

図 4-1 Pseudo Sampling Error の求め方①

組符号	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月	4月
A-1 (1年目)	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
A-2 (2年目)	5	6	7	8	5	6	7	8	5	6	7	8	5	6	7	8
B-1 (1年目)	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3
B-1 (2年目)	8	5	6	7	8	5	6	7	8	5	6	7	8	5	6	7
C-1 (1年目)	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2
C-1 (2年目)	7	8	5	6	7	8	5	6	7	8	5	6	7	8	5	6
D-1 (1年目)	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1
D-1 (2年目)	6	7	8	5	6	7	8	5	6	7	8	5	6	7	8	5

図 4-2 Pseudo Sampling Error の求め方②

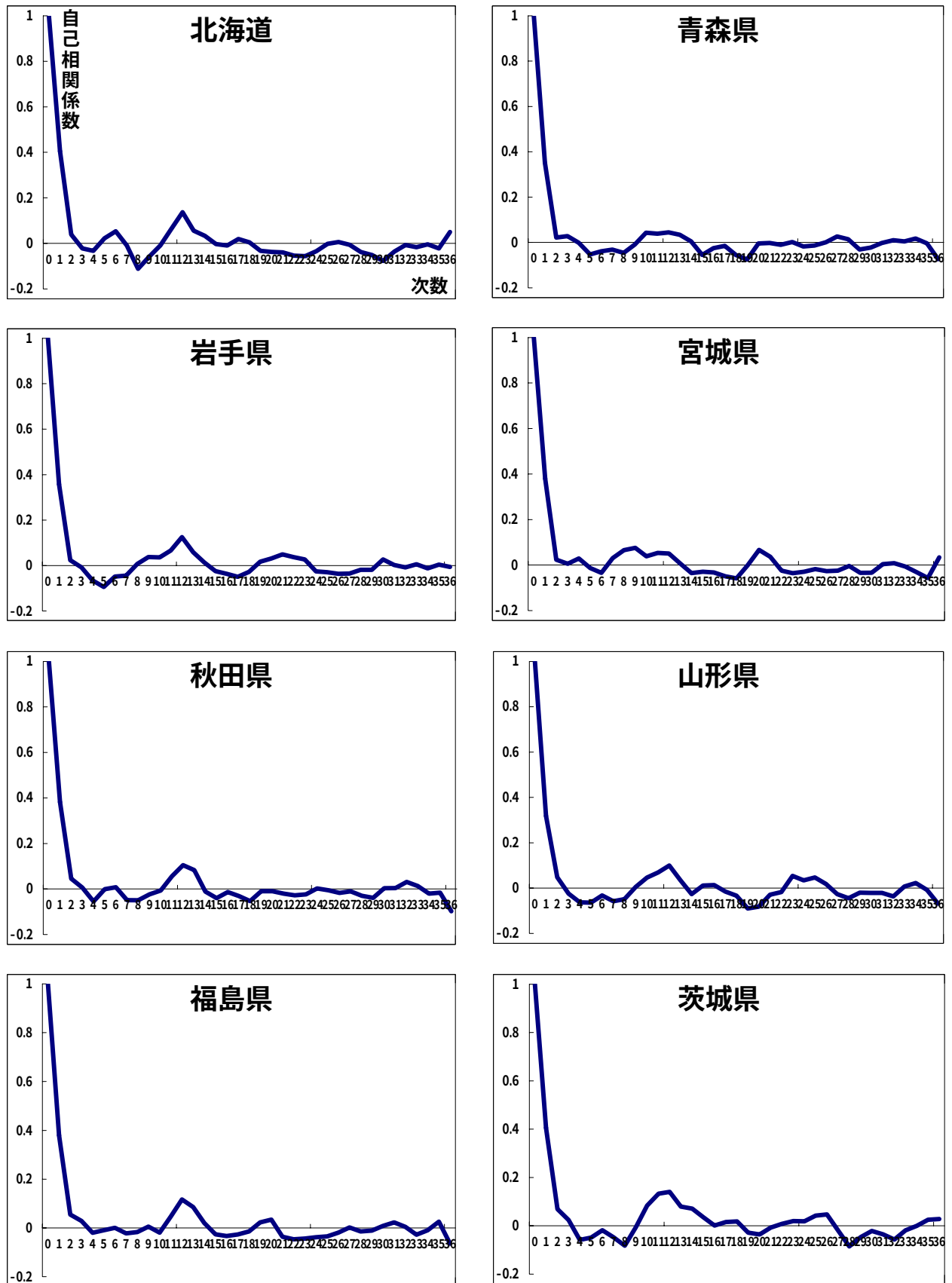
調査回数	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月	4月
調査回数1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
調査回数2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
調査回数3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
調査回数4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
調査回数5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
調査回数6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
調査回数7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
調査回数8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8

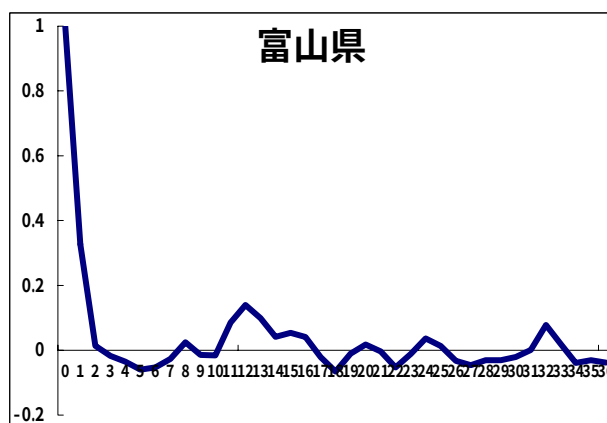
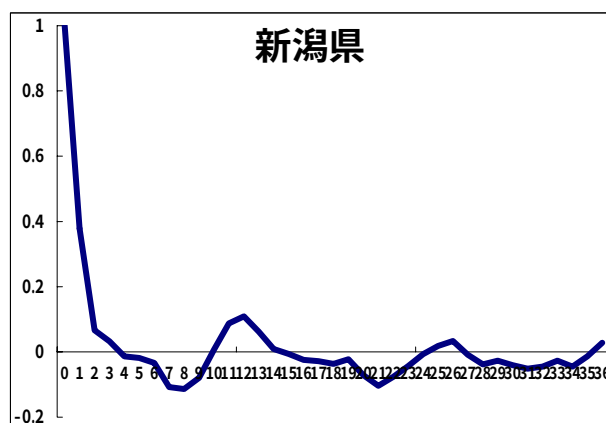
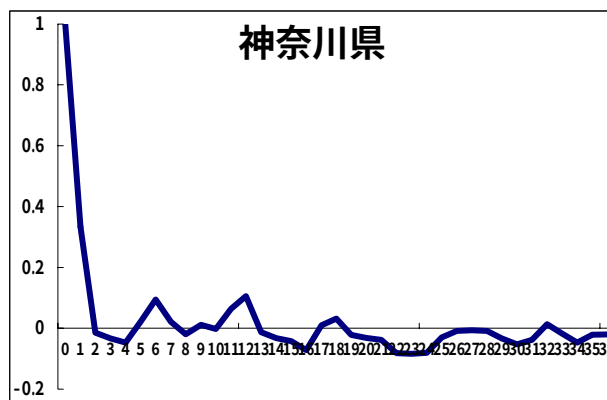
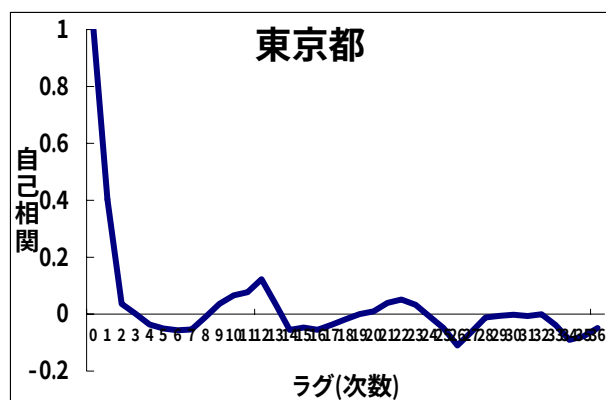
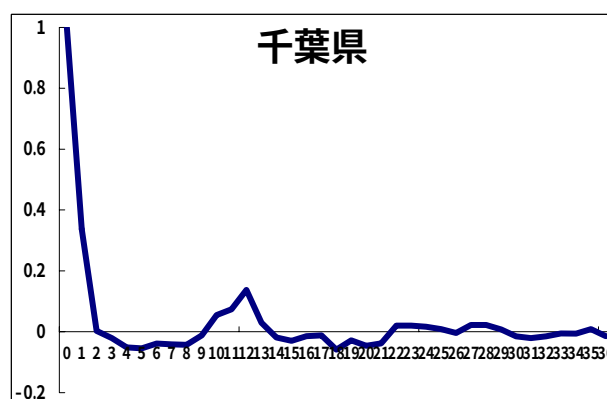
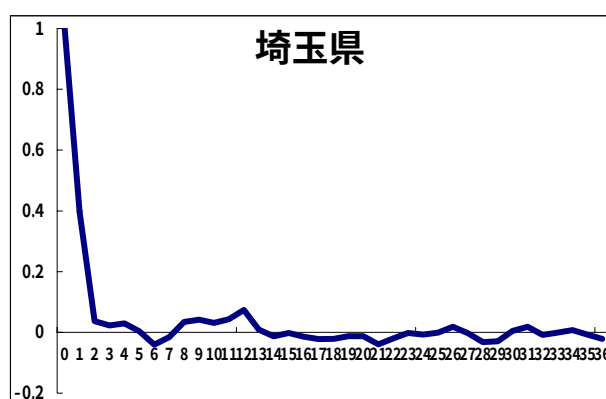
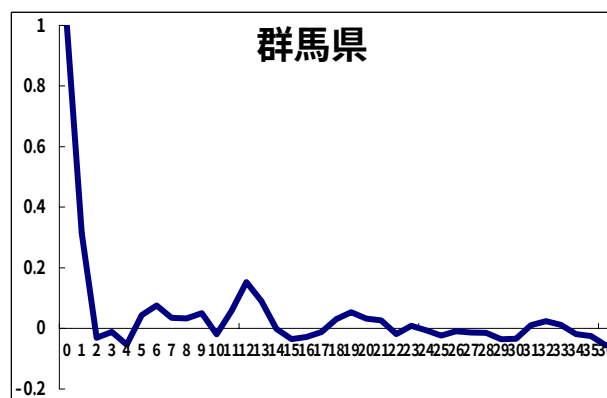
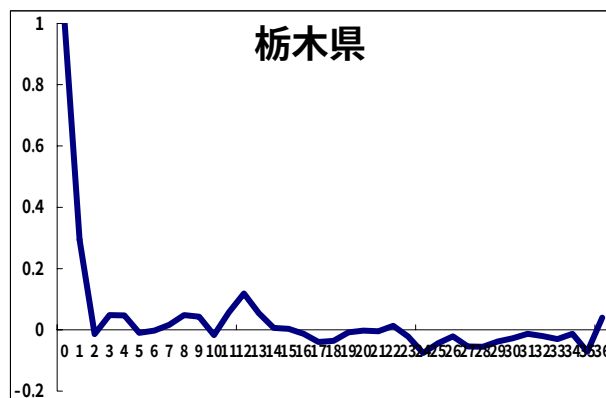
図 4-3 Pseudo Sampling Error の求め方③

