経済統計・政府統計の数理的基礎と応用-IV[1]

国友直人・山本拓 共編

2015 年 1 月

[1] この報告集は文部科学省・科学研究費プロジェクト「経済統計・政府統計の数理的基礎と応用」（2011 年度～2014 年度）が開催した研究集会における講演内容をまとめたものである。

前書き

本報告書は、日本学術振興会・科学研究プロジェクト「経済統計・政府統計の数理的基礎とその応用」（2011 年度－2014 年度、研究代表者：山本　拓）が、2015 年 1 月 30 日(金)に東京大学小島ホールにおいて開催した 2014 年度の研究集会における講演内容をまとめたものである。なお研究集会は毎年度開催しているので、今回は４度目で最終回となる。

本プロジェクトの目的は、経済統計・政府統計における主要な課題の、技術的および制度的問題を、統計学的な立場から理論的・学術的に検討し、具体的解決策を提言することである。

経済統計、とりわけ政府統計は、経済・社会の動向を理解し、政策を実施、評価するためには不可欠な情報であることは言うまでもない。最近は evidence-based policy ということもよく言われ、政府統計の重要性は一般に広く認識されつつあると思われる。しかし、経済統計・政府統計への信頼性は、近年必ずしも増しているとは言えない状況である。経済社会の急激な変化に伴い、政府統計の質の確保が困難になりつつある。マクロ経済統計の側面では、GDP 統計などに代表されるマクロ公表系列の質と信頼性の問題、信頼性の高い将来人口の推計の問題、地域による経済情勢のばらつきの把握などの問題を挙げることができる。またミクロ経済データにおいては、統計調査をとりまくプライバシー意識の高まりから、調査精度の確保が難しくなりつつあるという問題や、情報開示と秘密保持の両立という匿名化問題などを挙げることができる。

新しい統計学的知見の導入に関しては、日本の政府統計部局が分散化されているために、これまでは、個別の担当部局あるいはその時々の担当者に個別に招かれた研究者によって知見や助言が提供されることが多かった。政府統計を巡る重要な論点について、担当部局をまたいでその知見が共有されることは少なかったと思われる。またそれらの話題が広く研究者間で議論されることも少なかった。そのような意味で、経済統計・政府統計の技術的・制度的問題点を、統計学的立場から総括的に検討していくという本研究プロジェクトは、一つの新しい方向性を目指したものである。

本プロジェクトの研究集会は、プロジェクトのメンバーと実際に経済統計・政府統計に作成者または利用者として携わっている方々との積極的な交流をその重要な目的の一つとしている。したがって研究集会における研究報告は、それらの方の報告を多く含むように構成されている。

2011 年度の第 1 回目の研究集会の特徴は、外部の報告者として、実際に主要な政府統計を作成されている厚生労働省，内閣府，総務省の担当者を招き、作成上のポイント

や課題を報告して頂いたことである。さらに人口統計の推計ならびにパネルデータ作成上の課題に関しても報告して頂いた。2012 年度の第２回目の研究集会の特徴は、外部の報告者としては地方政府において統計に関わっている方に、そのあり方や課題などについて報告して頂いたことにある。さらにマクロ経済統計の作成者および利用者としての立場からその問題点や改善の方向性についての報告を頂いた。2013 年度の第３回目の研究集会は、外部の報告者として雇用・失業統計、人口統計、ならびに生産性統計についての報告をして頂いた。さらに季節調整の様々な問題について、海外からの招聘研究者の報告とともに、プロジェクトのメンバーの報告が英語セッションとして行われた。

これら３回の研究集会での報告は、それぞれ東京大学大学院経済学研究科付属・日本経済国際共同研究センター（CIRJE）研究報告書シリーズの CIRJE-R-10, CIRJE-R-12 ならびに CIRJE-R-16 にまとめられている。

今回の研究集会は、本プロジェクトの最終年度ということもあり、プロジェクト・メンバーの報告を主体とした構成となっている。ただし、外部からの報告として日次データによる物価指数の作成についての問題・課題を報告して頂いた。また最近の重要なトピックである小地域統計の問題を特別セッションとして取り上げた。このセッションにおいて外部の方から小地域統計の応用例について報告を頂いた。

以上は，主に各研究集会における外部の方の貢献を明示的に述べたが、本プロジェクトのメンバーは、これら４回の研究集会を通じて活発に研究報告を行ってきた。その主たる領域は以下のようにまとめられる。それらは、標本調査の実際と課題、消費統計ならびに物価統計にまつわる問題、匿名化と開示リスクの理論と課題、小地域統計の理論と応用、季節調整に関する様々な問題についての理論と応用、ベンチマーク問題、人口統計の推計問題等である。

このような研究集会が情報交換ならびに刺激となり、経済統計・政府統計の今後の改善の一助になることを期待する次第である。

2015 年 2 月

編者

プログラム

<セッションⅠ>　経済統計・政府統計を巡る諸問題
Chair:　山本拓
13:00-13:30　「調査票デザインに関する実験」 土屋隆裕（統計数理研究所）
13:30-14:00　「住宅・土地統計調査の開示リスク評価」 星野伸明（金沢大学）
14:00-14:30　「Estimating Daily Inflation Using Scanner Data」 渡辺努（東京大学）
14:30-15:00　「消費関連統計の比較」 宇南山卓（財務省財務総合政策研究所）

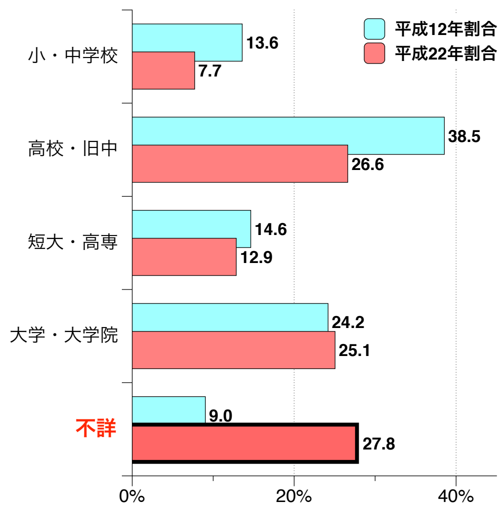
<セッションⅡ>　特別セッション：小地域統計の理論と応用
Chair:　国友直人
15:10-15:30　「変量分散モデルを利用した小地域推定」 久保川達也（東京大学）
15:30-15:50　「乗法モデルとベンチマーク問題」 川久保友超（東京大学大学院）
15:50-16:10　「データ変換と小地域推定」 菅澤翔之助（東京大学大学院）

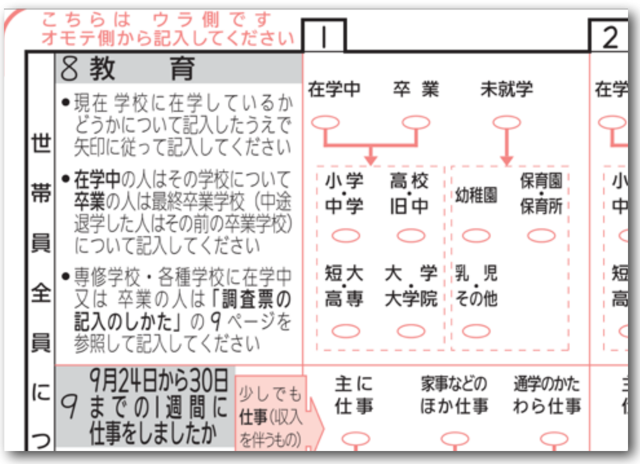
<セッションⅢ>　経済時系列の基礎
Chair:　高岡慎
16:20-16:50　「観測誤差と線形制約を伴う真のデータの推定に関する新たな近接法」 千木良弘朗・山本拓（東北大学・日本大学）
16:50-17:10　「トレンド・季節性とマクロ時系列」
佐藤整尚・国友直人（東京大学・東京大学）
17:10-17:40　「地域統計の季節調整問題」
川崎能典・国友直人（統計数理研究所・東京大学）

調査票デザインに関する視線追跡実験

土屋 隆裕（統計数理研究所）

国勢調査の卒業者学歴（東京都）





→ 低学歴層が「不詳」に？

八王子市民を対象とした郵送による比較実験調査

大学共同利用機関法人 情報・システム研究機構
統計数理研究所

整理番号
(謝礼をお送りするための番号です)

多摩地域　住民意識調査

・ご回答は、当てはまる**数字**を○で囲んでください。
・ご記入の終わった調査票は、同封の返信用封筒でご返送ください。後日QUOカードをお送りいたします。

はじめに、現在お住まいの市町村との関わりについておうかがいします

問１　あなたは、現在お住まいの市町村に住んで何年くらいになりますか？（○は一つ）
1 20年以上
2 10年以上 20年未満
3 5年以上 10年未満
4 5年未満

問２　あなたは、現在お住まいの市町村にこれからも住み続けたいと思いますか？（○は一つ）
1 住み続けたい
2 できれば住み続けたい
3 できれば他の市町村に移りたい
4 他の市町村に移りたい
5 その他

問３　あなたがお住まいの地域は、近隣地域と比べたとき、以下のことがらが当てはまりますか、それとも当てはまりませんか？（○はそれぞれ一つずつ）

	当てはまる	やや当てはまる	あまり当てはまらない	当てはまらない	その他
自然が多い	1	2	3	4	5
物価が安い	1	2	3	4	5
交通の便が良い	1	2	3	4	5
治安が良い	1	2	3	4	5
騒音が少ない	1	2	3	4	5
商業施設が充実している	1	2	3	4	5

大学共同利用機関法人 情報・システム研究機構
統計数理研究所

多摩地域 住民意識調査

・ご回答は、当てはまる**数字**を○で囲んでください。
・ご記入の終わった調査票は、同封の返信用封筒でご返送ください。後日 QUO カードをお送りいたします。

はじめに、現在お住まいの市町村との関わりについておうかがいします

問１　あなたは、現在お住まいの市町村に住んで何年くらいになりますか？（○は一つ）
1 20年以上
2 10年以上 20年未満
3 5年以上 10年未満
4 5年未満

問２　あなたは、現在お住まいの市町村にこれからも住み続けたいと思いますか？（○は一つ）
1 住み続けたい
2 できれば住み続けたい
3 できれば他の市町村に移りたい
4 他の市町村に移りたい
5 その他

問３　あなたがお住まいの地域は、近隣地域と比べたとき、以下のことがらが当てはまりますか、それとも当てはまりませんか？（○はそれぞれ一つずつ）

	当てはまる	やや当てはまる	あまり当てはまらない	当てはまらない	その他
自然が多い	1	2	3	4	5
物価が安い	1	2	3	4	5
交通の便が良い	1	2	3	4	5
治安が良い	1	2	3	4	5

整理番号
(謝礼をお送りするための番号です)

問４　あなたは、「町内会・自治会」に所属していますか？（○は一つ）
1 所属している
2 所属していない
3 町内会･自治会はない
4 わからない

1 を選んだ方は**問4-A**もお答えください

問4-A　どのような理由から所属しましたか？（○は一つ）
1 慣習・決まりなので
2 メンバーに勧誘されたので
3 関心があったので
4 その他

問５　あなたは、町内会・自治会などの地域の自治組織が行っている以下の活動に参加していますか？（○はそれぞれ一つずつ）

	だいたい参加している	ときどき参加している	全く参加していない
地域防犯に関連する活動（例：防犯パトロールや屋外灯の夜間点灯、あいさつ運動等）	1	2	3
地域防災に関連する活動（例：防災訓練、火災訓練、防災マップ作成、災害用備蓄等）	1	2	3
地域の環境保全に関連する活動（例：地域の掃除、リサイクル活動等）	1	2	3
地域の親睦に関連する活動（例：住民相互の連絡、スポーツ、文化祭、祭り、盆踊り等）	1	2	3

問５で一つでも 1 または 2 を選んだ方は**問5-A**もお答えください

問5-A　**問５**で参加している地域活動について、以下のことがらが当てはまりますか？（○はそれぞれ一つずつ）

	当てはまる	やや当てはまる	あまり当てはまらない	当てはまらない
活動を通じて地域をより良くしたいと思う	1	2	3	4
地域活動そのものが好きだ	1	2	3	4
地域活動がうまくいった時には、満足感を感じる	1	2	3	4

