小地域推計と労働力調査への適用

元山　斉
山口幸三

はじめに

小地域推計は、国における地域の経済状況の把握や地域振興などの施策の策定において、近年需要が高まっている。しかしながら、国勢調査などのセンサスは別として、標本調査では小地域別の結果はそれほど多くない。これは、全国又は地域別の結果を表章できる標本規模で設計されている標本調査が多く、これより細かい小地域の結果を推定するには、標本が小さくなるため、結果精度を確保するのが難しくなるためである。

そのため、標本規模を拡大せずに、結果精度を高める統計的な手法の研究が現在盛んになされている。アメリカでは以前から州ごとの労働市場の不均衡を把握するために、小地域推計に取り組んでいる。イギリスでは地域レベルで犯罪、失業、教育等の問題を解消するために小地域推計の重要性が高まり、研究を進めている。国際会議の場でも小地域推計の議論が行われており、国際的な関心も高くなっている。

Ⅰ　小地域推計について

統計局においても、小地域推計について注目し、平成 15 年度から 17 年度に諸外国の適用事例や参考文献を基に小地域推計の手法について調べるとともに、山口と元山を（メンバーに）含む雇用統計地域推計研究会を設置して検討してきたところである。種々の統計的手法についての検討は研究会で行い、検討した手法を実際に適用する際の理論面及び技術面を主に元山が担当し、試算等の実務面を統計局（山口、高部他）が担当した。

まず、統計局において検討した種々の統計的手法を簡単に解説した後、実際に適用した事例について紹介する。

１　時系列回帰モデル

諸外国で実際に統計調査に適用している手法についてみてみる。最初はアメリカの労働統計局で採用している時系列回帰モデルである。これは回帰項、トレンド項、季節変動項、不規則変動項、標本誤差項から構成されるモデルを仮定している。このモデルからカルマン・フィルタを用いて、標本誤差項を推計し、観測値から標本誤差項を除去し、推計値を求める。実際にわが国の労働力調査に適用した事例は、この手法を基本に改良したモデルであり、その手法の詳細は「Ⅱ」の「２　推計方法」を参照されたい。

この手法は、地域ごとにモデルを設定するので、それぞれの地域の特性を反映させることができ、時系列データの推計には有効な手法といえる。

２　ロジスティック回帰モデル

次に、イギリスの中央統計局の手法は、失業率などの比率のデータを推計するため、ロジスティック変換を用いるロジスティック回帰モデルで、アメリカが地域ごとの時系列情報を用いるのに対して、空間（地域）情報から推計している。

地域別に、観測値を被説明変数とし、説明変数が１つの場合のロジスティック回帰モデルは次の式で表される。

$$\mathrm{logit}\left(\pi_i\right) = \beta_0 + \beta_1 x_i + \beta_2 D_A$$

$$\mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

π_i ：i 地域における観測値
x_i ：i 地域における説明変数
D_A ：ダミー変数

ロジスティック回帰モデルの式は分散不均一であるので、i 地域においてウエイト w_i を用いた加重回帰を行う。

この手法は、細かい地域の補助情報が得られるときに、不完全でもよいからその小地域の推計が必要である場合に用いられ、精度の高い推計は困難である。イギリスではより細かい Unit Level のデータを用いており、男女、地域別に事後層別を行い、それを足しあげて推計値を求めている。反対に説明変数のデータが細かい地域情報が得られない場合は、回帰の当てはまりが悪くなり、あまり有効ではない。

なお、イギリスでは後述する EBLUP タイプの推定量についても検討していたようであるが、算式が複雑であること、効果に疑問があることから採用せずに、単純なロジスティック回帰モデルを採用しているようである。