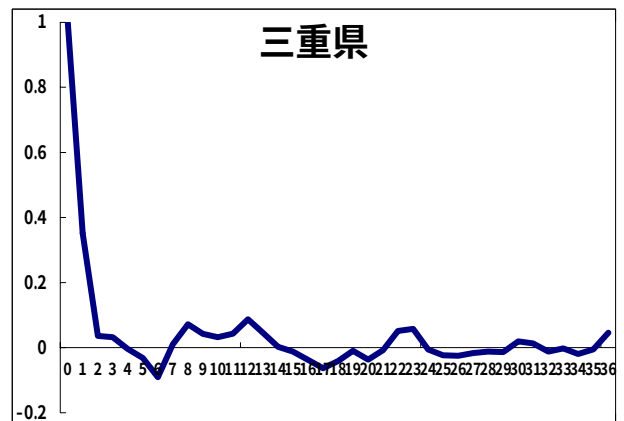
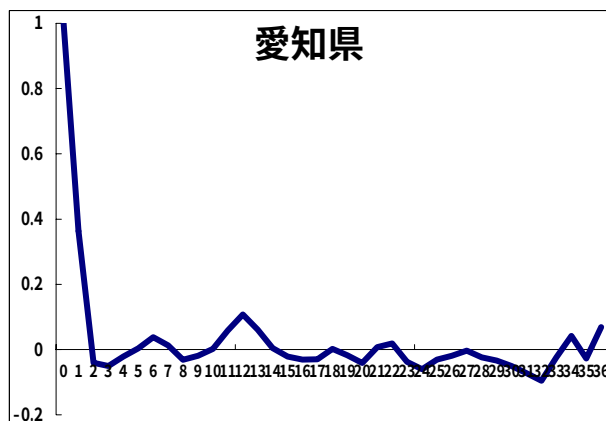
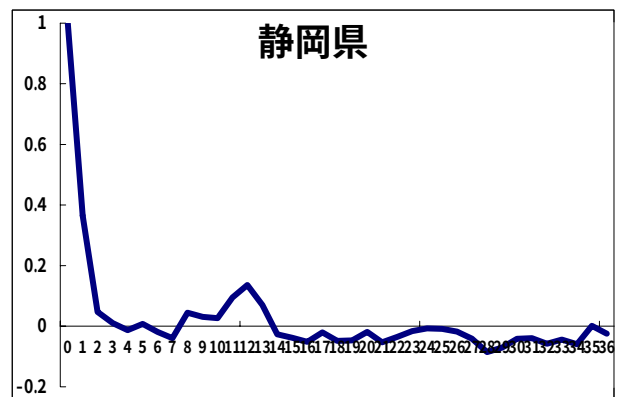
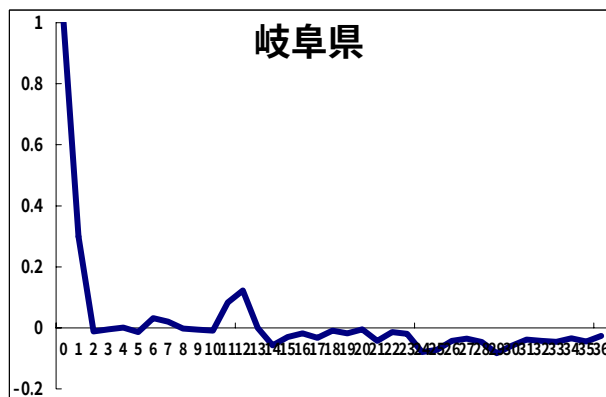
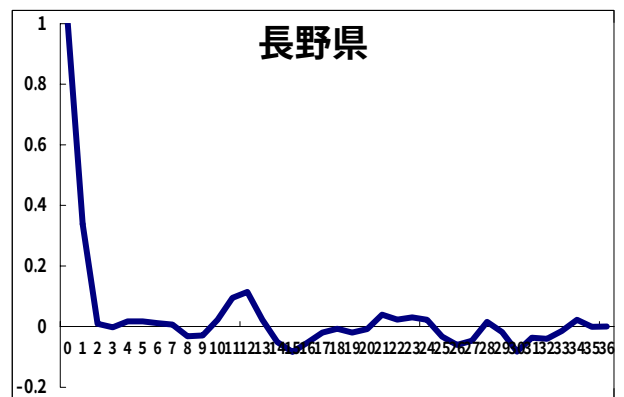
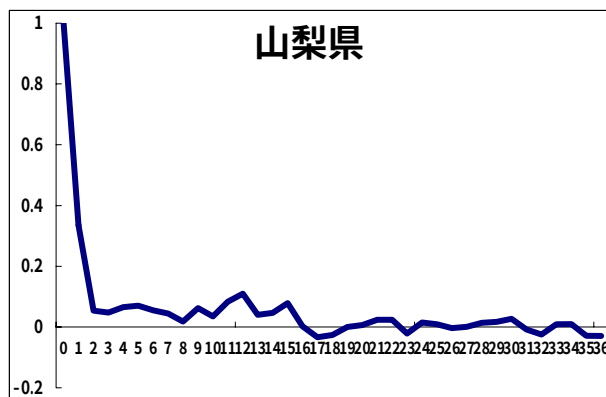
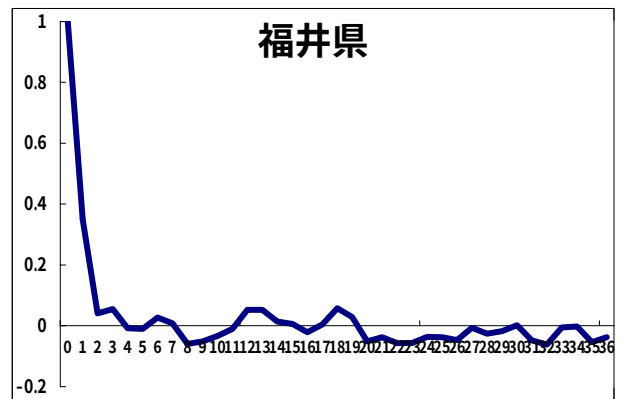
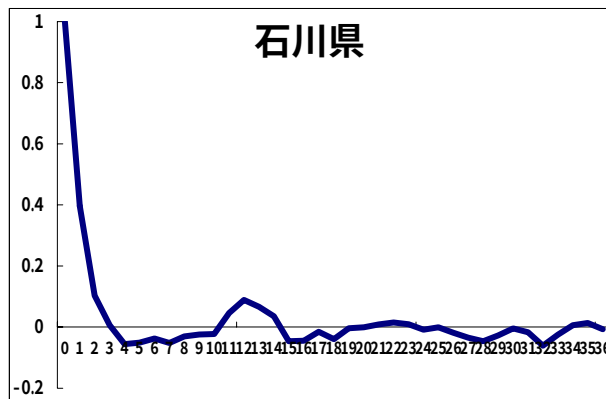
stream	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月	4月
stream1	1	2	3	4	1	2	3	4	5	6	7	8	5	6	7	8
stream5	5	6	7	8	5	6	7	8	1	2	3	4	1	2	3	4
stream8	8	1	2	3	4	1	2	3	4	1	2	3	4	5	6	7
stream4	4	5	6	7	8	5	6	7	8	5	6	7	8	1	2	3
stream7	7	8	1	2	3	4	1	2	3	4	1	2	3	4	7	8
stream3	3	4	5	6	7	8	5	6	7	8	5	6	7	8	1	2
stream6	6	7	8	1	2	3	4	1	2	3	4	1	2	3	4	5
stream2	2	3	4	5	6	7	8	5	6	7	8	5	6	7	8	1