地域活動にかかわることは、自分自身の価値

問４　あなたは、「町内会・自治会」に所属していますか？（○は一つ）

1 所属している　　2 所属していない　　3 町内会･自治会はない　　4 わからない

1 を選んだ方は**問4-A**もお答えください

問4-A　【**問４**で「1 所属している」と答えた方のみ】
どのような理由から所属しましたか？（○は一つ）

1 慣習・決まりなので
2 メンバーに勧誘されたので
3 関心があったので
4 その他

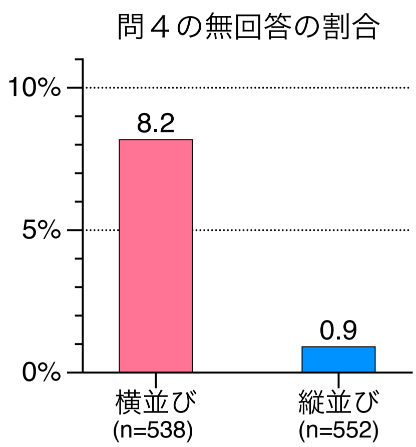
問４　あなたは、「町内会・自治会」に所属していますか？（○は一つ）

1 所属している
2 所属していない
3 町内会･自治会はない
4 わからない

1 を選んだ方は**問4-A**もお答えください

問4-A　どのような理由から所属しましたか？（○は一つ）

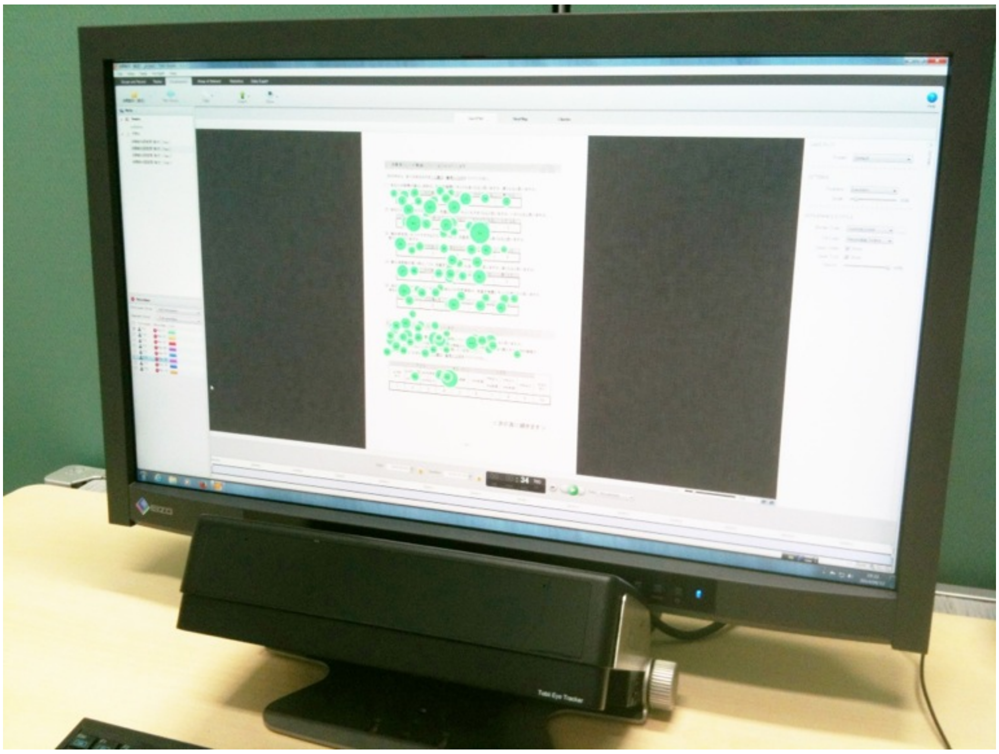
1 慣習・決まりなので
2 メンバーに勧誘されたので
3 関心があったので
4 その他



調査票の設計にあたって

- 一般的な方法として
  - ✓予備調査
  - ✓回答者へのインタビュー
- 一方で
  - ✓商品のパッケージデザインやWebデザインの設計には視線追跡装置

視線追跡装置



消費動向調査

- 調査客体は一般世帯と単身世帯合わせて8,400世帯
- 毎月15日を調査時点として、毎月１回調査を実施
- 調査世帯は15ヶ月継続で、15グループによるローテーションを行う
- 平成25年3月までは訪問留置法
- 平成25年4月からは郵送調査法（１ヶ月目世帯は調査員による訪問・回収）

政府統計

統計法に基づく国の統計調査です。調査票情報の秘密の保護に万全を期します。

㊙ 総務省承認　一般統計調査

内　閣　府

調査時期 年 月	都道府県番号	市町村番号	調査単位区番号	世帯番号

「消費動向調査」調査票

（月調査用）

【ご記入にあたってのお願い】

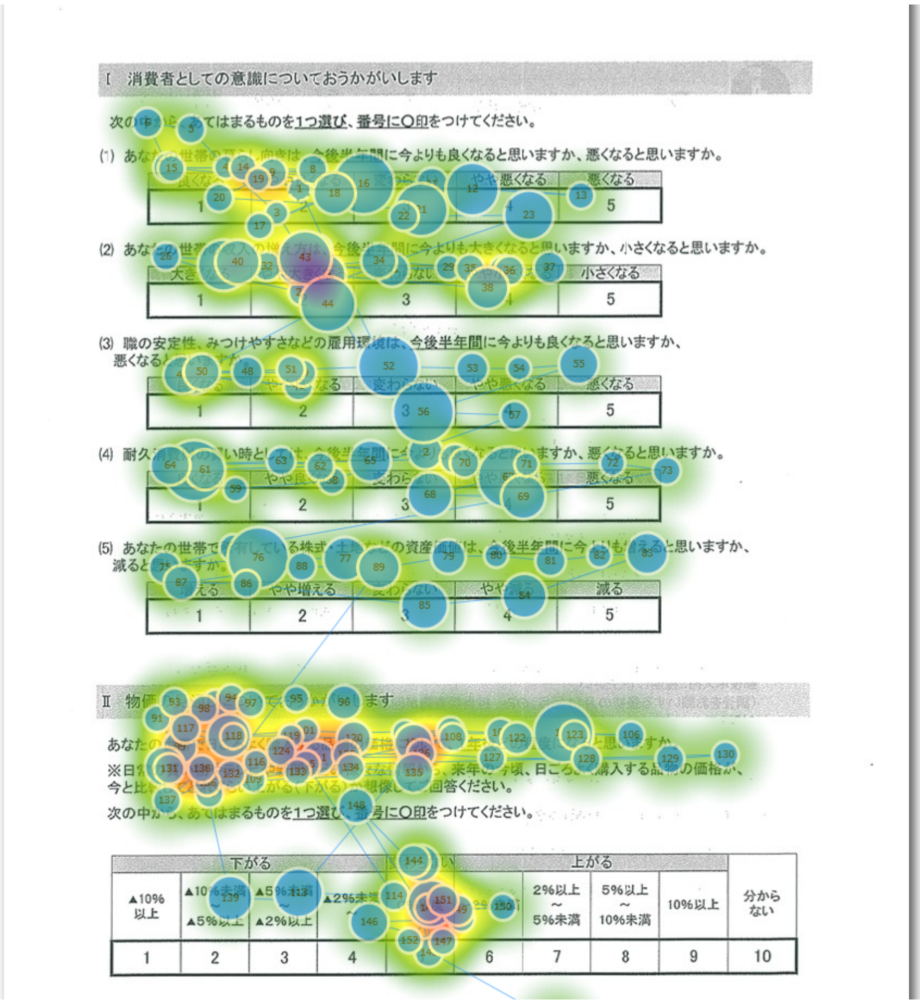
1. ご記入にあたっては、黒鉛筆を用いてはっきりと記入してください。
2. この調査は、平成26年３月１日現在を基準に行います。
3. この調査票に回答していただいた内容は、統計作成以外の目的、例えば税金の徴収などに使用されることは絶対ありませんので、ありのままご記入ください。
4. ご回答は、選択肢の番号に○をつける場合と、数字などを記入していただく場合があります。
5. 質問によっては、次に回答していただく質問を示す矢印（→）やことわり書きなどがあります。それらにしたがって、ご回答ください。
6. 調査票の記入が終わりましたら、同封の返信用の封筒に入れて、平成26年３月15日までに郵便ポストに投函してください。
（調査をお願いする最初の月（1回目）のみ、担当調査員が回収に伺いますので、調査票を直接お渡しください。）

ご回答いただく上でご不明な点、調査に関するお問い合わせは、下記までお願いいたします。

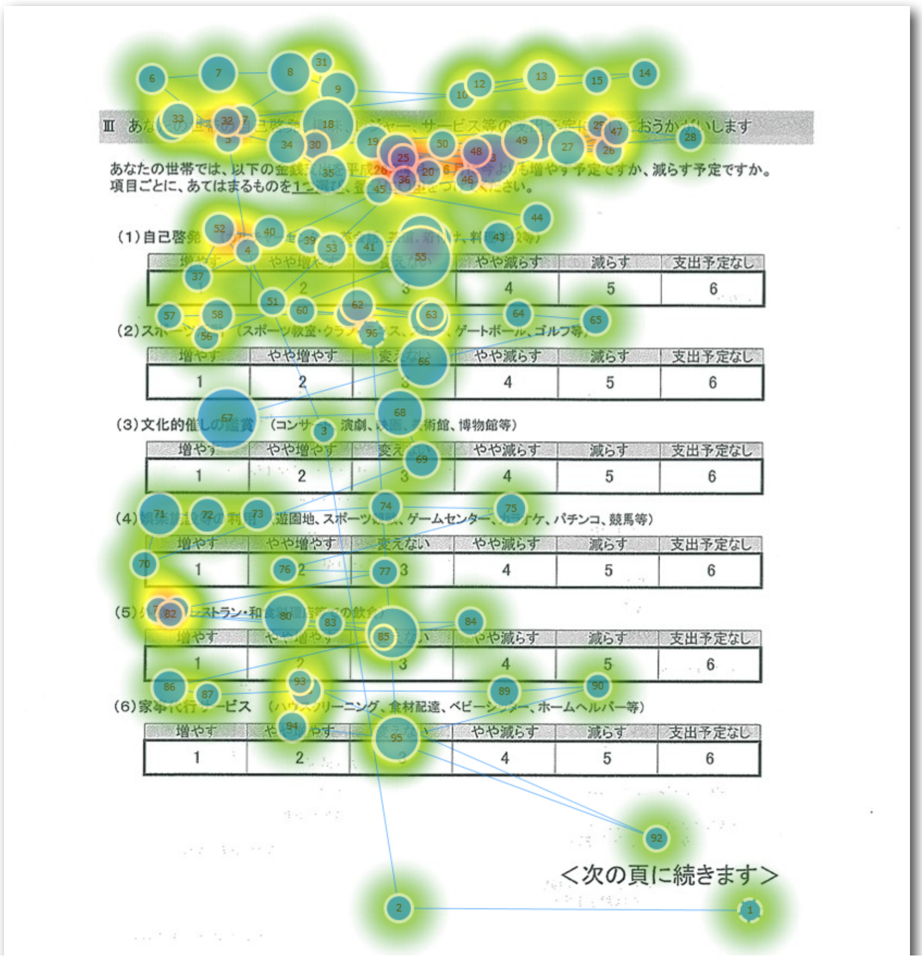
問合せ先：一般社団法人　新情報センター

住所：〒150-0013　東京都渋谷区恵比寿1-19-15

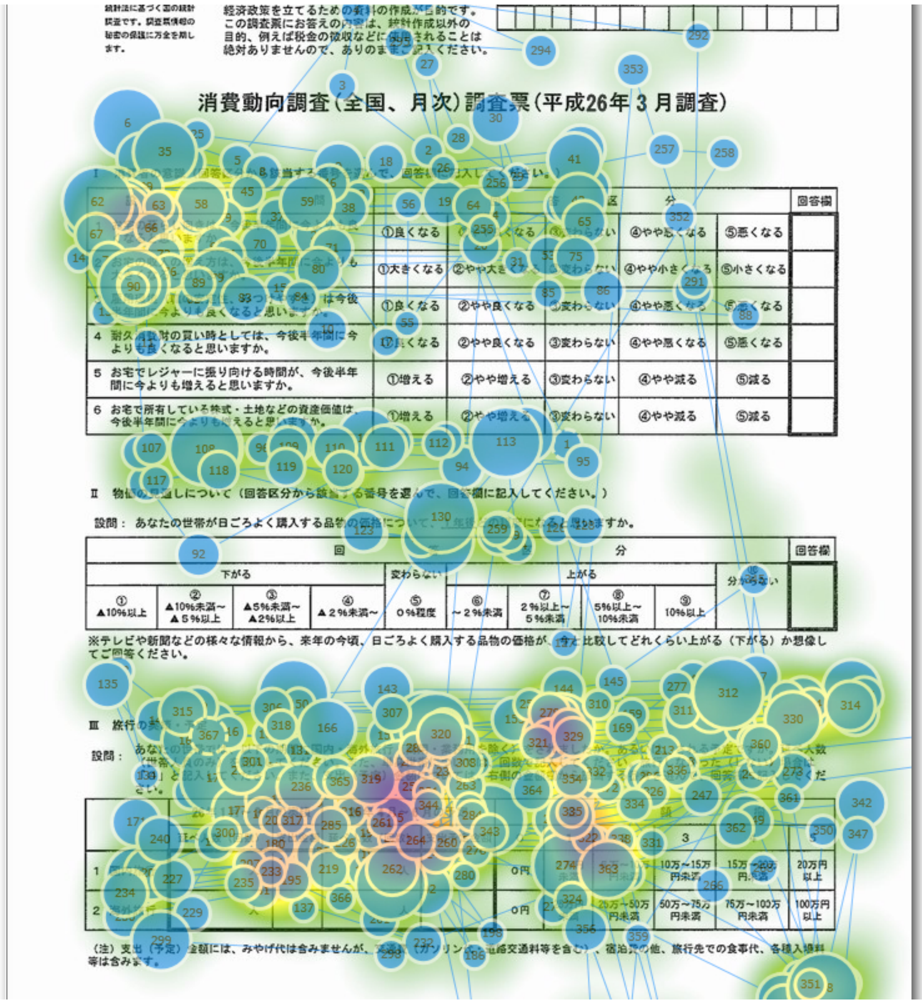
電話：ﾌﾘｰﾀﾞｲﾔﾙ　0120-78-5231　（受付時間：平日9～17時　※12～13時除く（担当：平栗、牛島））