3 EBLUP (Empirical Best Linear Unbiased Prediction)

イギリスが検討していたという EBLUP 推定量についてみると、データの背後にある構造として回帰直線を考え、観測値が次の Fay-Herriot モデルに従うと仮定している。このモデルでは、$x_i\beta+v_i$ が回帰直線上に乗るのではなく、地域の効果 v_i だけ離れていると考えている。

$$y_i = x_i\beta + v_i + e_i$$
$$e_i \sim N(0, \psi_i) \quad v_i \sim N(0, \sigma_v^2)$$

i ：都道府県
y_i ：観測値
x_i ： 説明変数（補助情報）
v_i ： 地域ごとの効果
e_i ： 標本誤差

ここで標本誤差 e_i の分散 ψ_i は既知と仮定する。実際は、推定量 $\hat{\psi}_i$ を用い、あらかじめ推計しておく。地域の効果 v_i の分散 σ_v^2 は、各地域で等しいと仮定する。

σ_v^2 を既知と仮定したとき、θ_i の線形不偏な予測量の中で平均二乗誤差が最小となるものは、BLUP と呼ばれ、BLUP は、以下の式で求めることができる。

$$\theta_i^{BLUP} = \gamma_i y_i + (1-\gamma_i) x_i \tilde{\beta}$$

$$\gamma_i = \frac{\sigma_v^2}{\psi_i + \sigma_v^2}$$

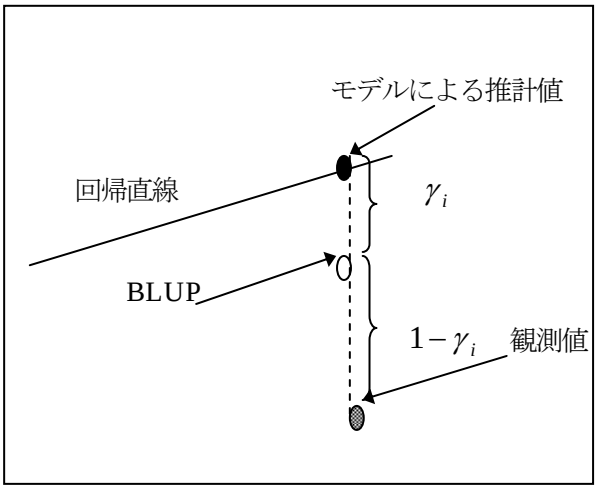
$$\tilde{\beta} = \left[\sum_{i=1}^{47} \frac{x_i^2}{\psi_i + \sigma_v^2}\right]^{-1} \left[\sum_{i=1}^{47} \frac{x_i y_i}{\psi_i + \sigma_v^2}\right]$$

BLUP の式を見るとウエイトを γ_i として、観測値 y_i と回帰モデルによる推計値 $x_i\tilde{\beta}$ とを加重平均した形になっており、BLUP 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。つまり、真の値は回帰直線と観測値の間にあると考えていることになる(図参照)。

観測値と回帰モデルそれぞれの変動の大きさにより観測値(あるいはモデルによる推計値)にかかるウエイトが決まってくる。観測値 $\hat{\theta}_i$ の誤差 ψ_i が大きくなると、モデルのウエイトが大きくなり、

地域の変動σ_v^2が大きくなると観測値のウエイトが大きくなる。観測値の誤差に応じて適当な割合でモデルの情報を取り込むという柔軟な推計値となっている。

図　BLUP 推定量の意味



実際の推計では、σ_v^2の値が未知なので推計する必要がある。BLUPにおいてσ_v^2の値を推計値$\tilde{\sigma}_v^2$で置き換えたものは、EBLUPと呼ばれる。$\tilde{\sigma}_v^2$は最尤法により推計する。

この手法は、ロジスティック回帰モデルと同じで回帰のあてはまりが悪いと有効でないものの、大規模標本調査について、センサス等の補助情報で推計する場合などには、最適な手法と思われる。また、EBLUPは、経験ベイズ推定量としての解釈も可能である。