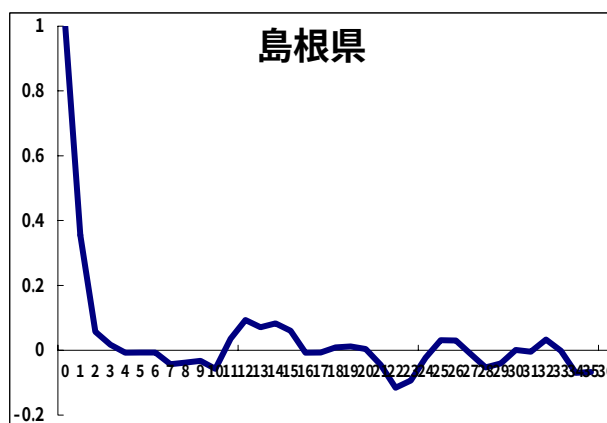
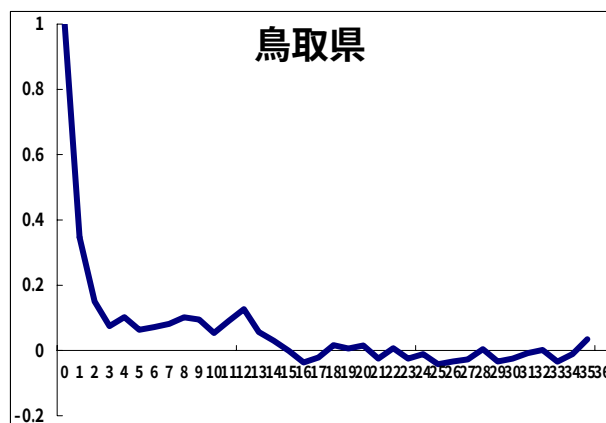
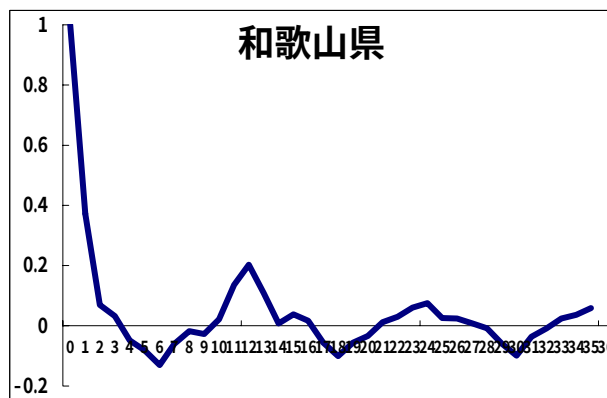
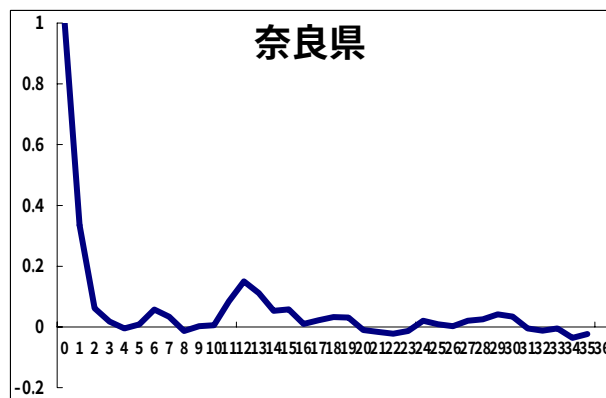
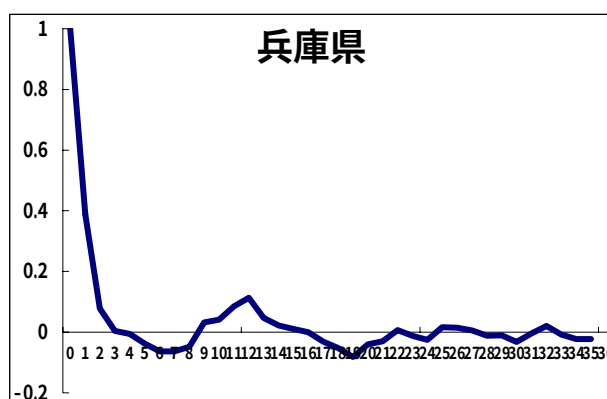
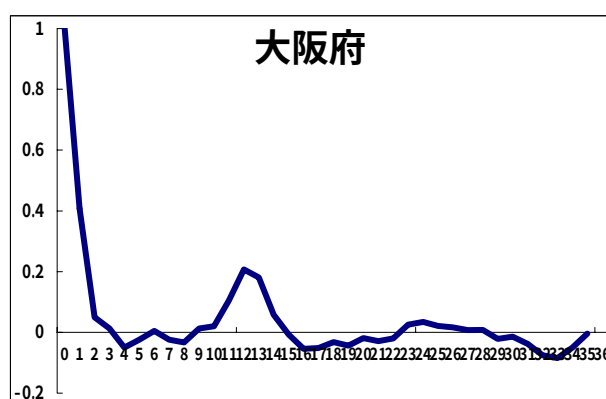
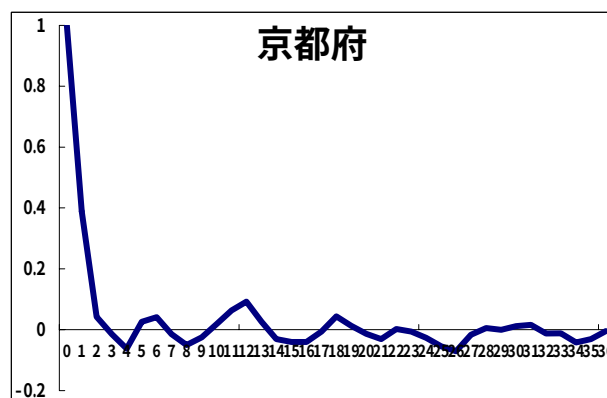
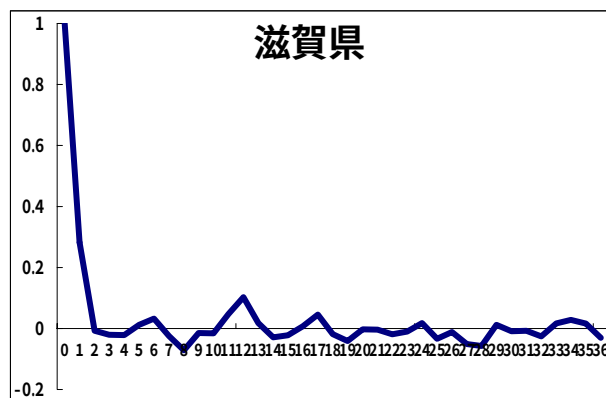
※ セル内の数値は調査回数を表す。