I　消費者としての意識についておうかがいします
(1) あなたの世帯の暮らし向きは、今後半年間に今よりも良くなると思いますか、悪くなると思いますか。
(2) あなたの世帯の収入の増え方は、今後半年間に今よりも大きくなると思いますか、小さくなると思いますか。
(3) 職の安定性、みつけやすさなどの雇用環境は、今後半年間に今よりも良くなると思いますか、悪くなると思いますか。
(4) 耐久消費財の買い時としては、今後半年間に今よりも良くなると思いますか、悪くなると思いますか。
(5) あなたの世帯で保有している株式や土地などの資産価値は、今後半年間に今よりも増えると思いますか、減ると思いますか。
Ⅱ　物価の見通しについておうかがいします
次の中から、あてはまるものを1つ選び、番号に〇印をつけてください。
下がる
上がる
▲10%以上
▲2%未満
2%以上 ～ 5%未満
5%以上 ～ 10%未満
10%以上
分からない
1
2
3
4
5
6
7
8
9
10

15


Ⅲ　あなたの世帯の主要なサービス等の支出予定についておうかがいします
あなたの世帯では、以下の金銭を支出予定ですか、減らす予定ですか。
(1) 自己啓発
(2) スポーツ活動（スポーツ教室・クラブ、ゲートボール、ゴルフ等）
(3) 文化的催しの鑑賞（コンサート、演劇、美術館、博物館等）
(4) 遊園地、スポーツ観戦、ゲームセンター、パチンコ、競馬等
(5) レストラン・和食等の飲食
(6) 家事代行サービス（ハウスクリーニング、食材配達、ベビーシッター、ホームヘルパー等）
増やす
やや増やす
変えない
やや減らす
減らす
支出予定なし
<次の頁に続きます>

17



まとめ

- 回答者によっては、調査票の表紙をほとんど読まないかもしれない
- 回答者によっては、段階評定の回答選択肢を全て読まないかもしれない
- 回答者によっては、調査票の右方向にはあまり注意を向けないかもしれない
- マトリックス形式の回答欄は、回答者には負担

☑ 自記式調査では、視線追跡装置を試みる価値あり

住宅・土地統計調査の開示リスク評価

星野　伸明
金沢大・経

2015 年 1 月 30 日

概要

1. 問題意識：データ匿名化の程度を合理的に定めたい。
   - 主観的問題とみなされているが、社会科学的モデル化できるはず。
2. 個体識別ができない状態の定式化
   (a) 個体識別の判別モデルと観測モデル
   (b) 母数推定
3. 開示リスク測度の計量
   (a) キー変数の選択方法
   (b) 住宅・土地統計調査 (H15) 匿名データの例

研究の方針

- 匿名化処理はどこまで減らせるか？
  - 匿名化が満たさなければならない条件は法律が定める。
    - 参）匿名データの定義（統計法第２条第１２項）:「一般の利用に供することを目的として調査票情報を特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む。）ができないように加工したもの」
    - 参）個人情報保護法による個人情報の定義：個人が識別できるデータ。
- ⇒ どの程度匿名化すれば（再）識別が出来ないのか？
  - 統計モデルを用いて観測から推定する。
    - 識別が可能な状態の定義から始める。

Marsh らの個体識別モデル

$$\Pr(\text{識別が実際に起きる}) = \Pr(\text{識別成功} \mid \text{識別を試みる}) \Pr(\text{識別を試みる}) \tag{1}$$

$$\Pr(\text{識別成功} \mid \text{識別を試みる}) = \Pr(a) \Pr(b|a) \Pr(c|a, b) \Pr(d|a, b, c) \tag{2}$$

ただし

(a) 攻撃用ファイルと公開ファイルのキー変数が同じ基準で記録されている。

- 照合をかける対象範囲が「キー変数」。

(b) 公開ファイルに個体が含まれている。

(c) 個体が母集団一意である。

- 一意に照合される個体が母集団でも一意ということ。

(d) 個体が母集団一意と確証出来る。

- 既存情報で一意数は評価される。
- 本報告では追加情報による攻撃を確証の一種として考慮。

個体識別が不可能という状態

- 「絶対的な匿名性」: 識別の可能性が疑う余地なく除かれている状態 $\Leftrightarrow$

$$\Pr(\text{識別成功} \mid \text{識別を試みる}) = 0 \tag{3}$$

- 「事実上の匿名性」: 識別の費用が便益を上回る状態 $\Leftrightarrow$(1) 式の確率が低い。
- 以下では状態 (3) が「個体識別が不可能」とみなす。

$$\Pr(\text{識別成功} \mid \text{識別を試みる}) = \Pr(a, b, c)\Pr(d|a, b, c) \tag{4}$$

  - $\Pr(a, b, c)$ か $\Pr(d|a, b, c)$ のいずれかが 0 なら個体識別は不可能。

母集団一意の確証

- 通常は $\Pr(a, b, c) \neq 0$ なので、個体識別が可能か不可能かは $\Pr(d|a, b, c)$ が 0 か否かの問題になる。
- $\Pr(d|a, b, c) = 0$ とは母集団一意の確証が不可能ということ。
- 母集団一意の確証方法：
  - 一意たらしめているキー変数の組み合わせ（指紋）について全数情報を集める。
  - 全数 $\subseteq$ 母集団。部分集団で一意なら母集団でも一意。
    - 例）日本の弁護士集団で一意なら、日本人でも一意。
  - 全数名簿が存在したり作りやすい場合は、匿名化で手当するのが前提。

個体識別可能性の判別モデル

- 匿名化等によって母集団一意の確証要因はコントロールする。
  - それでも残る不確実性を母数 (β) に集約。
- 統計モデル化：適当な非負の β について

$$\Pr(a,b,c) \leq \beta \Leftrightarrow \Pr(d|a,b,c) = 0 \tag{5}$$

  - データが情報豊富なら、$\Pr(a,b,c)$ が高い。
  - 確証可能性は、データ情報度の単調関数と思われる。
- $\Pr(a,b,c)$ は母集団一意確証の「容易度」。

個体識別の観測モデル

- モデル (5) の母数 β を統計的に推定するには観測が必要。
- 個体識別が可能か否かは直接観測できないので、識別成功の社会的認知の有 ($X=1$) 無 ($X=0$) を観測：

$$
\begin{aligned}
\Pr(X=1) &= \Pr(\text{識別の社会的認知} \mid \text{個体識別が実際に起きる}) \\
&\quad \times \Pr(\text{個体識別が実際に起きる})
\end{aligned}
$$

- $\Rightarrow \Pr(a,b,c)$ の評価値を γ で表せば

$$
\Pr(X=1)=\begin{cases} p(\gamma) & \gamma>\beta \text{ の場合} \\ 0 & \gamma\le\beta \text{ の場合} \end{cases} \tag{6}
$$

- 適当な条件の下で $p(\gamma)>0$.

閾値の最尤推定量 $\hat{\beta}$

- 過去の（匿名化した）データ公開事例 ($i = 1, 2, \ldots, n$) をモデル (6) からの独立標本とみなす。i 番目の事例について $\mathrm{Pr}(a, b, c)$ の評価値 γ_i と個体識別発生認知の有無 x_i は観測できる。
- 過去に個体識別が認知されていない事例の中で $\mathrm{Pr}(a, b, c)$ の最も高い評価値を $\bar{\gamma}$ と書けば、β は $\bar{\gamma}$ 以上（かつ個体識別発生が認知されている事例の評価値未満）と最尤推定される。
- 過大推定（$\hat{\beta} > \beta$）の確率は、$p(\cdot)$ が 0 に近いほど高い。それから真の β より γ が高い事例が少ないほど高い。
  - 新規に公開するデータの $\mathrm{Pr}(a, b, c)$ を $\bar{\gamma}$ と等しくすれば、真の β の位置によらず、過大推定の確率は単調非増加。

閾値 β が共通する範囲

- モデル (6) からの独立標本とみなせる、つまり β が共通する事例の範囲が問題。
- （計量しないので）β を変化させる要因？
  - β は母集団一意の確証に関する不確実性を集約しているので、確証能力（全数情報収集力）の変動要因を考える。
    1. 母集団が違えば（全数情報も違うので）β も変わるはず。
    2. 公知の変数の種類の変化、変数の精度の変化、変数が既知な個体量の変化のうち、個体量の変化は $\mathrm{Pr}(a, b, c)$ の評価に反映されない。
       * 個体量が増えれば全数情報との差が縮小し、確証しやすくなるかもしれない。
- 計量しない要因の定量評価は難しいので、変化がなければ β が共通と判断。

匿名データの作成に関する含意

- 匿名データについては既公開の事例が存在し、$\bar{\gamma}$ が求められる。
  - 計量しない要因は「チェックリスト」で比較。
- 既公開の匿名データは識別が認知されていないので、最も識別が容易な匿名データの水準 ($\bar{\gamma}$) まで、他の匿名データの匿名化は緩和できる。
  - 推定誤差があるので、安全マージンは必要。
- 実際に $\Pr(a, b, c)$ が計れることを、平成 15 年の住宅・土地統計調査で実証。
  - 都道府県コードあり、攪乱なし、という特徴。
- $\Pr(a, b, c)$ の計量で問題は、キー変数の選択。

キー変数の選択方法について

- Elliot et al. (2011)：個体情報を広範に調査した上でキー変数を選択。
  - 攻撃用情報の見当がついたとして、いかにキーを選ぶか？
- Fung et al. (2010)："open problem".
- 本報告：匿名化水準の管理にとって最適に選ぶ。
  - 既存研究は使い方を定めないので選べない。
  - k 変数からキーを選ぶ方法は 2^k 通りで、そのうちどれを採用するかと考える。

キー変数の選択に係る一意数の変化

- 2^k 個の母集団一意数の順序データ: $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(2^k)}$
- 例）住宅・土地統計調査匿名データの部分評価 ($2^{11} = 2048$)
  - “11vars”：都道府県、住宅以外の建物の種類、住宅以外の建物の所有関係、建物の構造、建物の階数（うち一戸建て・長屋、うち共同住宅)、むねの建築時期、建築面積、敷地面積、エレベータの有無、高齢者対応か
  - “ex region”：都道府県削除
  - “ex date”：むねの建築時期削除