4　時系列・クロスセクションモデル（Time Series and Cross Section Models）

カナダ統計局で研究されている手法で、推計には Fay-Herriot モデルを拡張した Rao-Yu モデルと呼ばれるモデルを用いる。これは回帰モデルに、①地域ごとの特色を反映する確率変数 v_i、②AR(1)モデルに従い、ランダムな時系列の効果を表す確率変数 u_{it}、の両方が加わったモデルである。モデルの具体的な形を以下に示す。

$$y_{it} = \theta_{it} + e_{it} \qquad y_i \sim N(\theta_i, \Sigma_i)$$

$$\theta_{it} = x_{it}\beta + v_i + u_{it} \quad v_i \sim N(0, \sigma_v^2)$$

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it} \ \varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$$

$|\rho|<1$

$y_i = (y_{i1}, \cdots, y_{iT})'$　　$\theta_i = (\theta_{i1}, \cdots, \theta_{iT})'$

$e_i = (e_{i1}, \cdots, e_{iT})'$

i：地域、t：時点　$1 \leqq t \leqq T$

y_{it}：観測値、x_{it}：説明変数（補助情報）、

e_{it}：標本誤差、Σ_i：誤差の分散共分散行列

上記のモデルに、階層的ベイズの考えを適用する。分布に以下の仮定を置く。

$$y_i \mid \theta_i \sim N_T(\theta_i, \Sigma_i) \qquad \theta_i = (\theta_{i1}, \cdots, \theta_{iT})'$$

$$\theta_i \mid \beta, u_{it}, \sigma_v^2 \sim N(x_i\beta + u_{it}, \sigma_v^2)$$

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

$$\beta \propto 1$$

$\sigma_v^2 \sim IG(a_1, b_1)$　　N_T：T 変量正規分布

$\sigma_\varepsilon^2 \sim IG(a_2, b_2)$　　IG：逆ガンマ分布

上記のモデルから事後分布をつくり、パラメータの推定はマルコフ連鎖モンテカルロ法の一種である Gibbs Sampler を用いる。その際に、L 個の連鎖の平均を取るという手法を用いる。これにより、

初期値に依存しない、安定した推定を行うことができる。誤差の分散共分散行列Σ_iは、誤差の自己相関を元に推定する。θ_{it}の推定には、ラオ・ブラックウェル化推定量を用いる。これにより、初期値の影響を少なくして、シミュレーションの誤差を抑えることができると期待される。

この手法は、シミュレーションを用いて推計値を求めるので､我が国の政府統計で採用するのは、難しいと考えられる。

5　階層的ベイズモデル

カナダの手法は階層的ベイズの考え方を適用しているが、階層的ベイズモデルで推計する手法も考えられる。

例えば、全国、地域ブロック、都道府県の各段階において、階層的な構造になっているモデルを考える。推計には、以下の階層的ベイズモデルを用いる。やや強い仮定ではあるものの、全ての確率変数は独立であるとする。

$$y_{ij} \mid \theta_{ij} \sim N(\theta_{ij}, \psi_{ij})$$
$$\theta_{ij} \mid \alpha_i \sim N(\alpha_i, \sigma_i^2)$$
$$\alpha_i \mid \beta \sim N(\beta, \tau^2)$$