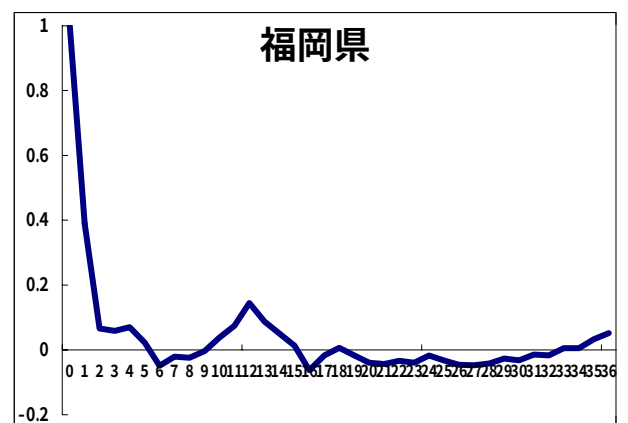
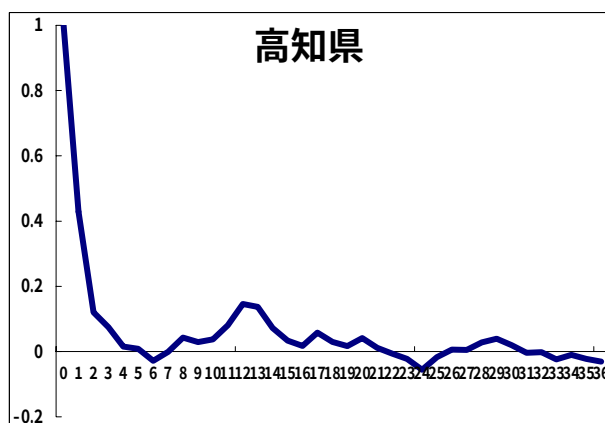
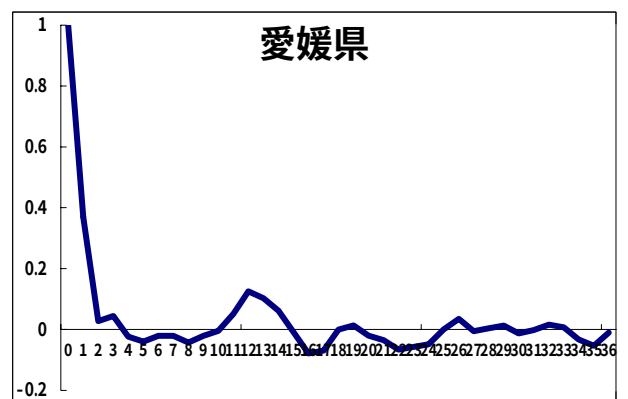
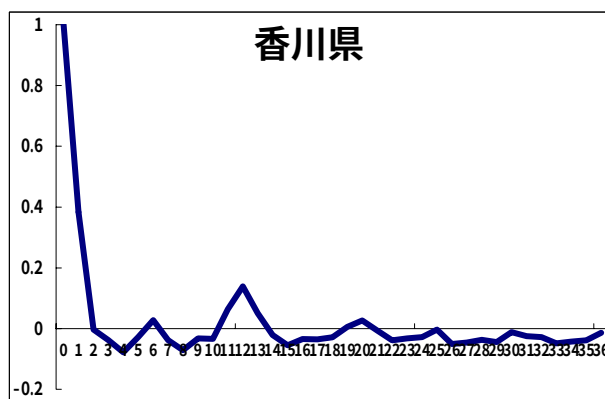
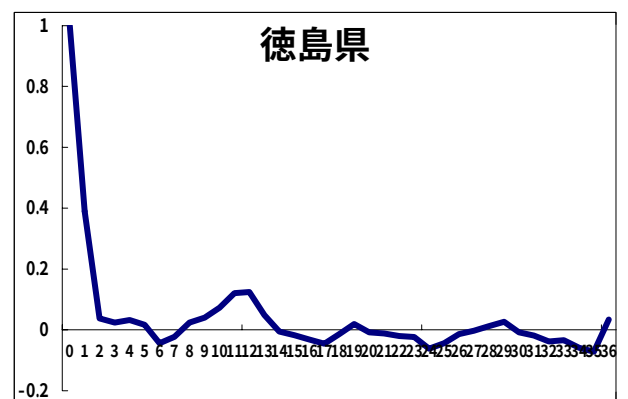
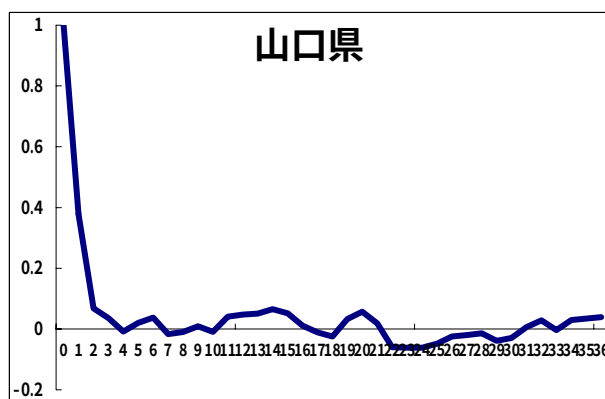
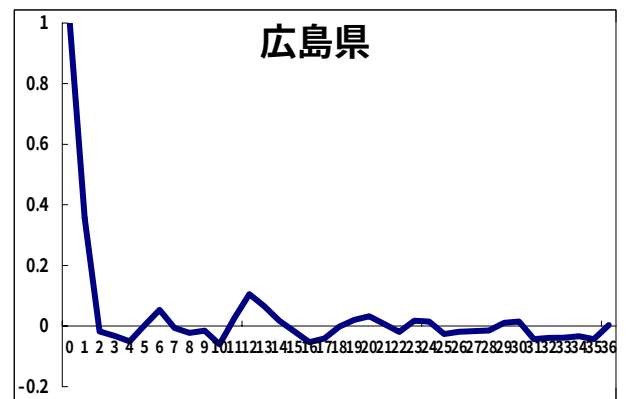
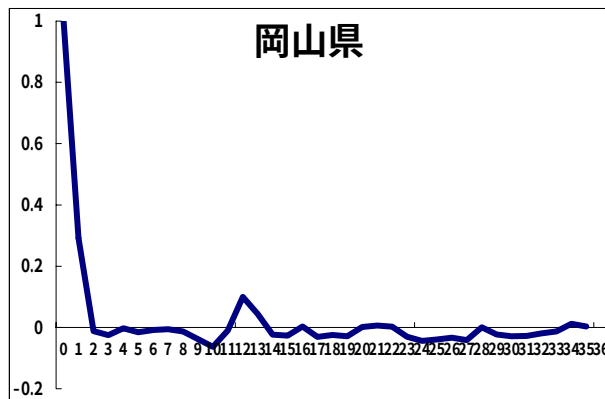
図 5 誤差の系列相関のグラフ