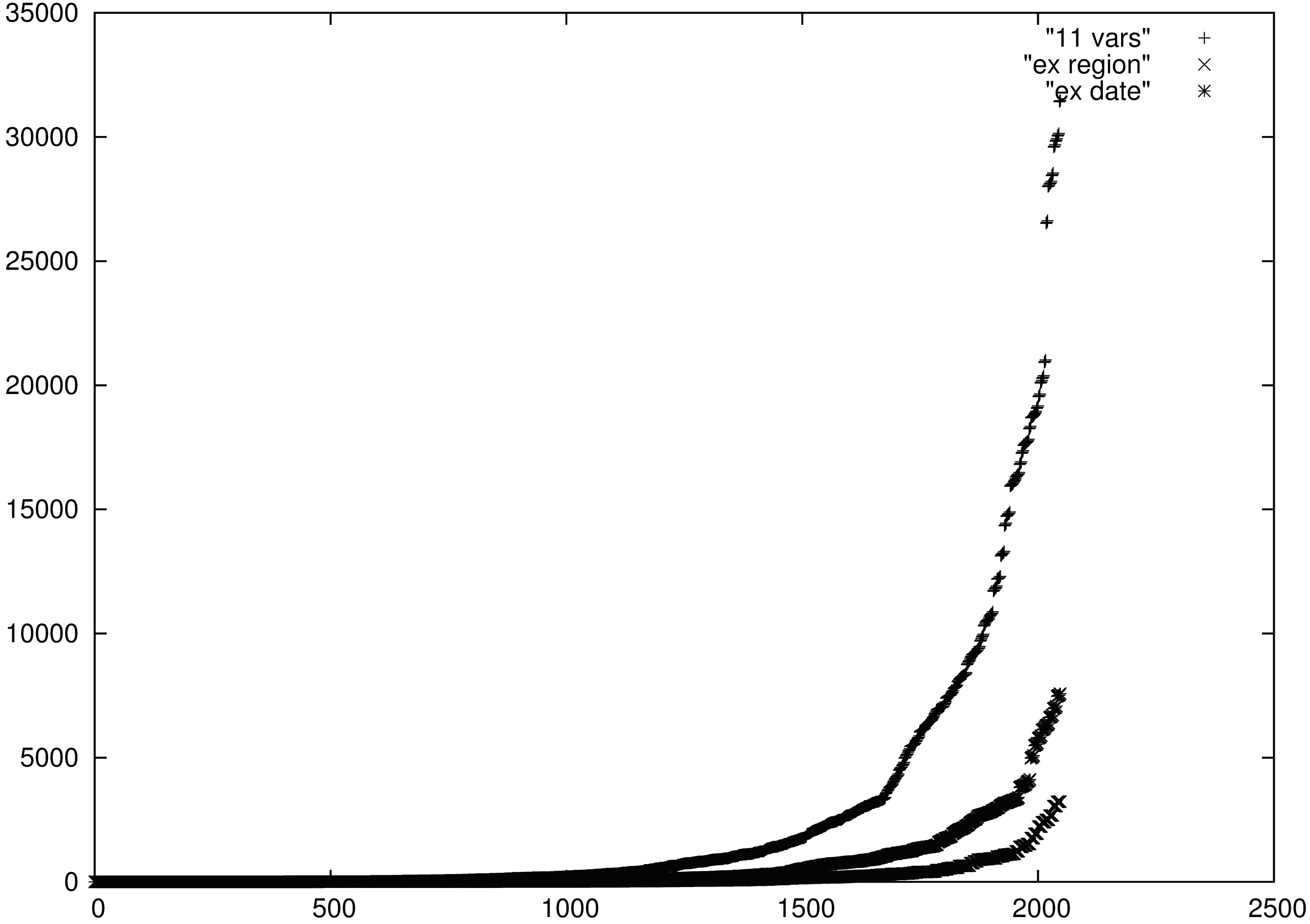


Figure 1: 標本一意数（縦軸）と順位（横軸）の関係

理論的なキー変数選択方針

- キー変数の選択とは、どの順位 r の $u_{(r)}$ を採用するか、という問題に他ならない。
- 選択した順位での一意数より小さい一意数を与えるキー変数しか使えない攻撃者は、（要因 (a,b) が一定なら）管理される。
- 順位 2^k を選べば全ての攻撃者を管理できる。しかし攻撃者がいない順位で評価したリスクは、実効リスクと違うので、識別不可能性の根拠にならない。
- $\Rightarrow$攻撃者が存在する最大の順位 で一意数を評価する。
- 順位 $(1, 2, \ldots, 2^k)$ 上の攻撃者の分布で、最大値を推定したい。
  - 攻撃者は能力の範囲内で最も一意数を多く得られる順位に存在すると考える。

実際的なキー変数選択方針

- しかし分布の最大値の推定は困難で、分位点推定の方が現実的。
  - 資本規制でも 99%分位点を管理 (VaR)。
- 実際にはデータがないので、攻撃者分布の分位点は定量的に推定できない。
- 考察の主旨を活かせば、「大半」の攻撃者を管理するという方針でキー変数を選ぶのが実際的。つまり「公知」の変数をキーとする。
- 大半の外の攻撃者は、匿名化では管理できない。
  - 匿名化以外の手法が有効。例えば攻撃者分布の右裾に位置するような主体（名簿業者、個人情報収集組織等）にデータを渡さなければよい。
  - 識別事故が起きたときにうまく対応すれば当局への信認は上がると MacKey (2009) は議論。

例）住宅・土地統計調査 (H15) 匿名データ

- 公表サンプルサイズ： 31 万 266（世帯）; 居住世帯ありのレコードのみ。
  - リサンプリング率 10%
- 母集団サイズ：4726 万（世帯）
- 標本抽出率=$\Pr(b|a)$：0.66 パーセント；単純無作為抽出とみなす。
- 撹乱は使われていないので $\Pr(a) = 1$ とみなす。

住宅・土地統計調査のキー変数

- Case 1：都道府県、世帯の種類、同居世帯の有無、夫婦の組数、家族類型、世帯の型、65 歳以上の世帯員の有無、75 歳以上の世帯員の有無、65 歳以上の世帯員のみか、75 歳以上の世帯員のみか、高齢夫婦の有無、世帯内の最高年齢
- Case 2：Case 1-都道府県
- Case 3：都道府県、世帯員各員について性別・年齢（15 歳未満は各歳）・配偶者の有無・続柄
- Case 4：Case 3+世帯主情報（性別、年齢、従業上の地位）
- Case 5：Case 4+現在の居住形態、所有の形態
- Case 6：Case 5+建物に関する事項、むねに関する事項、住宅の種類、所有関係、民営借家の所有区分、住宅の建て方、建築の時期
- Case 7：Case 6+地下室有無、自動車所有の有無、駐車スペースが敷地内、敷地外、住宅の購入・新築・建て替え等の別、H11 年以降の増改築有無
- Case 8：Case 7+台所、トイレ、浴室の設備状況

	S_1	$\Pr(c\|a,b)$	$\Pr(a,b,c)$
Case 1	4918819	.104	.00068
Case 2	1683983	.036	.00023
Case 3	5038968	.107	.00070
Case 4	6871365	.145	.00096
Case 5	9374185	.198	.00130
Case 6	29082561	.615	.00404
Case 7	35610454	.753	.00495
Case 8	42962590	.909	.00597

Table 1: 個体識別の容易度評価

まとめ

- 少なくとも匿名データの開示リスク管理方法は理論化した。
  - 開示リスク測度（$\Pr(a,b,c)$）の選択と閾値の推定
  - キー変数の選び方
  - これらは統一的に議論するべき問題であった。
    - モデル分析によってそれが可能となった。
- 他の匿名データについても個体識別の容易度評価を行う予定。
- 本研究で使用した匿名データは統計法に基づいて（独行）統計センターから提供を受けた。

エビデンスに基づいた匿名化

星野　伸明*

平成 27 年 1 月 20 日

Evidence Based Anonymization

Nobuaki Hoshino*

概 要

匿名データについて、個体識別が可能か否かの判定は定義に関わる。しかしこの判定は明確に定式化されておらず、改善のための議論を阻んでいる。従って本論文は、個体識別可能性の判定方式を定量評価に基づいて明確化する。このような判定方式に関する既存研究は存在するが、個体識別可能性の測度について閾値を定める理論が欠けている。この点について本論文では、個体識別が起きていないという観測可能な事実に基づいて閾値を推定する。このような観測に基づいて意思決定する態度は、エビデンスに基づいた匿名化と呼ぶのがふさわしい。この立場から、匿名データ審査体制について改善点が指摘できる。

Anonymized Data are defined so that no individual shall be identified. This unidentifiability, however, is not clearly defined. Hence the decision process of the unidentifiability has not been clearly formulated, which prevents explicit arguments on its improvement. Therefore the present paper clearly states a method to decide whether given data are identifiable or not, based on a measure of disclosure risk. The existing theory of disclosure risk lacks the method of deciding its critical value; the present paper estimates it using a fact that identification is not observed. Our method based on observations should be called evidence based anonymization. This theory results in concrete improvements on the assessment of Anonymized Data.

キーワード: 母集団一意, プライバシー, 統計的開示制限.

1　はじめに

匿名データは、平成２１年度に総務省所管の四調査（全国消費実態調査、社会生活基本調査、 就業構造基本調査、 住宅・土地統計調査）から提供が開始された。平成２６年５月現在、総務省所管の国勢調査、労働力調査や厚生労働省所管の国民生活基礎調査の匿名データも提供されている。新しい制度がこのように実績を重ねてきたことは喜ばしい。今後は実績という経験を活かし、制度を継続的に改善する道筋をつけるべきである。特に利用者からのデータ改善要求に応える必要があろう。

*金沢大学経済学類, 〒 920-0927, 石川県金沢市角間町, E-mail: hoshino@kenroku.kanazawa-u.ac.jp

匿名データは元の個票を変換（匿名化[1]）して作られる。例えば全国消費実態調査等の匿名データでは、15 歳から 84 歳までの年齢を 5 歳階級別に変換している。また地域情報は「3 大都市圏」及び「その他の地域」の 2 区分に変換している。このような変換により、各歳別の分析や詳細な地域別分析は不可能となる。データ分析において、匿名化は明らかに望ましくない。故に利用者の要求として、匿名化の緩和は典型的である。

しかし全ての匿名化を外せるわけではない。匿名データの定義（統計法第 2 条第 12 項）を引用すると、「一般の利用に供することを目的として調査票情報を特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む。）ができないように加工したもの」である。元の個票（調査票情報）はこの定義を満たすように匿名化されなければ、匿名データとして提供不可能[2]である。従って匿名化は、個体識別が不可能な範囲で少ない方がよい。つまり匿名データの改善の多くは、個体識別が可能か否かという判断を必要とする。

この判断について、総務省政策統括官（統計基準担当）(2011) による「匿名データの作成・提供に係るガイドライン」（以下、ガイドライン）には、審査用資料として「チェックリスト」を作成することが定められている。そして「チェックリストに記載された内容等を基に」、「匿名化処理の妥当性等に係る審査を実施する」とある。参考として世帯調査のチェックリスト（H23/3/28 改正版）の要約を付録に収めた。チェックリストは個体識別に関係する要因を記載しているはずである。しかしその使い方は説明されていない。

結局「一律に匿名化の基準を設定することは困難」なので「一橋大学における匿名標本データの試行的提供の事例[3]及び諸外国の統計機関における同様の提供の事例等を参考に」匿名化せよとガイドラインは書く。同様とはどのような事例で、それをいかに参考にしたらよいか。この点について判断は審査担当者の見識に委ねられている。個体識別可能と不可能の区別は、不明確である。

この区別の明確化、精密化は匿名データに関してだけの課題ではない。いわゆる個人情報保護法において個人情報は「生存する個人に関する情報であって、当該情報に含まれる氏名、生年月日その他の記述等により特定の個人を識別することができるもの（他の情報と容易に照合することができ、それにより特定の個人を識別することができることとなるものを含む。）」と定義される（第 2 条）。このように個体識別が可能か否かを区分の基準とする例は外国法[4]でも見られる。いかに匿名化すれば個体識別が不可能かという問題は普遍的である。

ところがこの基本的な問題が蔑ろにされている。匿名化についての多くの研究は個体識別の危険性（開示リスク）の測り方は定める。しかし開示リスクの目標値について、せいぜい危険選好に応じてデータの分析価値（有用性）とバランスさせよ[5] としか言わない。このような態度は開示リスク測度の技術的研究には好都合である。しかし匿名データ等の場合、所与の環境において個体識別が不可能か判定したい。多くの研究はそのような要求に応えていない。

本論文では個体識別行為をモデル化し、個体識別が可能という状態をある種の開示リスクと関係づける。このようにモデル化すれば、評価される開示リスクの目標値が問題となる。この点について本論文は、個体識別が不可能な状態を過去の事例を基に決める方法を提案する。過去に公

[1]匿名化の正確な定義については星野 (2010) を見よ。

[2]本論文において統計法改正は手段として除外する。しかし個体識別性をデータ提供の基準とするのは必ずしも望ましくない。例えば個体識別されてもデータが悪用されなければよいという主張は妥当かもしれない。

[3]試行的提供の詳細については山口 (2008) を見よ。

[4]例えば U.S. Privacy Act など。U.S. Office of Federal Statistical Policy and Standards (1978, pp.3–5) の解説を見よ。

[5]例えば Duncan et al. (2001) や Domingo-Ferrer and Torra (2001) など。伊藤 (2012) のサーベイを参照のこと。

開されたデータについて個体識別が観測されていないとすれば、その事実は個体識別不可能ということについて情報を持っている。故に個体識別が観測されることの確率モデルを構成し、既公開の匿名データ等を匿名化の程度判断についての統計的証拠に転ずる。