$1 \le i \le 10$: 10 地域

$1 \le j \le 47$: 都道府県

y_{ij} : 第 i 地域内の第 j 県の観測値

ψ_{ij} : 第 i 地域内の第 j 県の観測値の分散

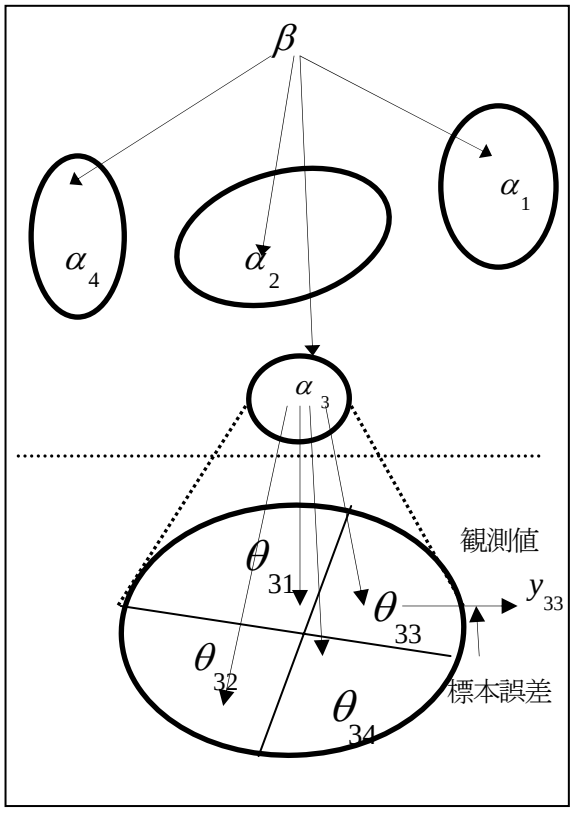
θ_{ij} は都道府県別のパラメータであり、これを推計する。モデルのイメージについては、以下の図を参照のこと。

上記のモデルでは次のような、3 段階の発生過程を考えていることになる。

第１段階 : 全国の平均的な値 β から、各地域の平均的な値 α_i が、分布 $N(\beta, \tau^2)$ に従って発生する。

第２段階 : 第１段階で発生した α_i から、地域内各県の平均値 θ_{ij} が、分布 $N(\alpha_i, \sigma_i^2)$ に従って発生する。

図　階層的ベイズモデルのイメージ



第３段階 : 第２段階で発生した各県の平均値 θ_{ij} に標本誤差が付加されて､最終的な観測値は分布 $N(\theta_{ij}, \psi_{ij})$ に従う確率変数　y_{ij} として観測される。

β、σ_i^2、τ^2、ψ_{ij} を既知の値であると仮定し、事後分布を導出する。求めた事後分布から、θ_{ij} の事後平均及び事後分散が、以下のように求められる。

$$E[\theta_{ij} \mid y_{ij}] = \left(\frac{1}{\psi_{ij}} + \frac{1}{\sigma_i^2 + \tau^2}\right)^{-1} \left(\frac{y_{ij}}{\psi_{ij}} + \frac{\beta}{\sigma_i^2 + \tau^2}\right)$$

$$V\left[\theta_{ij} \mid y_{ij}\right]=\left(\frac{1}{\psi_{ij}}+\frac{1}{\sigma_i^{2}+\tau^{2}}\right)^{-1}$$

θ_{ij} の推定には $E\left[\theta_{ij} \mid y_{ij}\right]$ を用いる。これはちょうど、観測値と全国の値を分散の逆数で加重平均した形になっている。

$$E\left[\theta_{ij} \mid y_{ij}\right]=\gamma\, y_{ij}+(1-\gamma)\beta$$

$$\left[\gamma=\frac{1}{\psi_{ij}}\left(\frac{1}{\psi_{ij}}+\frac{1}{\sigma_i^{2}+\tau^{2}}\right)^{-1}=\frac{\sigma_i^{2}+\tau^{2}}{\psi_{ij}+\sigma_i^{2}+\tau^{2}}\right]$$

観測値の誤差ψ_{ij}が大きい場合にはγがゼロに近くなり、観測値にかかるウエイトが小さくなる。逆に地域内や地域間のばらつきσ_i^{2}、τ^{2}が大きい場合には、全国の値βにかかるウエイトが小さくなる。

この手法は、モデルの設定やパラメータの推定において、かならずしも客観的な基準がないために、我が国の政府統計として適用するのは難しい面がある。

6　スタイン（Stein）タイプの推定量

統計局が以前に小地域推計の手法として検討していた手法にJames-Stein推定量がある。

$\theta_i\ (i=1,2,\cdots,k)\quad k\geq 3$ の推定量$\hat{Y}_i$が、独立に分散１の正規分布 $N(\theta_i,1)$ に従うとき、