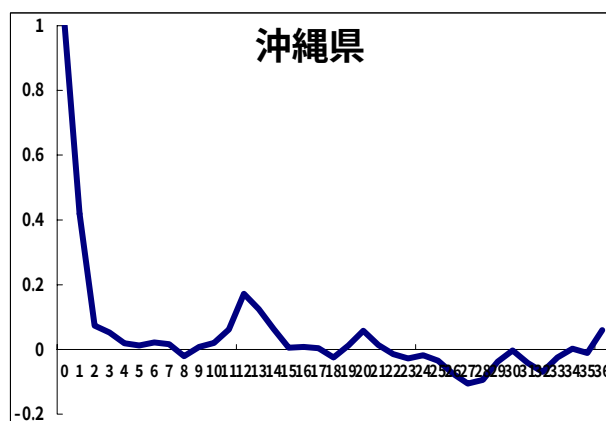
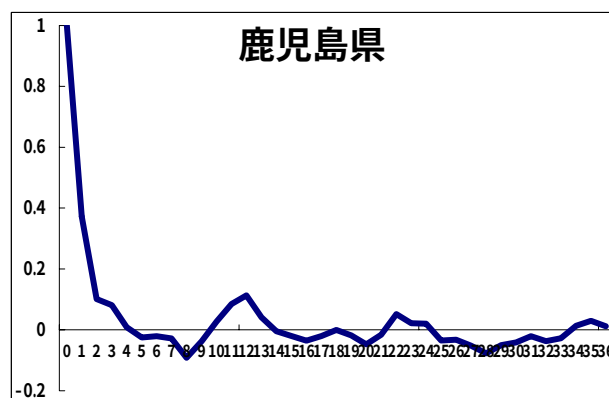
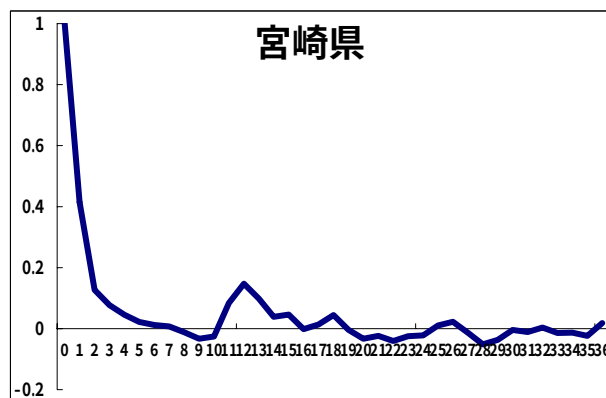
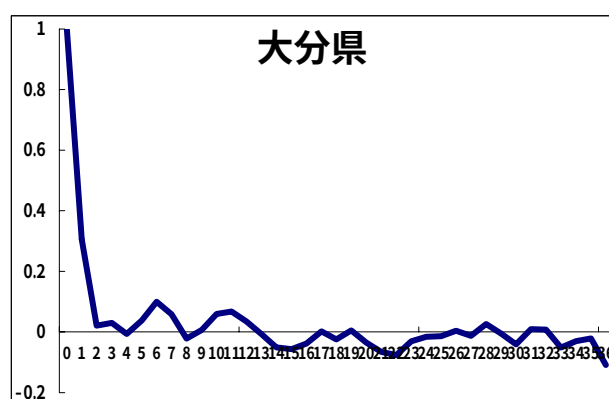
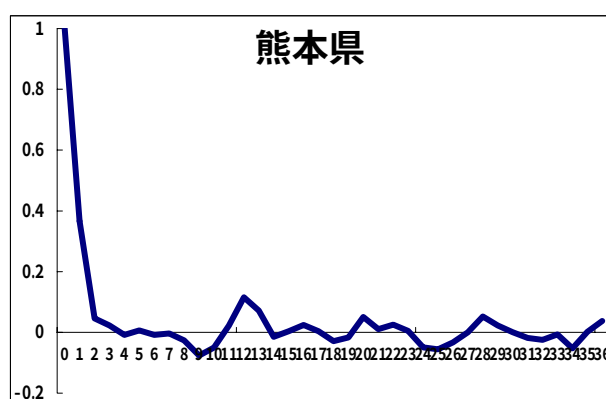
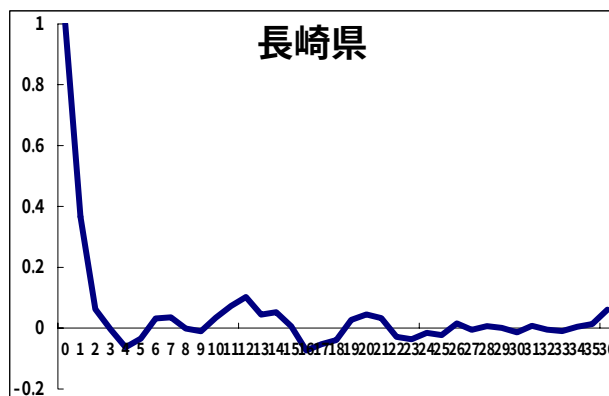
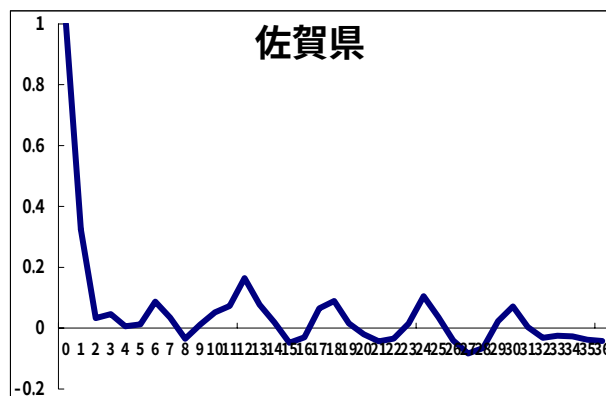