このような理論なくして、明確な匿名性審査はあり得ない。そもそもチェックリストの記載事項は、個体識別と理論的に関係する要因であるべきだ。そしてリストの使い方も理論が定める。また本論文の理論は観測結果と関係を持ち、実証の対象である。いかなる理論も実証を経ることで継続的に改善される。従って本論文は、個体識別に関する統計的証拠—エビデンスに基づいた匿名化 (Evidence Based Anonymization, EBA) を主張する。

本論文の構成は以下の通りである。2 章は全体で、個体識別が可能か否かの判定方式を明らかにする。まず 2.1 節において、個体識別が可能という状態について考察する。次に 2.2 節ではこの考察に基づき、個体識別が可能という状態を、個体識別の容易度と解釈できる開示リスク測度を用いて判別するモデルを構築する。2.3 節ではこのモデルと観測可能な事実を確率モデルでつなぎ、個体識別の容易度の許容範囲が推定できることを示す。2.4 節では、我々のモデルの実務的含意として、匿名化の程度を決定する望ましいプロセスを示す。3 章では全体で、個体識別の容易度の評価手法を明らかにする。3.1 節ではいわゆるキー変数の選択方針を議論する。結果として、大半の攻撃者が知る公知の変数をキーとする。3.2 節ではその方針に基づいて、個体識別の容易度の評価手法を定める。3.3 節は具体例として、平成 15 年の住宅・土地統計調査の匿名データについて個体識別の容易度を評価する。4 章では開示リスクの定量的管理と定性的管理を組み合わせることが主張される。4.1 節では個体識別の容易度の比較可能性を議論する。4.2 節では匿名データの開示リスク管理で重要な役割を果たすチェックリストについて、改善点を指摘する。5 章はこれまでの議論の補遺である。

2　個体識別可能性の判定方式

本章では所与のデータについて個体識別が可能か否か、判定する方式を定める。そのためには個体識別が可能という状態を定義する必要があるので、2.1 節で議論する。2.2 節では個体識別が可能という状態を、個体識別の容易度が閾値を超えない状態とモデル化する。2.3 節ではこの閾値を観測から推定する。2.4 節では本章のモデルの実務的含意を考察する。

2.1　個体識別が可能とは

個体識別が可能な状態の定義について、コンセンサスは存在しない。何故なら考察すべき要因が多岐にわたるからだと思われる。しかし定義されないものは技術的に判定不可能である。従って本節では個体識別が可能な状態を考察し、後の定式化につなげる。

大まかに言って個体が識別された状態とは、公開されたファイルの一レコードが現実の特定個体の情報と分かることである。このような個体識別の試みを「攻撃」、これを試みるものを「攻撃者」と呼ぶ。

攻撃者が公開データを入手できなければ、明らかに個体識別は不可能である。実際、匿名データの利用が審査を伴うのは、それを目的としている。統計法の 33 条申請のように利用審査がより

厳しければ、攻撃者に公開データが渡る可能性は減るだろう。しかし利用審査で完全に攻撃者を排除できるとは限らない。以下の議論では、攻撃者が公開データを入手しうるのを前提[6]とする。

攻撃方法は大きく分けて「順攻撃」と「逆攻撃」に分類される（竹村, 1997）。前者は「照合 (matching)」とも呼ばれ、攻撃者が素性を知る個体を公開ファイルから探す。後者は公開ファイルから珍しい属性を持つ個体を選び、そのような個体を世の中で探す。

様々な実証研究などから、順攻撃（照合）の方が危険[7]と言われている。照合においては、特定個体について攻撃者が知る属性（「キー変数」）をフィールドとするレコードを並べたファイルを「攻撃用ファイル」と呼ぶ。照合とは、攻撃用ファイルの各レコードについて公開ファイルからキー変数が同等のレコードを探すことと言える。そのようなレコードが一意に存在するなら、キー変数の組は個体の識別子として働く。名前のような直接識別子と区別して、そのようなキー変数の組を Dalenius (1986) は「疑似識別子」と呼んだ。

ここで一意に照合されれば個体識別と言えるかは、自明ではない。例えば公開ファイルに含まれる個体群が母集団の一部としか攻撃者は分からないとしよう。この場合、照合結果が真に一意[8]であっても、公開ファイルに含まれていない母集団の他の個体が攻撃用レコードの個体かもしれない。故に攻撃用レコードの個体の属性を持つ個体が母集団で一意に定まる場合を「母集団一意」と呼び、そのような個体についてのみ識別されうると考えるのが一つの立場である。

なお母集団の部分である公開ファイルにおいて、特定の属性を持つ個体が一意の場合を「標本一意」と呼び、母集団一意と区別する。公開ファイルは統計調査の標本の一部であることが多く、標本一意の「標本」は必ずしも統計調査の標本ではないことに注意するべきである。

このように母集団と標本を区別して個体識別行為を具体的にモデル化した先行研究としては、英国国勢調査匿名化標本の開示リスクを評価した Marsh et al. (1991)、及びこの論文を再考した Dale and Elliot (2001) が挙げられる。Marsh らは個体識別を以下のような確率モデルで表した。

$$\Pr(\text{識別が実際に起きる}) = \Pr(\text{識別成功} \mid \text{識別を試みる}) \Pr(\text{識別を試みる}). \qquad (1)$$

そして識別を試みた時にそれが成功する事態は、以下の 4 つの条件が成立する場合だという。

(a) 攻撃用ファイルと公開ファイルのキー変数が同じ（時点や分類の）符号化基準で記録されている。

(b) 公開ファイルに個体が含まれている。

(c) 個体が母集団一意である。

(d) 個体が母集団一意と確証出来る。

これらの条件が満たされる事象をそれぞれ a から d と書けば

$$\Pr(\text{識別成功} \mid \text{識別を試みる}) = \Pr(a) \Pr(b|a) \Pr(c|a,b) \Pr(d|a,b,c) \qquad (2)$$

ということになる。

---

[6] リサーチデータセンターなどでの監視付きデータ利用なら、攻撃は不可能とみなしてよいかもしれない。

[7] 匿名データや個人情報の定義で照合を明示的に問題とするのは当然であろう。

[8] 公開ファイルが攪乱されている場合などで、間違って照合されるかもしれない。

Marsh らの議論では、母集団一意であることを確証しないと個体が識別されたと言わない。真に母集団一意であっても、攻撃者がそれを分からないことも多いからだ。例えば被調査者本人が自分を探す場合、全ての調査項目がキー変数となるため、その本人のレコードは母集団一意だろう。しかし自分以外の情報も知らなければ、母集団一意と確証できない。

照合の結果、母集団一意の可能性が高い個体について、攻撃者は追加情報を得て確証しようとするかもしれない。それは逆攻撃の一種である。母集団一意であることとその確証を区別すれば、順攻撃と逆攻撃を統合して分析可能となる。

ただ、母集団一意は確証されなくても、可能性が高ければ問題とする立場もあり得る。より一般的に考えて、あるレコードが特定個体の情報という推測が当たる確率が高ければ、問題かもしれない。例えば母集団 i 意なら $1/i$ の確率で特定の個体と当たると言える。

しかし確率的に当たる場合も個体が識別されたと見なせば、それを不可能にするには、全てのレコードを実在しない個体とするしかない。それは現実の匿名化では受け入れられない。つまり確率的に当たる場合も個体が識別されたと解釈すれば、個体識別が不可能な状態は事実上達成できない。そして個人情報保護法や統計法が求める匿名化は、存在しないということになる。そのように結論される法解釈は不適当ではないか。

本稿の目的は法の下での匿名化の決定である。故に我々は Marsh らのように、あるレコードが特定の個体の情報と確定する場合[9] のみ個体識別と解釈しよう。このような法的な安全と人々の安心は違う概念で、法的に十分な匿名化が社会的に十分とは限らない。しかし本稿では、法的に十分な匿名化の決定に議論を絞る。

では法的に識別が不可能という状態は Marsh らのモデルとどのように関係するだろうか。匿名化の達成目標については、「絶対的な匿名性」と「事実上の匿名性」の概念が知られている（濱砂 (2000) の解説を参照のこと）。絶対的な匿名性は、識別の可能性が疑う余地なく除かれていることを要求する。この状態は (2) 式がゼロ、すなわち

$$\Pr(\text{識別成功} \mid \text{識別を試みる}) = 0 \tag{3}$$

と考えられる。それから事実上の匿名性は、識別の費用が便益を上回る状態を指す。それは (2) 式が正でも、損得計算をして識別を試みないかもしれない攻撃者についての安全性を意味する。つまり事実上の匿名性概念は、(2) 式に識別を試みる確率を掛けた (1) 式の確率が低いことを要求している。

事実上の匿名性では個体識別が可能でも起きない状態が許容される。故に統計法や個人情報保護法における個体識別が不可能な状態は、絶対的な匿名性と読むのが妥当であろう。従って以下では、(3) 式が成立する状態を個体識別が不可能と考える。この場合 Marsh らのモデルでは、(2) 式の右辺各項のいずれかが 0 の状態と、個体識別が不可能な状態は同値である。

確率 $\Pr(a, b, c)$ の評価については 3.2 節で詳細に検討するが、$\Pr(a, b, c) = 0$ となるように匿名化すれば、有用性を著しく損なう。故に通常の匿名化事例では $\Pr(a, b, c)$ は正となる。従って、(2) 式の右辺を書き換えると

$$\Pr(\text{識別成功} \mid \text{識別を試みる}) = \Pr(a, b, c)\Pr(d|a, b, c) \tag{4}$$

[9] $i \geq 2$ について、母集団 i 意でもそのうちの $i-1$ 個体が結託すれば残り 1 個体が確定するが、それは問題にされていない。

なので、$\Pr(d|a,b,c)$ が 0 なら個体識別が不可能と考えられよう。つまり Marsh らの枠組みにおいて通常の場合、個体識別が可能か否かは $\Pr(d|a,b,c)$ が 0 か否かという問題に縮退する。

ではどのようにすれば事象 d、すなわち母集団一意の確証が起きるだろうか。Marsh らは母集団一意の確証手法として、「公衆の目」と全数名簿の利用を検討している[10] 。公衆の目とは、珍しくて目立つ個体が有名な場合を言う。例えば現職の首相が母集団一意ということは知られている。それから全数名簿の利用とは、職業人等の全数名簿で一意な個体は母集団でも一意なことを利用する。例えば日本の弁護士会は強制加入団体なので、その名簿は弁護士を全数含む。その名簿から、ある地域の女性若手弁護士が一人と分かるとしよう。この場合、母集団が日本人でキー変数が { 職業, 地域, 性別 } なら、その弁護士は母集団一意と確証できる。

名簿が実際に存在しなくても、特定条件を満たす全数の情報があれば母集団一意と確証できる。例えば日本人の母集団で職業が現職の首相という条件を満たす個体の全数を公衆が知っているから、母集団一意と確証できるのである。全数調査における低次元クロス集計表でセルの度数が 1 と分かるような場合も該当する。正確に言えば、あるレコードを母集団一意としている最小のキー変数の組み合わせ（Willenborg et al. (1995) によれば「指紋」）について、個体と紐付けされた全数情報があれば母集団一意を確証できる。

我々は母集団一意の確証が成功することを、母集団で特定条件を満たす個体の全数データとの照合が一意に成功することと考えよう。このような全数データを「確証用ファイル」と呼ぶことにする。攻撃用ファイルが全数のデータなら確証用ファイルにもなるが、両者は概念として区別した方がよい。前者が既存の情報であるのに対し、後者は追加情報も含みうるからだ。

全数データセットは、母集団の小さい部分なら入手は必ずしも難しくないことに注意するべきである。例えば改札口が一つしかない駅で限られた時間帯の入場者を全て観察することは可能で、その結果は部分集団の全数名簿である。その入場者中に女性が一人しかいないとしよう。この場合、母集団が電車利用者でも { 駅, 入場時間帯, 性別 } というキー変数についてこの女性が一意と確証できる。

それから、もし統計調査の標本の全数が公開されるなら、それは確証用ファイルになる。何故なら母集団において調査されたという条件を満たす個体の全数が、統計調査の標本である。ある個体が公開ファイルに含まれることを攻撃者は知っており、統計調査の標本の全数が公開されているとしよう。その個体の属性と公開ファイルを照合して一意ならその個体は標本一意だが、母集団一意と確証される。調査の有無がキー変数になっており、その個体と同属性の個体は、調査されていない個体群には存在しない。

情報学では、攻撃用ファイルと公開ファイルの個体群が同じという前提で開示リスク評価をすることが多い。それは調査の有無をキー変数に使って条件付きで議論しているということに他ならない。だからこそ、母集団一意と標本一意を区別しないのである。