$\hat{\theta}_i=\hat{\bar{Y}}+\left(1-\frac{(k-2)}{S}\right)\cdot(\hat{Y}_i-\hat{\bar{Y}})$ は、平均平方誤差（MSE）（$\sum_i(\theta_i-\hat{\theta}_i)^2$ の期待値）を元の平均平方誤差（$\sum_i(\theta_i-\hat{Y}_i)^2$ の期待値）よりも小さくする。ただし、$\hat{\bar{Y}}=\frac{1}{k}\sum_{i=1}^{k}\hat{Y}_i$ 、$S=\sum_{i=1}^{k}(\hat{Y}_i-\hat{\bar{Y}})^2$ である。

さらに、この Stein タイプの推定量は、S < k-2 のときに$\left(1-\frac{(k-2)}{S}\right)$ が負になるので、負のときに０で打ち切った positive-part James-Stein 推定量を考え、式を展開すると、

$$\hat{\theta}_i=(1-\omega)\cdot\hat{\bar{Y}}+\omega\cdot\hat{Y}_i \qquad \omega=\left(1-\frac{(k-2)}{S}\right)^{+}$$

ただし、$a^{+}=\max(0,a)$

の形に変形することができる。すなわち、$\hat{Y}_i$ と$\hat{\bar{Y}}$をウエイトωを用いて配分している。式の形から、ω が小さければ全体の情報に引っ張られることになる。

ω が小さいとき、個々の推定量は全体の情報に引っ張られる。これは、全体の MSE を小さくするためである。個々の推定量の MSE をある程度保存するために、$\hat{\theta}_i$ の移動量を制限した推定量を用いることもある。その推定量は以下のようなものである。

$$\tilde{\theta}_i=\begin{cases}\hat{Y}_i-c & \hat{\theta}_i<\hat{Y}_i-c\\ \hat{\theta}_i & \hat{Y}_i-c\leq\hat{\theta}_i\leq\hat{Y}_i+c\\ \hat{Y}_i+c & \hat{Y}_i+c<\hat{\theta}_i\end{cases}$$

ここで、c は適当な正の定数である。（ただし、これは$\hat{Y}_i$ が分散1の分布に従うとした場合の形である。分散が１でない場合はc に分散の平方根を乗ずる形になる。）

この手法は、地域の特性を反映した推計値であって有効であるが、全体の MSE を改良する推定量

であり、個々の地域のMSEを改良するわけではないので、地域によってはかえって精度を悪くすることもある。また、James-Stein 推定量は、経験ベイズ推定量としての解釈も可能である。

7 SPREE (Structure Preserving Estimation)

ここまで述べてきたような複雑な小地域推計の手法を検討する前に、単純な手法で小地域推計を試みており、そのときの手法がSPREEである。

SPREEは、繰り返し比例補正（Iterative Proportion Fitting）の方法を用いて、複数のクロスがある表のセルの合計が、特定の周辺値に一致するように調整する。周辺値は調査から得られた信頼できる推定値である。表の各セルの数値は直近のセンサスの値から得られる。

このようにセンサス時に得ることのできる小地域の値をセンサス間でも推計できる。

３つのクロスがあるセンサスの値について、hは小地域、iはある変数、gはiに関連する変数を表す。$h\times i\times g$ の三次元の分割表を作成する。未知の数値を$x_{h,i,g}$ で表し、i，gについての和、hについての和をそれぞれ$x_{h,\bullet,\bullet}$、$x_{\bullet,i,g}$に対応する分割表の周辺分布は$m_{h,\bullet,\bullet}$、$m_{\bullet,i,j}$ となる。

以下の繰り返し計算によって、$x_{h,i,g}$ の値を調整する。

STEP1 ：$x_{h,i,g}$ の初期値 $x_{h,i,g}^{(0)}$ に直近のセンサスの値 $N_{h,i,g}$ を代入する。

STEP2： $${}_1x_{h,i,g}^{(k)}=\frac{x_{h,i,g}^{(k-1)}}{x_{\bullet,i,g}^{(k-1)}}m_{\bullet,i,g}$$ （列方向の調整）