表 1 誤差の系列相関の表

		1次	2次	3次	4次	5次	6次	7次	8次	9次	10次	11次	12次	13次	14次
1	北海道	0.40	0.04	-0.02	-0.03	0.02	0.05	-0.01	-0.11	-0.06	-0.01	0.06	0.14	0.06	0.03
2	青森県	0.35	0.02	0.03	0.00	-0.05	-0.04	-0.03	-0.05	-0.01	0.04	0.04	0.04	0.03	0.00
3	岩手県	0.36	0.02	-0.01	-0.07	-0.09	-0.05	-0.04	0.01	0.04	0.04	0.07	0.13	0.06	0.01
4	宮城県	0.38	0.02	0.01	0.03	-0.01	-0.03	0.03	0.07	0.08	0.04	0.05	0.05	0.01	-0.03
5	秋田県	0.38	0.05	0.01	-0.05	0.00	0.01	-0.05	-0.05	-0.02	-0.01	0.06	0.10	0.08	-0.01
6	山形県	0.32	0.05	-0.02	-0.06	-0.06	-0.03	-0.06	-0.05	0.01	0.05	0.07	0.10	0.03	-0.03
7	福島県	0.38	0.05	0.03	-0.02	-0.01	0.00	-0.02	-0.02	0.01	-0.02	0.05	0.12	0.08	0.02
8	茨城県	0.41	0.07	0.02	-0.06	-0.05	-0.02	-0.05	-0.08	-0.01	0.08	0.13	0.14	0.08	0.07
9	栃木県	0.29	-0.01	0.05	0.05	-0.01	0.00	0.02	0.05	0.04	-0.02	0.05	0.12	0.06	0.01
10	群馬県	0.31	-0.03	-0.01	-0.05	0.04	0.08	0.03	0.03	0.05	-0.02	0.06	0.15	0.09	0.00
11	埼玉県	0.40	0.04	0.02	0.03	0.00	-0.04	-0.02	0.03	0.04	0.03	0.04	0.07	0.01	-0.01
12	千葉県	0.34	0.00	-0.02	-0.05	-0.05	-0.04	-0.04	-0.04	-0.01	0.05	0.07	0.14	0.03	-0.02
13	東京都	0.40	0.04	0.00	-0.04	-0.05	-0.06	-0.05	-0.01	0.04	0.07	0.08	0.12	0.04	-0.06
14	神奈川県	0.33	-0.01	-0.03	-0.05	0.02	0.09	0.02	-0.02	0.01	0.00	0.06	0.11	-0.01	-0.03
15	新潟県	0.38	0.07	0.03	-0.01	-0.02	-0.03	-0.11	-0.11	-0.08	0.01	0.09	0.11	0.06	0.01
16	富山県	0.32	0.01	-0.02	-0.04	-0.06	-0.05	-0.02	0.03	-0.02	-0.02	0.08	0.14	0.10	0.04
17	石川県	0.39	0.09	0.00	-0.06	-0.05	-0.04	-0.05	-0.04	-0.04	-0.03	0.04	0.09	0.07	0.03
18	福井県	0.35	0.04	0.06	-0.01	0.00	0.03	0.01	-0.06	-0.05	-0.03	-0.01	0.05	0.05	0.01
19	山梨県	0.32	0.03	0.03	0.06	0.07	0.04	0.04	0.02	0.06	0.01	0.05	0.09	0.03	0.04
20	長野県	0.34	0.01	0.00	0.02	0.02	0.01	0.00	-0.03	-0.03	0.03	0.10	0.11	0.02	-0.05
21	岐阜県	0.30	-0.01	-0.01	0.00	-0.02	0.03	0.02	0.00	-0.01	-0.01	0.08	0.12	0.00	-0.06
22	静岡県	0.36	0.05	0.01	-0.01	0.01	-0.02	-0.04	0.04	0.03	0.03	0.09	0.14	0.07	-0.03
23	愛知県	0.36	-0.04	-0.05	-0.02	0.00	0.04	0.01	-0.03	-0.02	0.00	0.06	0.11	0.06	0.00
24	三重県	0.35	0.04	0.03	0.00	-0.03	-0.09	0.01	0.07	0.04	0.03	0.04	0.09	0.05	0.00
25	滋賀県	0.28	0.00	-0.02	-0.02	0.01	0.03	-0.02	-0.06	-0.01	-0.01	0.04	0.09	0.01	-0.03
26	京都府	0.39	0.04	-0.01	-0.06	0.02	0.04	-0.01	-0.05	-0.02	0.02	0.06	0.09	0.03	-0.03
27	大阪府	0.41	0.05	0.01	-0.05	-0.02	0.00	-0.02	-0.03	0.01	0.02	0.10	0.21	0.18	0.06
28	兵庫県	0.39	0.08	0.00	0.00	-0.04	-0.06	-0.06	-0.05	0.03	0.04	0.09	0.11	0.05	0.02
29	奈良県	0.33	0.06	0.02	-0.01	0.00	0.06	0.03	-0.02	0.00	0.00	0.08	0.15	0.11	0.06
30	和歌山県	0.36	0.05	0.03	-0.05	-0.09	-0.14	-0.06	-0.02	-0.03	0.01	0.13	0.20	0.10	0.00
31	鳥取県	0.33	0.10	0.07	0.09	0.08	0.09	0.07	0.07	0.07	0.05	0.06	0.12	0.04	0.00
32	島根県	0.36	0.06	0.02	-0.01	-0.01	-0.01	-0.04	-0.03	-0.03	-0.06	0.03	0.08	0.06	0.07
33	岡山県	0.29	-0.01	-0.03	-0.01	-0.02	-0.01	-0.01	-0.01	-0.04	-0.06	-0.01	0.10	0.04	-0.02
34	広島県	0.36	-0.02	-0.03	-0.05	0.00	0.05	-0.01	-0.02	-0.02	-0.06	0.03	0.11	0.07	0.02
35	山口県	0.38	0.06	0.04	0.00	0.02	0.03	-0.02	-0.01	0.00	-0.01	0.03	0.05	0.05	0.06
36	徳島県	0.39	0.04	0.03	0.04	0.02	-0.05	-0.02	0.03	0.04	0.07	0.12	0.13	0.05	0.00
37	香川県	0.38	0.00	-0.04	-0.08	-0.03	0.02	-0.04	-0.07	-0.03	-0.03	0.06	0.14	0.05	-0.02
38	愛媛県	0.37	0.03	0.04	-0.02	-0.04	-0.02	-0.02	-0.04	-0.02	-0.01	0.05	0.13	0.10	0.06
39	高知県	0.43	0.12	0.07	0.01	0.00	-0.03	0.00	0.04	0.03	0.04	0.08	0.14	0.14	0.07
40	福岡県	0.39	0.07	0.06	0.07	0.02	-0.05	-0.02	-0.02	0.00	0.04	0.07	0.14	0.09	0.05
41	佐賀県	0.30	0.01	0.03	-0.03	-0.01	0.05	0.00	-0.05	-0.01	0.02	0.03	0.13	0.06	-0.01
42	長崎県	0.36	0.05	-0.01	-0.06	-0.04	0.03	0.04	0.01	-0.01	0.03	0.07	0.11	0.04	0.05
43	熊本県	0.36	0.04	0.02	-0.01	0.00	-0.01	0.00	-0.03	-0.08	-0.05	0.02	0.11	0.07	-0.01
44	大分県	0.30	0.02	0.03	0.00	0.04	0.10	0.06	-0.01	0.01	0.06	0.07	0.04	-0.01	-0.05
45	宮崎県	0.41	0.13	0.07	0.04	0.01	0.01	0.00	-0.02	-0.03	-0.03	0.08	0.14	0.09	0.04
46	鹿児島県	0.37	0.11	0.08	0.01	-0.02	-0.02	-0.03	-0.09	-0.04	0.02	0.08	0.11	0.04	0.00
47	沖縄県	0.42	0.07	0.05	0.02	0.01	0.02	0.02	-0.02	0.01	0.02	0.06	0.17	0.12	0.06

IV 推定に用いたプログラム

標本誤差の系列相関の計算に用いたプログラムを、以下に示す。

- ・プログラム①は、系列相関の計算用のデータをまとめるための SAS プログラムである。毎月のデータは、プログラム中のパス表示（'D:\data\H03\02.TXT'）のとおり、各年を表すフォルダの中に月ごとに保存してあるものとする。プログラムの途中 … となっている部分では、必要な時系列の分だけ入力を繰り返す。マクロ「dt_input」の引数の yr は、データレイアウトフォームが平成 14 年 1 月から変更されたことに対応したもので、平成 13 年以前のデータは yr = 13、平成 14 年以降のデータは yr = 14 とする。
- ・プログラム②は、プログラム①でまとめたデータに対して、PSE を用いた SEA の推定を行うプログラムである。入力するデータの月数が T の場合には、計算結果が T 行 47 列の行列の形で出力される。入力するデータは必ず 1 月から始まるようにする必要がある。データの終点は何月でも良い。

```
*****
* PROGRAM ①
* SOFTWARE: 「SAS」
* TITLE: RATIO ESTIMATION FOR EACH ROTATIN GROUP
* DATE: 2004/1/19
*****
;
%MACRO IN/ITAL;
    DATA DT00;
        DO I = 1 TO 47;
            DO J = 1 TO 8;
                PRF = I;
                RGN = J;
                OUTPUT;
            END;
        END;
    KEEP PRF RGN;
    RUN;
    DATA RESULT;
        SET _NULL_ ;
    RUN;
%MEND INITIAL;

%MACRO DT_INPUT (PATH, YR);
    DATA DT01;
        INFILE &PATH;
```

```

%IF &YR = 13 %THEN %DO;
    INPUT
        PRF    15 - 16
        RGN    24
        STT    43
        MLT    179 - 192;
    %END;
%IF &YR = 14 %THEN %DO;
    INPUT
        PRF    15 - 16
        STT    51
        RGN    105
        MLT    190 - 203;
    %END;
LF = 0;
UE = 0;
    IF 1 <= STT <= 5 THEN LF = MLT;
    IF STT = 5 THEN UE = MLT;
RUN;
%MEND DT_INPUT;

%MACRO RATIO;
    PROC SORT DATA = DT01;
        BY PRF RGN;
    RUN;
    PROC MEANS DATA = DT01 NOPRINT;
        BY PRF;
        VAR MLT;
        OUTPUT OUT = DT02 SUM = BCH;
    RUN;
    PROC MEANS DATA = DT01 NOPRINT;
        BY PRF RGN;
        VAR LF UE MLT;
        OUTPUT OUT = DT03 SUM = LF UE OV15;
    RUN;
    DATA DT04;
        MERGE DT02 DT03;
        BY PRF;
        LF = LF * BCH / OV15;
        UE = UE * BCH / OV15;
    RUN;
%MEND RATIO;