このような事態は、全数調査の結果を全数公開する場合に近い。また標本調査においても、部分的には妥当である。調査区内で複数の標本が抽出される場合、標本に選ばれた個体は、同一調査区に他の標本が存在することが分かる。また調査区内で全数調査（集落抽出）なら、標本に選ばれた個体は他の標本が分かるし、そうでなくても調査員を尾行するなどして他の標本を一部でも突き止め、攻撃者となるかもしれない。このようなケースは現実性がないとは言えない。また

[10]他に母集団一意の確率を統計モデルで求めることを挙げているが、それでは母集団一意の確証にならない。Dale and Elliot (2001) による Marsh らの議論の再評価でも、統計的推測は母集団一意の確証として扱われていない。

そもそも被調査者は、自分が調査されていることを知っている。

従って標本を全数公開するのは避けなければならない。そのために標本の一部だけ公開する匿名化手法を「サンプリング」もしくは「リサンプリング」と呼び、匿名データにも施されている。ただしリサンプリングで公開ファイルから除外された個体が分かる攻撃者には、リサンプリングは期待した効果を持たない[11]。しかしそのようなことが分かるのは、匿名化処理の担当者くらいであろう。その担当者にしても、内部統制により、確率的にしか結果がわからないようにすることは可能だ。

統計当局は、内部統制や匿名化により、母集団一意とその確証を分離させるべきである。まずリサンプリングは必須と思われる。それから Marsh らも言うように、弁護士会会員名簿のような全数名簿の存在を知っているなら、それを利用した確証が出来ないように匿名化するべきである。例えば公表データの職業を「弁護士または司法書士」と再符号化しよう。これにより弁護士会会員名簿は、開示リスク評価上の全数名簿でなくなる。それからレコードを公開ファイルから削除する手法もよく使われる。

では攻撃用ファイルが全数でないとして、確証用ファイル、つまり母集団一意を確証したいレコードを含むような全数名簿をいかに作るか。それには全数の外枠の中をしらみつぶしに調査することになる。調査項目は（母集団一意を確証したいレコードが所与で）、キー変数全てではなく指紋だけでよい。ただししらみつぶしのためには、その外枠がある程度狭い必要がある。外枠を狭めるために属性の条件を用いるのだが、うまく選ばないと外枠は母集団になってしまう。例えば身長 170cm という属性条件で、しらみつぶしの範囲は絞れない。

全数名簿が作れるような外枠を絞る属性条件としては、所属団体や居住地域、職場などが挙げられる。これらは広義の地理情報と考えられ、そもそもその範囲の全数名簿は存在することが多い。狭義の地理情報が一意を生みやすい強力なキー変数であることは良く知られているが、全数名簿を作らせないという意味で、広義の地理情報の精度を匿名化で管理する必要がある。

このように匿名化の過程を管理すれば、母集団一意の確証を妨げることができる。ではそのような管理に依存して $\Pr(d|a,b,c)$ を評価し、個体識別が不可能、すなわち $\Pr(d|a,b,c)=0$ と言えるだろうか。

これまでの議論を振り返れば、母集団一意が確証できる確率 $\Pr(d|a,b,c)$ や母集団一意数は、攻撃者の能力や知識、情報に依存している。従って攻撃者についての想定を「シナリオ」と呼び、シナリオ依存で個体識別の可能性を議論することが多い（例えば Paas (1988) など）。

実務で個体識別が可能か判断する際も、シナリオは必要である。ただ実務を規定する法に基づいてシナリオを構成したいところだが、個人情報保護法や統計法には、個体識別が可能ということが誰にとっての問題か書かれていない。つまり攻撃者が誰か、法解釈の問題が存在する。

以下の議論では、攻撃者は実在する人間と考える。研究レベルでは想定の現実性は必ずしも問題にならないので、完全な情報を持つ理想的攻撃者が想定されることも多い。しかしそのような想定の下では、匿名化を強く施さなければならない。この場合に有用なデータは提供しづらくなる。一方、攻撃者が弱ければ有用なデータとなるが、現実に個体識別が起きては問題である。従ってデータの有用性を確保しながら情報を現実的に保護するには、実在という条件で攻撃者の能力を限るのが望ましい。このような限定は法の主旨に反しないはずだ。

[11]本質的にはそのような理由で、Chaudhuri and Mishura (2006) はリサンプリングが差分プライバシーを達成しないと議論する。

実在が基準ならば、シナリオは現実の制度をモデル化して構成することになる。例えば匿名化を施す（統計当局の）人間を潜在的な攻撃者とみなすか否かで、匿名化の結論は大きく異なる。この点について我々のシナリオ構成では、内部統制が適切ならば匿名化を施す人間は攻撃者とみなさない、と考える。一般に契約や研修などの制度も、個体識別の可能性管理に使われている。従って匿名化の実務的判断において、現実の制度は無視するべきではない。

では実在を基準として、$\Pr(d|a,b,c)$ を評価できるだろうか。その評価のためには実在の攻撃者の能力を知る必要がある。しかしそれの調査はコスト制約の下で限界がある。従って $\Pr(d|a,b,c)$ の確定的評価[12]は現実には難しい。

従って我々は攻撃者についての情報が限られているという前提で、個体識別が不可能となる匿名化水準を現実的に定める方法を考察する。基本的なアイデアは、未知の攻撃者の能力を統計的に推定するということである。制度に影響される攻撃者の能力は現実を観測した結果に反映されているはずなので、観測可能な事実は利用する。それから我々は問題を $\Pr(d|a,b,c)=0$ か否かの判断に絞る。それは $\Pr(d|a,b,c)$ の評価より易しいはずである。

2.2 個体識別可能性の判別モデル

本節では母集団一意の確証が可能か、すなわち $\Pr(d|a,b,c)=0$ か否かの判別モデルを構成する。前節の議論によると、条件 (a,b,c) を満たすレコードについて確証用ファイルとの照合が成功することと、母集団一意の確証は同値と定義されている。従って $\Pr(d|a,b,c)=0$ か否かの判断は、そのような確証用ファイルが用意できるか否かを判断すればよい。

確証用ファイルとは、指紋に関する全数データであった。その存在は不確実だが、定性的に考えて、指紋が多ければ適切な確証用ファイルを用意出来る可能性は上がるのではないか。そして条件 (a,b,c) を満たす公開データが多ければ、照合されうる指紋も増えるはずだ。そもそも確証能力は一種の情報収集能力だから、キー変数を多く持ち母集団一意を多く得る攻撃者は、確証能力も高いと考えられる。故に条件 (a,b,c) を満たす公開データの量を基準化した $\Pr(a,b,c)$ が増加すれば、母集団一意の確証は単調に易しくなると考える。

我々はこの単調性から、$\Pr(a,b,c)$ が適当な閾値を下回ることと母集団一意の確証が不可能なことが同値と考えよう。すなわち適当な非負の β について

$$\Pr(a,b,c) \leq \beta \Leftrightarrow \Pr(d|a,b,c)=0 \tag{5}$$

と考える。個体識別が不可能という状態は、$\Pr(a,b,c)=0$ または $\Pr(d|a,b,c)=0$ なので、(5) 式が成立していれば個体識別が不可能である。

ここで $\Pr(a,b,c)$ は、母集団一意の確証の「容易度」と解釈できる。基本的に $\Pr(a,b,c)$ は母集団一意数に比例するので、その挙動は母集団一意数のそれが支配的である。

母集団一意数は局所的匿名化[13] において無意味な場合があるし、全ての母集団一意数を等しく

---

[12]Marsh らは $\Pr(d|a,b,c)$ について「非常に多くのキー変数について事前情報が無いはずなのでゼロと信ずるが 0.001 と仮定」している。これでは評価が出来ているとは言えない。

[13]一律の基準で全レコードを匿名化するのではなく、レコード毎に匿名化手法を変える場合を「局所的」と呼ぶ。例えば De Waal and Willenborg (1994) は指紋の位置に削除を施している。一般には Hoshino (2009, Section 6) で議論したとおり、匿名化された表現が互いに排他的でないと母集団一意は照合成功の条件にならない。ただし互いに排他的でない匿名化は余り使われない。

危険と考えることには異論[14]もある。例えば Elliot et al. (1998) は、より低次元の周辺分割表で一意になるレコードの方が危険とみなし、そのようなレコードを「スペシャル・ユニーク」と呼ぶ。スペシャル・ユニークは高次元の周辺分割表でも一意なので、高次元で見れば同じ一意でも差があると考えられている。

それでも母集団一意数の利用は積極的に支持される。何故なら、まず母集団一意概念は非専門家でも理解が可能で有名である。従って匿名化に関する説明責任を果たす上で、有用なツールとなる。それから我々の議論においては、以下で説明される匿名化の程度についての単調性を持つことが重要な理由となる。

主に使われる匿名化手法は「再符号化[15]」と呼ばれ、個体属性をより粗く分類する。再符号化を分割表への操作と見れば、セルの併合と同等である。母集団でも標本でも一意数は分割表で度数 1 のセル数なので、セルの併合について一意数は非増加である。つまり一意数は、再符号化について単調に変化する。

従って再符号化を追加すれば、母集団一意数に比例する $\mathrm{Pr}(a,b,c)$ は増えることがない。故にモデル (5) の下で、個体識別が不可能なデータに再符号化を追加して、個体識別が可能と判定されることはない。

再符号化が追加されたデータは、元のデータより直感的には匿名性が高いはずだ。このような匿名性の高低と、$\mathrm{Pr}(a,b,c)$ の大小関係は矛盾しない。何らかの匿名性の測度を「個票開示リスク測度」と呼ぶが、$\mathrm{Pr}(a,b,c)$ はそのような測度として望ましい性質を持っている。実は我々は最終的に $\mathrm{Pr}(a,b,c)$ を事例間で相対比較するので、この美点は貴重である。

なおモデル (5) による判定は、必ずしも母集団二意以下を無視するわけではない。二意以下の数は一意数に連動するので、通常は $i \geq 0$ について標本 i 意のレコード数を用いて母集団一意数を推定する。この場合、標本 i 意が持つ母集団 i 意の情報は無視していない。この点は情報学で有名な k-匿名性 (Sweeney, 2002) 基準が、$i \geq k$ について標本 i 意のレコード数を無視するのと異なる。

さて、個票開示リスク測度の理論では、匿名化したデータについてその測度の計量値が適当な範囲に収まるなら、そのデータを公開可能と判断するのが普通である。モデル (5) に基づく判定方式は、個票開示リスク測度として $\mathrm{Pr}(a,b,c)$ を用いる場合の意思決定と同じことになる。

問題は $\mathrm{Pr}(a,b,c)$ の閾値 β の決定である。多くの個票開示リスクの研究では、測度の許容範囲つまり閾値の決定は、主観の問題とみなされる。そのため匿名化の実務では、判断の責任を統計委員会のような権威に負わせるのが普通である。しかしそのような権威の判断は、透明に説明出来るのが望ましい。また匿名化の可否判断に何らかの権威が必要なら、権威にアクセス出来ない人々は匿名化を利用できないことになる。個票開示リスク測度の許容範囲を、客観的に定める方法はないだろうか。

我々は $\mathrm{Pr}(a,b,c)$ の閾値 β を統計的に推定しよう。統計的推定である以上誤差を伴い、誤差に関する信頼係数の選択に主観的判断が残る。しかし信頼係数の社会的管理は容易だろう。次節では、統計的推定に必要な個体識別の観測モデルを構成する。

[14]そのように考えてレコード毎に測られる開示リスクを「レコードレベルリスク」と呼ぶ。一方、母集団一意数はファイルの性質であり「ファイルレベルリスク」を表す。
[15]情報学では「一般化」と呼ばれる。トップコーディングや削除 (suppression) も再符号化の特殊ケースである。

2.3 個体識別の観測モデル

本節ではモデル (5) について、個体識別が可能となる閾値 β を観測と関係づける。個体識別が可能かどうかは直接観測できないので、観測可能な事象をモデル化しなければならない。

個体識別に関する観測モデルを構成するとして、個体識別が実際に起きても必ずしも観測されないことに注意するべきである。攻撃者は個体識別に成功したとしても、隠れているかもしれないからだ。

攻撃者が真に識別を成功させた場合、その事実を公表して得られる利益と、識別を隠して得る利益がある。まず識別成功を公表した場合、攻撃者は有名になるだろう。そして統計当局の面目が失われることに魅力を感じるかもしれない。しかし識別された個体は情報の漏洩を知ることになり、識別によって入手した情報を用いた詐欺、ストーキング等は難しくなる。そのように識別で得た情報を実用するには、識別成功は公表しない方がよい。また識別成功を公表すれば法的、社会的制裁の対象[16] になるかもしれない。故に例えば商業目的なら識別成功を公表せず、攻撃者は精度の良いマーケティングの利益を享受するだろう。