STEP3： $$x_{h,i,g}^{(k)}=\frac{{}_1x_{h,i,g}^{(k)}}{{}_1x_{h,\bullet,\bullet}^{(k)}}m_{h,\bullet,\bullet}$$ （行方向の調整）

以下STEP２とSTEP３を収束するまで凝り返す。

この手法は、推計値の初期値（センサスの値）に依存しており、センサス間で構造があまり変化しなければ不偏な推計となるが、仮定が満たされない場合には、精度が低下する。

II 労働力調査への適用

1 背景

雇用情勢が悪化していた時期に、各方面から雇用・失業情勢の詳細な把握のため、現在公表している労働力調査の地域（北海道、東北、南関東、北関東・甲信、北陸、近畿、中国、四国、九州の10ブロック）別結果より細かい小地域（都道府県）別結果の公表が要請されるようになった。そうした状況をかんがみ、統計局では、月又は四半期ごとの都道府県別結果を公表できるようにするために、前述のような統計的な手法による推計を検討してきたところである。

2 推計方法

各推計手法を労働力調査に適用し、都道府県別結果の推計を行い、その推計結果に基づき、労働力調査の都道府県別結果に最適な推計手法について検討した。各推計手法について、推計手法として一般性（認知されているか否か）、再現性、簡明性（理解が得られるか否か）、労働力調査に適用することについて、政府統計としての適用性、推計結果からみた適用性、実用性（必要な時期に推計可能か）の６つの観点から、総合的に評価を行った。

その結果、推計方法としては、アメリカの労働統計局で採用している時系列回帰モデルを基本に、我が国の労働力調査に適用させるために、回帰項に空間（地域）情報を取り入れたモデルとした。時系列回帰モデルとして、次のような回帰項、トレンド項、季節変動項、不規則変動項、標本誤差項によるモデルと仮定する。

$$y(t) = X(t)\beta(t) + T(t) + S(t) + I(t) + e(t)$$

各項は次のようなモデルで表現される。

回帰項：$\beta(t) = \beta(t-1) + v_\beta(t)$

$$v_\beta(t) \sim N(0, \sigma_\beta^2)$$

ﾄﾚﾝﾄﾞ項：$T(t) = T(t-1) + v_T(t)$

$$v_T(t) \sim N(0, \sigma_T^2)$$

季節変動項：$S(t) = -\sum_{j=1}^{11} S(t-j) + v_S(t)$

$$v_S(t) \sim N(0, \sigma_S^2)$$

不規則変動項：$I(t) = v_I(t)$

$$v_I(t) \sim N(0, \sigma_I^2)$$

標本誤差項：$e(t) = \gamma(t) e^*(t)$

（ $\mathrm{var}(e(t)) = \sigma_e^2(t)$ ）

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

$$v_e(t) \sim N(0,1)$$

$e^*(t)$ は労働力調査のローティション・サンプリングによる標本誤差項 $e(t)$ 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。

上記のモデルは、次のように状態空間表現で表すことができる。

$$y(t) = H(t)\alpha(t) + w(t) \quad \{\text{観測方程式}\}$$
$$\alpha(t) = F\alpha(t-1) + GV(t) \quad \{\text{遷移方程式}\}$$

$\alpha(t)$：状態変数ベクトル，F：遷移行列，
$H(t)$：観測行列，$w(t)$：観測ノイズ，
G：駆動行列，$V(t)$：システムノイズ

これにカルマン・フィルタを適用して各確率過程に分解し、標本誤差項を推計する。観測値から推計した標本誤差項を除去した値が求める推計値である。時系列回帰モデルについて、当初、回帰項のための適当な説明変数を見つけられなかった。都道府県別に結果が得られる雇用・失業に関する統計データとしては、職業安定業務統計、雇用保険業務統計及び毎月勤労統計などがある。これらの統計データを説明変数として試算した結果は、推計精度を高める結果とはならなかった。回帰の当てはまりが悪い場合には、回帰項を用いない推計もありえ、実際にその方がよいと思われる結果となっていた。