```

```
%MACRO IMPUTE;  
  PROC MEANS DATA = DT04 NOPRINT;  
    BY PRF;  
    VAR LF UE;  
    OUTPUT OUT = DT05 MEAN = MLF MUE;  
  RUN;  
  DATA DT05;  
    MERGE DT00 DT05;  
    BY PRF;  
  RUN;  
  DATA DT06;  
    MERGE DT04 DT05;  
    BY PRF RGN;  
    UR = 100 * UE / LF;  
    IF UR = 0 | UR = . THEN  
      UR = 100 * MUE / MLF;  
    KEEP UR;  
  RUN;  
%MEND IMPUTE;
```

```
%MACRO TRANS;  
  PROC TRANSPOSE DATA = DT06 OUT = DT06;  
  RUN;  
  DATA RESULT;  
    SET RESULT DT06;  
  RUN;  
%MEND TRANS;
```

```
%MACRO ESTIM(PATH, YR) ;  
  %DT_INPUT(&PATH, &YR) ;  
  %RATIO;  
  %IMPUTE;  
  %TRANS;  
%MEND ESTIM;
```

```
%INITIAL;  
%ESTIM(' D:¥DATA¥H03¥01.TXT' , 13) ;  
%ESTIM(' D:¥DATA¥H03¥02.TXT' , 13) ;  
  .  
  .  
  .  
%ESTIM(' D:¥DATA¥H15¥12.TXT' , 14) ;
```

```
DATA _NULL_;
  SET RESULT;
  FILE 'D:\UR.SEA.TXT';
  PUT COL1 - COL376;
RUN;
```

```
#####
## Program ②
## Software 「R」
## Title: Estimation of the Survey Error Autocorrelation
## Date: 2005/1/25
#####
library(ts)
```

```
U.R <- matrix(scan("D:/UR.SEA.TXT"), ncol = 376, byrow = TRUE)
len <- nrow(E)
```

```
I1 <- matrix(NA, 12, 8)
I1[9, ] <- I1[5, ] <- I1[1, ] <- c(1, 7, 5, 3, 2, 8, 6, 4)
I1[10, ] <- I1[6, ] <- I1[2, ] <- c(3, 1, 7, 5, 4, 2, 8, 6)
I1[11, ] <- I1[7, ] <- I1[3, ] <- c(5, 3, 1, 7, 6, 4, 2, 8)
I1[12, ] <- I1[8, ] <- I1[4, ] <- c(7, 5, 3, 1, 8, 6, 4, 2)
I2 <- matrix(NA, 24, 8)
I2[5:12, ] <- I1[5:12, ]
I2[1, ] <- c(1, 8, 6, 4, 2, 7, 5, 3)
I2[2, ] <- c(3, 1, 8, 6, 4, 2, 7, 5)
I2[3, ] <- c(5, 3, 1, 8, 6, 4, 2, 7)
I2[4, ] <- c(7, 5, 3, 1, 8, 6, 4, 2)
I2[13, ] <- c(2, 7, 5, 3, 1, 8, 6, 4)
I2[14, ] <- c(4, 2, 7, 5, 3, 1, 8, 6)
I2[15, ] <- c(6, 4, 2, 7, 5, 3, 1, 8)
I2[16, ] <- c(8, 6, 4, 2, 7, 5, 3, 1)
I2[21, ] <- I2[17, ] <- c(2, 8, 6, 4, 1, 7, 5, 3)
I2[22, ] <- I2[18, ] <- c(4, 2, 8, 6, 3, 1, 7, 5)
I2[23, ] <- I2[19, ] <- c(6, 4, 2, 8, 5, 3, 1, 7)
I2[24, ] <- I2[20, ] <- c(8, 6, 4, 2, 7, 5, 3, 1)
```

```
Split <- function(dt)
{
  res <- array(NA, c(len, 8, 47))
  for (i in 1:47) {
    j <- 8 * (i - 1) + 1
    res[, , i] <- dt[, j:(j + 7)]
  }
}
```

```
    }
    res
  }

Trans <- function(dt, num)
{
  res <- matrix(NA, len, 8)
  switch (num,
    J <- I1,
    J <- I2
  )
  for (i in 1:len) {
    k <- 12 * num
    mod <- i %% k
    if (mod == 0) {
      r <- k
    } else {
      r <- mod
    }
    L <- J[r, ]
    for (j in 1:8) {
      switch (num,
        res[i, j] <- dt[i, L[j]],
        res[i, L[j]] <- dt[i, j]
      )
    }
  }
  res
}

Center <- function(dt)
{
  work <- dt - apply(dt, 1, mean)
  res <- t(t(work) - apply(work, 2, mean))
  res / sqrt(apply(res, 1, var))
}

Auto.corr <- function(dt)
{
  work <- matrix(NA, 36, 8)
  for (i in 1:8) {
    acf.res <- acf(dt[, i], lag.max = 35,
```

```
        type = "covariance", plot = F)
    work[, i] <- acf.res$acf
  }
  res <- apply(work, 1, sum)
  res / res[1]
}

SEA <- function(dt)
{
  res <- matrix(NA, 36, 47)
  dt <- Split(dt)
  for (i in 1:47) {
    work <- Trans(dt[, , i], 1)
    work <- Center(work)
    work <- Trans(work, 2)
    res[, i] <- Auto.corr(work)
  }
  res
}

result <- SEA(U.R)
```

おわりに

今回は労働力調査に対して、いくつかの小地域推定の手法を適用し、都道府県別月次の完全失業率を推定した。そして手法ごとに結果の比較及び分析を行った。各手法による推定結果をもう一度まとめると、以下のようになっている。

- ・アメリカ BLS における小地域推定

モデルによる推定値を見ると、観測値の誤差が上手く取り除かれている。また誤差率も観測値に比べて低くなっており、安定している。ベンチマーキングによって偏りも上手く修正されている。説明変数と原型列の季節性のずれも、季節項により修正されている。

- ・EBLUP(Empirical Best Linear Unbiased Prediction) 推定量を用いた小地域推定

誤差率や ARE を比較すると、モデルによる推定値の方が観測値のものよりも低くなっており、精度が改善されているといえる。EBLUP 推定量の加重平均のウェイトが変化することで、モデルと観測値を上手く組み合わせた推定を行うことができている。

- ・Time Series & Cross Sectional Model を用いた小地域推定

誤差率や ARE を比較すると、モデルによる推定値の方が観測値のものよりも低くなっており、精度が改善されているといえる。マルコフ連鎖モンテカルロ法を用いることにより、複雑なモデルの推定も簡単に行うことができている。

どの手法についても、モデルや補助情報を上手く選択することで、もとの観測値データよりも良好な結果が得られている。しかし補助情報と完全失業率の相関が年々弱まっていることや、外れ値の問題など、今後検討すべき課題がいくつかあることもわかった。

また本稿では分析を行わなかったものの、手法間の比較に関する検討も必要であると考えられる。手法間の比較は、その基準となる絶対的な指標が存在しないので、実際かなり難しい問題である。しかし信頼できるベンチマークと比較する方法やモンテカルロシミュレーションによる分析などを組み合わせた総合的な判断を、今後行っていく必要があると思われる。

本稿で紹介した様々な小地域推定の手法に更なる改良を加えて、他の標本調査へ応用し、小地域の結果に関して有益な情報が得られるようにしたい。

参考文献

- [1] P.A.Bell. and A.M.Carolan, Australian .Bureau of Statistics Working Paper, 98/1, 1998.
- [2] W.G.Cochran, Sampling Techniques 3rd, Wiley New York, 1977
- [3] G.S.Datta, P.Lahiri, T.Maiti, and K.L.Lu, Hierarchical Bayes estimation of unemployment rate for the states of the U.S. Journal of the American Statistical Association. 94, 1999.
- [4] F.T.Denton, Adjustment of Monthly Quarterly Series to Annual Totals:An Approach Based on Quadratic Minimization, The Journal The American Statistical Association, 66, 1971.
- [5] J.Durbin and S.J.Koopman, Time Series Analysis by State Space Model, Oxford University Press, 2001.
- [6] R.E.Fay. and R.A.Herriot, Estimation of Income from Small Places: An application of Tames-Stein Procedures to Census Data, Journal of The American Statistical Association, 74, 1979.
- [7] M.Ghose and J.N.K.Rao., Small Area Estimation: An Appraisal, Statistical Science, 9, 1994.