従って個体識別に成功した攻撃者が隠れる可能性を想定し、個体識別の発生とその社会的認知は区別する。我々は個体識別発生の社会的認知有りの場合、確率変数 $X = 1$ とし、認知なしの場合 $X = 0$ とする。つまり

$$\Pr(X = 1) = \Pr(\text{識別の社会的認知} \mid \text{個体識別が実際に起きる}) \Pr(\text{個体識別が実際に起きる})$$

と考えよう。さらに (1) 式と (2) 式より

$$\begin{aligned} \Pr(X = 1) &= \Pr(\text{識別の社会的認知} \mid \text{個体識別が実際に起きる}) \\ &\quad \times \Pr(a, b, c, d) \Pr(\text{識別を試みる}) \end{aligned} \tag{6}$$

と書ける。

モデル (6) の下でモデル (5) を用いるとして、未知母数 β と X の関係を見よう。まず個票開示リスク測度 $\Pr(a, b, c)$ の評価値を γ で表す。この時もし条件

$$\Pr(\text{識別の社会的認知} \mid \text{個体識別が実際に起きる}) > 0 \tag{7}$$

と条件

$$\Pr(\text{識別を試みる}) > 0 \tag{8}$$

の両方が成立するなら、γ に依存する正の確率 $p(\gamma)$ について

$$\Pr(X = 1) = \begin{cases} p(\gamma) & \gamma > \beta \text{ の場合} \\ 0 & \gamma \leq \beta \text{ の場合} \end{cases} \tag{9}$$

[16] 匿名データの利用者については、統計法第 43 条第 2 項に「当該匿名データをその提供を受けた目的以外の目的のために自ら利用し、又は提供してはならない」とある。個体識別の成功を公表することは（識別目的でのデータ提供は行われないので）本条項に違反するが、直ちに罰則が適用されるわけではない。匿名データの利用者についての罰則は「匿名データを、自己又は第三者の不正な利益を図る目的で提供し、又は盗用した者」に対して「五十万円以下の罰金に処する」(61 条 3 項）とだけ定められている。例えば匿名データの不備を指摘するための個体識別の公表は不正な利益を図る目的と必ずしも言えないので、罰則は適用できないのではないか。なお 33 条申請によって調査票情報を手に入れた者が個人又は法人の秘密を漏らした場合は「二年以下の懲役又は百万円以下の罰金」(57 条 2 項 3 号)、自己又は第三者の不正な利益を図る目的で提供又は盗用した場合は「一年以下の懲役又は五十万円以下の罰金」(59 条 2 項)、と罰に差がつけられている。ところが匿名データの利用者が（個体識別によって入手した）秘密を漏らした場合の罰則規定はなく、そのような事態を統計法は想定していないように思われる。

あるいは

$$\Pr(X=0)=\begin{cases}1-p(\gamma) & \gamma>\beta \text{ の場合}\\ 1 & \gamma\le\beta \text{ の場合}\end{cases} \tag{10}$$

となる。ここで

$$p(\gamma)=\gamma\Pr(d|a,b,c)\Pr(\text{識別の社会的認知}\mid\text{個体識別が実際に起きる})\Pr(\text{識別を試みる}) \tag{11}$$

だが、以下の母数推定の議論は $p(\gamma)$ の具体形に依存しない。また $\Pr(a,b,c)$ 以外の個票開示リスク測度を用いても成立する。つまりモデル (9) を用いて個票開示リスクの許容範囲を推定する方法は、汎用性を持つ。

まず条件 (7) は妥当と思われる。先に検討したように、攻撃者は識別成功を公表する動機がある。識別成功が公表されて社会的に認知される可能性は 0 ではないはずだ。また個体識別の成功で得た情報を隠れて実用するような場合でも、それが発覚すれば社会的に識別成功は認知される。発覚の可能性は 0 ではないだろう。

なお公開データが隠れて実用できるような情報を含まなければ、攻撃動機は識別成功の公表による利益のみになる。そのような場合、確率 Pr(識別の社会的認知 | 個体識別が実際に起きる) は上がるはずだ。従って匿名化によって、この確率は変えられる。例えば病歴という情報は、削除したり罹患時期を区間表示したりすることで、実用困難にできる。

次に条件 (8) も妥当と考える。識別を試みるか否かは、識別成功の損得や容易性に依存すると考えられる。先に検討したとおり、公開データが実用可能な情報を含まなくても、攻撃に成功してそれを公表することの利益は存在する。そのような動機が考えられる以上、識別を試みる確率は正のはずだ。

なお Marsh et al. (1991) が指摘するように、Pr(識別成功 | 識別を試みる) $=\Pr(a,b,c,d)$ の減少は Pr(識別を試みる) を減少させるだろう。Marsh らは結局「識別を試みた例を知らないので識別を試みる確率の最良の推定値は経験からゼロ」と述べているが、無知は根拠にならない。他に Elliot et al. (2010) は、もっともらしい攻撃シナリオの考察こそが、Pr(識別を試みる) の妥当なモデル化につながると主張している。

さて、これまでの考察から条件 (7) と (8) が成立すると考える。故にモデル (9) あるいは (10) を採用し、このモデルの下で母数 β を最尤推定する。匿名化したデータが公開された n 件の事例を、共通のモデル (9) からの標本とみなそう。この時 $i, i=1,2,\ldots,n,$ 番目について個票開示リスクの評価値 γ_i と個体識別発生認知の有無 x_i は少なくとも観測される。単純化のため $\gamma_1>\gamma_2>\cdots>\gamma_n$ としよう。

この中で個体識別発生が認知された（$x_i=1$ となる i が存在する）場合は $\gamma_i\le\beta$ の尤度 $\ell(\beta)$ はゼロである。故に例えば $x_n=1$ なら、最尤推定量は $\hat\beta<\gamma_n$ となる。では適当な $m\le n$ について $x_{m-1}=1, x_m=x_{m+1}=\cdots=x_n=0$ の場合はどうか。この時 $\beta\ge\gamma_{m-1}$ なら $\ell(\beta)=0$、$\gamma_{m-1}>\beta\ge\gamma_m$ なら $\ell(\beta)\propto p(\gamma_{m-1})$、$i\ge m$ について $\gamma_i>\beta\ge\gamma_{i+1}$ の時 $\ell(\beta)\propto p(\gamma_{m-1})\prod_{j=m}^{i}(1-p(\gamma_j))$ である。ここで $p(\gamma)$ は正なので、最尤推定量は $\gamma_{m-1}>\hat\beta\ge\gamma_m$ となる。そしてもし個体識別発生が全く認知されていなければ、最尤推定量は $\hat\beta\ge\gamma_1$ である。

つまり過去に個体識別発生が認知されていない事例[17] の中で個票開示リスクの最も高い評価値

[17] データの公開直後に個体識別発生の認知が起きなくても、ある程度後で認知が起きることはあり得る。閾値の推定をする時点に依存して各 $p(\gamma_i)$ は変化するかもしれない。しかし $0<p(\gamma_i)$ なら最尤推定量は変わらない。

を γ^* と書けば、閾値 β は γ^* 以上（かつ個体識別発生が認知されている事例の評価値未満）と最尤推定される。なお個票の公開で先行する海外でも、個体識別が発生した事例[18] はほとんど認知されていない。識別が起きないように匿名化しているからであろう。

閾値 β の最尤推定量 $\hat{\beta}$ の精度を評価しておこう。推定値が過大ならば、個体識別が可能なデータを不可能と判断する誤りが起きる。過小の場合は逆である。前者は情報漏洩による被害につながるのに対し、後者はデータの有用性を損なう。統計当局の意思決定問題としては、前者の損失の方が重い。従って過大推定（$\hat{\beta} > \beta$）の確率を求めよう。個体識別が可能にも関わらず認知されない（$X = 0$）ことが誤りを引き起こすことに注意すれば、もし $\gamma_1 > \beta \geq \gamma_2$ ならこの確率は $1 - p(\gamma_1)$ である。もし $\gamma_2 > \beta \geq \gamma_3$ なら、この確率は $(1 - p(\gamma_1))(1 - p(\gamma_2))$ となる。同様に考えて、一般に過大推定の確率は、$p(\cdot)$ が 0 に近いほど高い。それから真の β より γ が高い事例が少ないほど過大推定の確率は高い。

当然だが、誤りを減らすには実績を積めばよい。統計当局は過大推定の確率を減らすために、公開するデータの個票開示リスク γ を決められる。モデル (9) からの新しい標本（新規公開するデータ）の個票開示リスクを γ^* と同じにすれば、個体識別が認知されないとして、$\hat{\beta} > \beta$ の確率は増えない。もし同じでなければ、β の真の位置によっては過大推定の確率は増えないとは言えない。

なお $p(\cdot)$ を増やしても過大推定の確率は減る。しかし (11) 式を見ると、Pr(識別の社会的認知 | 個体識別が実際に起きる) を上げる以外の方法は必ずしも社会的に望ましくない。

これまでの議論で、観測に基づいて個票開示リスクの許容範囲を定めることができた。しかし重要な問題が残っていて、それは閾値 β が共通する事例の範囲である。

例えば英国国勢調査匿名化標本と日本の国勢調査匿名データでは、閾値 β が違うのではないか。もう一例挙げよう。日本法では制度化されていない一般目的汎用ファイル (Public Use File, PUF) が仮に作られるとして、PUF の閾値と匿名データの閾値は異なる[19] のではないか。このような制度によって閾値 β が変わることを許すには、制度所与の条件付きで β を推定すればよい。そこでは β の推定値の差が制度効果の差であり、制度が β を変動させていると考えられる。

同様に制度以外の個体識別行為の要因でも、$\Pr(a, b, c)$ の計量に反映されなければ閾値 β を変動させると考えられる。故にどの範囲の要因所与で β の条件付き推測をするかは、$\Pr(a, b, c)$ の計量方法を定めないと決められない。従って $\Pr(a, b, c)$ の計量方法は章を改めて考察し、その後で閾値 β が共通する事例の範囲を定める。

次節ではこれまでの議論を振り返り、開示リスク管理実務について我々のモデルが示唆することを述べる。

2.4　個体識別モデルの実務的含意

前節までの理論的考察から、匿名化実務では個体識別の容易度 $\Pr(a, b, c)$ を γ^* とすることが推奨される。そのようにすれば、過去の事例が個体識別が不可能なことの統計的証拠になり、新規公開するデータも今後の事例の統計的証拠となる。

---

[18] Sweeney (2002) による個体識別の成功は有名である。

[19] 星野 (2010) の考察の通り、PUF は匿名データよりも強い匿名化を必要とする。しかし個体識別が可能か不可能かという統計法の二分法では、PUF と匿名データを区別できない。ところが個体識別の容易度という概念を用いれば、個体識別が不可能という状態の中で匿名データと PUF を区別出来る。

このように開示リスクを管理する方法は、匿名化手法ではなく開示リスクについての前例踏襲と言える。1 節で確認したように、現行のガイドラインは別の事例を参考に匿名化せよと定めている。開示リスクについての前例踏襲はガイドラインの方針と軌を一にしており、実務的に受容しやすいのではないか。

ただし開示リスクについて前例踏襲をする場合、匿名化の緩和判断が問題になるかもしれない。我々の基準では、匿名化を緩和しても開示リスクが γ^* を上回らなければ、そのような緩和は受容される。しかし真の β は γ^* より大きいかもしれず、匿名化は更に緩和できるかもしれない。ところがこれまでのモデル分析からは、開示リスクの許容範囲を広げる論理は得られない。過去の事例がなければ安全性の証拠もないからである。

では（$n = 0$ の場合も含めて）過去の事例がなければどうしたらよいか。情報を補うため β の事前分布、モデルを用いることは考えられる。しかし我々は β の真の位置を余り知らないように思われる。故にモデルを用いた判断は将来的な課題とし、当面は観測情報の増加を工夫する方が健全である。

例えば公開を予定する匿名化したデータへの攻撃実験は、仮想的なデータ公開事例となろう。ただし実験では、社会に隠れている潜在的な攻撃者の能力は分からない。

それから観測する量 X を個体識別発生の有無の二値と限らないことも、β の位置を定める参考となる。例えば個体を識別できたと誰かが誤って主張することを観測すれば、個体識別を試す気にさえならない水準より開示リスクが上がっていることが分かる。一般に 1 件の重大事故の陰には 300 件のヒヤリ・ハットが起きているという。データ提供への社会的反応が、データの開示リスクと閾値の距離を示唆するはずだ。

このような考え方は、治験薬の用量反応関係の推測で実績がある。薬の臨床試験では、人体に決定的な悪影響を及ぼしてはならない。このような制約下では、動物実験で薬の人体への影響を仮想的に確かめる。その後で人体へ薬の投与を少量から始めて徐々に増やし、危険な兆候が見られれば中止する。そこでは生死の二値ではなく、投与量が死亡の閾値と近いことを示す兆候（心拍や呼吸の異常等）を観測するのである。