そこで、カナダ統計局の時系列情報と空間情報の二次元の回帰の考え方をヒントとして、回帰項の説明変数に時系列情報を用いるのではなく、近隣地域情報を用いて試算したところ、回帰項なしよりも精度が向上したため、この方法を採用することとした。労働力調査の 10 地域ブロックのそれぞれのトレンドをあらかじめ推計し、そのトレンドを近隣地域情報とした。

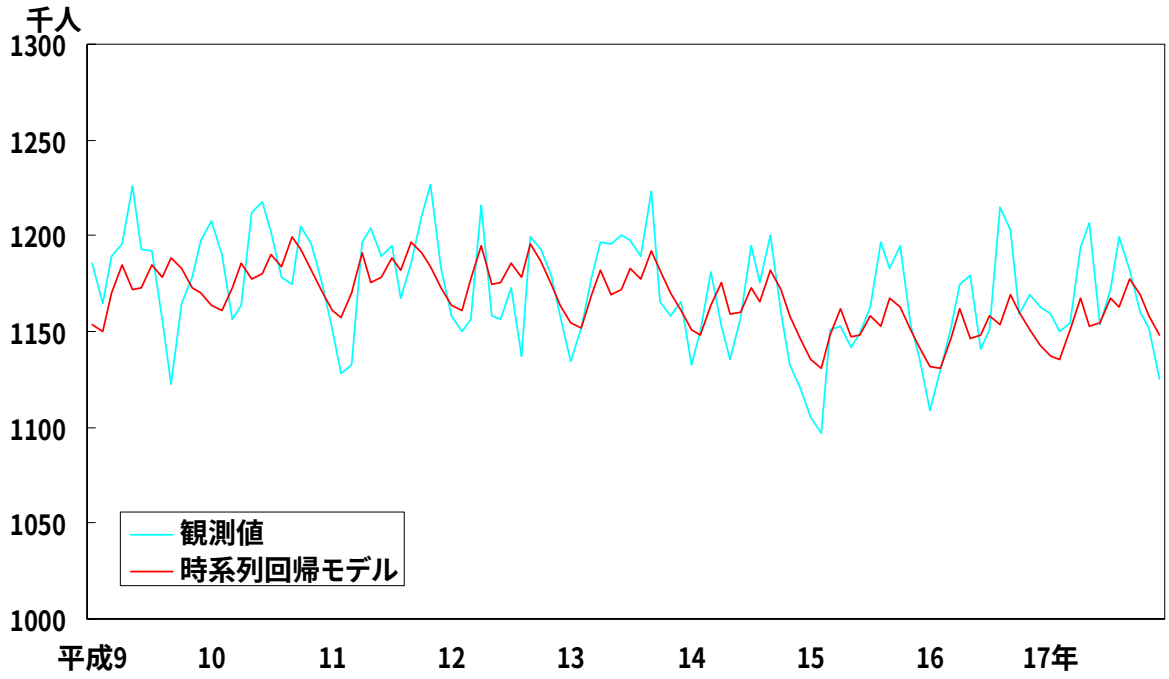
3　公表

推計した都道府県別四半期結果は、通常の推定値と区別するために「モデル推計値」として、平成 18 年 5 月 30 日に公表された。公表内容は、労働力人口、就業者、完全失業者、非労働力人口、完全失業率で、四半期ごとに公表される。なお、利用者の便を配慮し、平成 9 年まで遡及した結果を公表している。

（もとやま　ひとし
青山学院大学(総務省統計局 雇用統計地域推計研究会 元委員)
やまぐち　こうぞう
一橋大学(元総務省統計局)）

労働力調査都道府県別結果の推計例

就業者



完全失業者