- [8] G.Kitagawa, and W.Gersch., A smoothness priors state space approach to the modeling of time series with trend and seasonality, The Journal The American Statistical Association, 79, 1994.
- [9] D.V.Lindrey, A.F.M.Smith, Bayes Estimates for The Linear Model, 1972.
- [10] D. Pfeffermann, M.Feder and D.Signorelli, Estimation of Autocorrelations of Survey Errors With Application to Trend Estimation in Small Areas, Journal of Business & Economic Statistics , 16, 1998.
- [11] D. Pfeffermann, and R Tiller, ,State-Space Modeling with Correlated Measurements with Application to Small Area Estimation Under Benchmark Constraints, Southampton Statistical Sciences Research Institute S3RI Methodology Working Paper , 2003
- [12] J.N.K.Rao., Some Recent Advances in Model Based Small Area Estimation, Survey Methodology, 25, 1999.
- [13] J.N.K.Rao , Small Area Estimation, Wiley, 2003.
- [14] R.Tiller, A Kalman Filter Approach to Labor Force Estimation Using Survey Data, Proceeding of Survey Research Methods Section, American Statistical Association, 1989.
- [15] R.Tiller, An Application of Time Series Methods to Labor Force Estimation Using CPS Data, Proceeding of Survey Research Methods Section, Journal of Official Statistics. 8, 1989.
- [16] R.Tiller, Time Series Modeling of Sample Survey Data from the U.S. Current Population Survey Journal of Official Statistics, 6 , 1992
- [17] U.S.Bureau of Census, Local Area Unemployment Statistics Program Manual, 1996.
- [18] U.S.Bureau of Census, Current Population Survey : Design and Methodology, Technical Paper 63RV, 2002.
- [19] U.S.Bureau of Census, Proposed Improvement in Estimating and Benchmarking State Labor Force Estimates, 2004
- [20] K.M.Wolter, Introduction to Variance Estimation, Springer-Verlag, 1985.
- [21] Y.You, J.N.K.Rao, and J.Gambino, Model-Based Unemployment Rate Estimation for the Canadian Labour Force Survey: A Hierarchical Bayes Approach, Survey Methodology, 29, 2003.
- [22] T.Zimmerman, R.Tiller and T.Evans, State Unemployment Rate Time Series Models, Proceeding of Survey Research Methods Section, American Statistical Association, 1994.
- [23] 青木正直, 時系列解析と日本経済, 東洋経済新報社, 1984.
- [24] 青木正直, 日本経済のシステム論的接近, 朝倉書店, 1984.
- [25] 石井 達男, 労働力調査の標本誤差の一試算, 統計局研究彙報第 46 号, 総務庁統計局, 1988
- [26] 石黒真木夫 松本隆 乾敏郎 田邊國士, 統計科学のフロンティア第 10 巻 一階層ベイズモデルとその周辺一, 岩波書店, 2004.
- [27] 岩永琢磨, 労働力調査の都道府県別結果の試算について, 統計局研究彙報第 49 号, 総務庁統計局, 1991
- [28] 岩永琢磨, 古橋正宏, 労働力調査の標本誤差, 統計局研究彙報第 49 号, 総務庁統計局, 1991
- [29] 大駒誠一, 改定 FORTRAN77, サイエンス社, 1982.
- [30] 岡田昌 編, The R Book, 九天社, 2004.
- [31] 尾崎統 北川源四郎 編, 時系列解析の方法, 朝倉書店, 1998
- [32] 小泉英希, いくつかの Stein タイプの推定量の導入と評価方法, 統計研究彙報第 61 号, 総務省統計局, 2005

- [33] 坂元慶行 石黒真木夫 北川源四郎, 情報量統計学, 共立出版, 1983
- [34] 柏木宜久, 状態空間接近による季節変動中性に関する一考察, 統計数理, 45 巻 2 号, 統計数理研究所, 1997.
- [35] 片山徹, 応用カルマンフィルタ, 朝倉書店, 1989
- [36] 北川源四郎, 時系列の分解 —プログラム DECOMP の紹介—, 統計数理, 34 巻 2 号, 統計数理研究所, 1986.
- [37] 北川源四郎, 時系列解析プログラミング, 岩波書店, 1993.
- [38] 北川源四郎, 季節調整プログラム DECOMP とその後の展開, 統計数理, 45 巻 2 号, 統計数理研究所, 1997.
- [39] 北川源四郎 小西貞則, 情報量基準, 岩波書店, 2004.
- [40] 北川源四郎, 時系列解析入門, 岩波書店, 2005.
- [41] SAS User's Guide Basics 1~3, SAS ソフトウェア, 1986.
- [42] 佐和隆光, 回帰分析, 朝倉書店, 1979.
- [43] 繁枅数男, ベイズ統計入門, 東京大学出版会, 1985.
- [44] 篠崎信夫, Stein タイプの縮小推定量とその応用, 応用統計学, 20, 1991.
- [45] 下平英寿 伊藤秀一 久保川達也 竹内啓, 統計科学のフロンティア第 3 巻 —モデル選択—, 岩波書店, 2004.
- [46] 鈴木武 山田朔太郎, 数理統計学, 内田老鶴圃, 1996.
- [47] 杉原左右吉, 時系列の統計的研究, 東洋経済新報社, 1984.
- [48] 杉原敏夫, 経営・経済のための時系列分析と予測 —カルマンフィルタ適用を中心として— 税務経理協会, 1994.
- [49] 鈴木雪夫, 統計学, 朝倉書店, 1987.
- [50] 総務省統計局, 労働力調査 標本設計の解説, 2005.
- [51] 総務省統計局, 労働力調査の解説, 2002.
- [52] 竹内啓 監修 市川伸一 大橋靖雄 岸本淳司 浜田知久馬, SAS によるデータ解析入門(第 2 版), 東京大学出版会, 1989.
- [53] 戸川隼人, ザ FORTRAN77, サイエンス社, 1986.
- [54] 中妻照雄, ファイナンスのための MCMC 法によるベイズ分析, 財団法人三菱経済研究所, 2003.
- [55] 中村健策 古田裕繁, 賃金センサス・毎勤統計活用の手引き, 労働法令協会, 2004.
- [56] 中山和彦, SAS によるデータ解析 基礎編 上下, 丸善, 1990.
- [57] 標本誤差推計研究会 編, 標本誤差の推計方法 - 最新時代の理論と実証 -, 財団法人統計情報研究開発センター, 1999.
- [58] 広松毅 浪花貞夫, 経済時系列分析, 朝倉書店, 1994.
- [59] 船尾暢男, The R Tips, 九天社, 2005.
- [60] 宮脇典彦 酒井和男 共著, SAS によるデータ解析の基礎(Windows 版 SAS 準拠) 培風館, 1997
- [61] 山本拓, 経済の時系列分析, 創文社, 1995.
- [62] A.C.ハーベイ 著 国友直人 山本拓 訳, 時系列モデル入門, 東京大学出版会, 1988.
- [63] W.N.ヴェナブルズ, B.D.リプリー 著 伊藤 幹夫 大津 泰介 戸瀬 信之 中東 雅樹 訳, S による統計解析, シュプリンガー・フェアラーク東京, 2001.