3　個票開示リスクの計量

前章ではモデル分析に基づいて、個体識別の容易度 $\Pr(a, b, c)$ を事例間で比較する管理手法を提案した。しかし $\Pr(a, b, c)$ の具体的計測方法は定めなかったので、本章で議論する。3.1 節では、母集団一意数を評価する際のキー変数の選択方針を考察する。3.2 節ではその方針に基づき、$\Pr(a, b, c)$ の計算手法を明らかにする。3.3 節は現実の匿名データを例にとり、計算手法を実証する。

3.1　キー変数の選択方針

個体識別の容易度 $\Pr(a, b, c)$ の計測において最大の問題は、攻撃者が照合に使う個体属性、すなわちキー変数の選択である。我々は攻撃者が実在することをシナリオの選択基準としているが、実在しうる攻撃者像は多様である。例えば全ての変数について匿名化前の元データを知る最強の攻撃者かもしれないし、たまたま隣人の落とした給与明細を見た攻撃者かもしれない。どのような攻撃者を想定してキー変数を選べばよいだろうか。

キー変数の選択方法に関する既存研究で挙げられるのは、Elliot et al. (2010, 2011) くらいである。Elliot らは社会で利用可能な個体情報の網羅的調査が必要と述べている。具体的には (i) アクセス制限付きデータベースの調査項目、(ii) 公になっている個体データの形態、(iii) ネットショッピング等での web 上データ収集項目、(iv) 商業データベースの情報、(v) 個人情報の収集実験結果、(vi) 情報保有組織における個体情報の取り扱い慣行、(vii) ソーシャルネットワーキングサービス (SNS) での個体データの形態、の調査が提案されている。そして複数のソースにまたがって存在する個体情報について、符号化基準の違いを調整して体系的に統合したデータベースの作成が主張される。つまりこのようなデータベースを参照すれば、キー変数の種類や符号化基準の一致する程度が分かると考えられている。また Elliot らは、個体情報の入手が可能となる形態の分類[20] も将来課題として議論している。この入手形態の知識があれば、強い攻撃者から弱い攻撃者まで、多様な攻撃形態を想定しやすい。その上で Elliot らは多様な開示リスク評価の必要性を指摘している。

Elliot らの言うように、社会で利用可能な個体情報が分からないと現実的な攻撃者は想定できない。しかし例えば SNS で公開する個体情報などは常に変化するので、法外なコストをかけても攻撃に利用できる個体情報を完全には調査できない。故に Elliot らの主張するデータベースは不完全にならざるをえない。従って現実的な攻撃者を想定してキー変数を選択する方法は不確実性を伴う。だからこそ多様な開示リスク評価が必要と主張されるわけだが、どの程度多様ならよいのか。Elliot らの定性的議論では分からない。Fung et al. (2010) によれば、キー変数の選択方法は「未解決問題」である。

データが k 個の変数からなるとして、そこからキー変数を選ぶ方法は 2^k 通りある。もし 2^k 通りの全ての場合について、「多様な」リスク管理ができれば、キー変数の選択の問題は生じない。しかし k が多い場合にそれは現実的ではない。そして公的統計は k が多い。故に我々は 2^k の場合からリスク評価の対象を選択せざるをえないとして、その方針を考察しよう。

仮に 2^k の全ての場合について、母集団一意数が得られるとする。それらを順序データ $(u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(2^k)})$ として表す。母集団一意数 $u_{(i)}$ を i の関数と見てできる曲線を「(母集団) 一意数曲線」と呼ぼう。キー変数の選択は、一意数を評価する順位 i を決定する問題に他ならない。

例えば平成 15 年の住宅・土地統計調査の匿名データで説明しよう。ただしこのデータは変数を 170 ほど含むので、11 変数（都道府県、住宅以外の建物の種類、住宅以外の建物の所有関係、建物の構造、建物の階数（うち一戸建て・長屋、うち共同住宅）、むねの建築時期、建築面積、敷地面積、エレベータの有無、高齢者対応か）しかないとする。そして $2^{11} = 2048$ 通りの変数の組み合わせについて、j 番目の組み合わせの標本一意数を u_j とする。本来 u_j は母集団一意数だが、計算が簡単なので標本一意数を用いる。その順序データの順位 i を横軸にとり、縦軸に $u_{(i)}$ をとってプロットした結果が図 1 の “11 vars” である。

このような一意数曲線は再符号化によって単調に下方へシフトすることを確認しておこう。例えば図 1 の例で都道府県コードを匿名化によって削除（都道府県コードが全て同じとみなす）した場合の一意数曲線が “ex region”、同様にむねの建築時期を削除した場合が “ex date” である。都道府県コードという強力なキー変数を削除する方が、曲線を大きくシフトさせていることが確認

[20]他者によって個体情報が非自発的に収集される場合、自発的参加で個体情報が他者により収集される場合、それから個体情報を自発的に公開する場合の三分類が例示されている。他に外観から分かる情報もある事や、情報取得に必要な資源の多寡も重要な要素という事などにふれている。

できる。

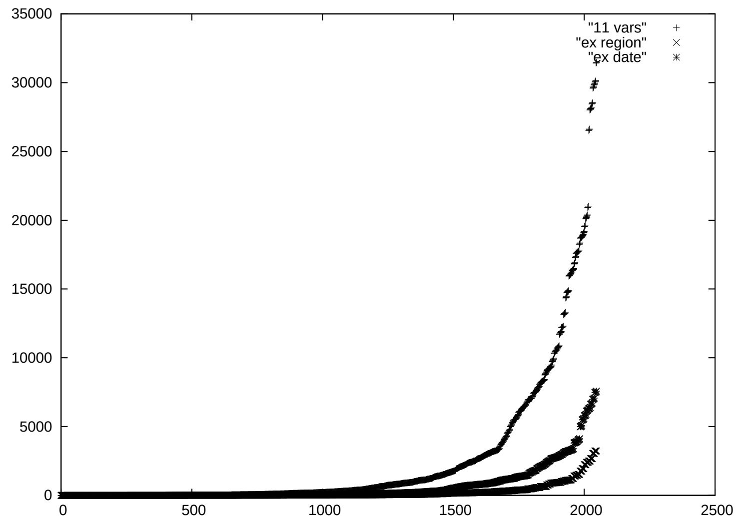


図 1: 標本一意数（$u_{(i)}$）と順位（i）の関係

なお一意数を計算する際にキー変数の区分（カテゴリー分け）にも選択の余地があるなら、一意数を評価する順位は非整数もありうることになる。一意数を計算するための各変数の区分は、データ公表に用いる区分と一致させるのが常識的である。ただ攻撃用情報が公開データより粗ければ、公表データの区分方法で算出した一意数は、攻撃者にとっての一意数にならない。例えば攻撃者が 8 地方区分のデータしかもっていなければ、都道府県別でデータが公表されても地方区分内で識別できない。このように考えて変数の区分を攻撃者の知識に合わせるなら、キー変数の選択候補は 2^k 個だけでなく、これらの順位間の内点も候補ということになる。何故なら、ある変数を最も粗く単一の分類とすることは、キー変数として使わないことと同値だからだ。従って理論上は、一意数を評価する順位は区間 $[0, 2^k]$ 内の実数ということになる。しかし我々は整数以外の順位を無視し、一意数を計算するための変数の区分を公表用の区分に一致させる。以下の議論での推定誤差を考えれば、手間をかけて内点を検討する価値はない。

さて、一般に順位 r を選択した時、それより小さい順位 $i \leq r$ について $u_{(i)} \leq u_{(r)}$ である。従って最大の順位を選択し、全ての変数をキーとして母集団一意を評価すれば、全ての（より部分的な情報しか持たない）攻撃者にとっての母集団一意はそれより小さい。このように全ての変数をキーとして使えるような攻撃者を想定するミニマクス基準を採用している研究はよく見かける。しかし問題は、そのような攻撃者が存在しないかもしれないことである。存在しない攻撃者の能力は、観測が証拠にならない。

例えば今回公開して個体識別が社会的に認知されていないデータがあるとしよう。このデータについて、全ての変数をキーとして計測した母集団一意数が c とする。ところが実は一部の変数しかキーに使えない攻撃者しか存在せず、その攻撃者の能力に合わせて計測した母集団一意数が d とする。そして全ての変数をキーとして、次回公開するデータの母集団一意数を c 以下にするとしよう。しかし件の攻撃者にとって次回公開するデータの母集団一意数 d' は、c 以下だが d 以下とは限らない。故に他の要因が一定でも、個体識別は可能となるかもしれない。

このように考えると、現実に存在する最強の攻撃者を想定してキー変数を選択するという指針が得られる。ここで「最強」とは、所与のデータについて母集団一意数を最も多く得られるという意味である。そしてその攻撃者にとってのキー変数が前提で母集団一意数を管理すれば全攻撃者について母集団一意数が管理され、毎回のデータ公開事例が個体識別可能性の統計的証拠となる。

キー変数の選択問題は理論的にはこのように解決される。この方針は統計当局にとって最悪の場合を現実の範囲で管理する点が特徴である。考えてみれば当然だが、計測方法は目的にふさわしく定めるべきである。キー変数を選択する上記の方法は、エビデンスに基づいた匿名化という体系の一部になっている。

しかし現実に存在する最悪の攻撃者の能力は未知で、実際には推定しなければならない。キー変数の組み合わせに関する母集団一意数の順位 $r \in (1, 2, \ldots, 2^k)$ 上で攻撃者の分布を考えれば、分布関数 $F(r)$ が 1 となる最小の順位を推定するということになる。

なおこの攻撃者分布の構成では、全ての攻撃者は手持ちの情報の範囲で母集団一意数を最も多く得られる順位に所属すると考える。ただし実際には、ある変数の集合をキーとして使える攻撃者はその部分集合もキーとして使える。従ってある順位のキー変数に対して攻撃をかけられる者の厳密な数は、攻撃者分布だけでは分からない。傾向としては順位が小さい方がキー変数が少ないので攻撃者の数は多く、順位が大きいほど攻撃者の数は減ると考えられる。

スペシャル・ユニーク概念が少ないキー変数で一意になるレコードの方を危険と考えるのは、より多くの攻撃者の標的になるのが理由と解釈できる。それからガイドラインは一変数の周辺分布の裾に再符号化をほどこす目安を与えているが、一変数を利用できる攻撃者は多いという意味で妥当である。

攻撃者分布の最大順位の推定について、Elliot らが主張するような個体情報の調査をすればデータが得られるだろう。しかし例えば津波の最大波高の推定と同じで、極値の統計理論はあるが、データが有っても最大値の推定は難しい。推定精度が悪い最大順位における $\mathrm{Pr}(a, b, c)$ を管理対象とすれば、個体識別が可能となる閾値 β の推定精度にも悪影響を及ぼす。

より実行可能なリスク管理の対象は、分位点である。分位点の推定は最大値の推定に比べて易しい。例えば金融規制のヴァリュー・アット・リスク (VaR) という考え方では、損失額の最大値ではなく損失額分布の 99 パーセント分位点に見合う資本の蓄積が要求される。同様に、キー変数の組み合わせに関する母集団一意数の順位 $(1, 2, \ldots, 2^k)$ 上の攻撃者の分布における適当な分位点を、管理対象としたい。

この管理対象が α パーセント分位点なら、他のリスク要因が一定として定量的に開示リスクが管理される攻撃者の割合が α パーセントということである。従って α は大きい方が望ましい。しかし攻撃者の分布を定量的に評価するのは困難である。実務上は定性的な方針で「大半」に相当する分位点を選ぶしかないだろう。

結局、開示リスクを管理する分位点は、「公知」の個体情報をキー変数とする場合がよい。公知とは多くの人が知っているということであり、（潜在的な）攻撃者の多さを意味する。Elliot らが主張するような網羅的調査でようやく存在が分かる個体情報は、知る人が少ないので公知とは考えない。また、個人的に知りうる詳細な個体情報も、公知とは考えない。たまたま泊めてもらって友人の詳細な生活時間を知ったような攻撃者は少ないはずだ。

それから実は、個体が調査されたか否かをキー変数として扱うか決めなければならない。2.1 節で議論したように、それがキー変数として使えると、母集団一意の確証はかなり易化する。しか

し個体が調査されたか否かを知る攻撃者は少ないはずで、その情報は公知とは考えない[21] ことにしよう。

以上の方針によれば、管理対象の順位に攻撃者が存在しない可能性を減らすこともできる。しかし分位点によるリスク管理の最大の問題は、分位点より外に管理されない攻撃者が残ることである。ただ、そのような攻撃者が常に危険というわけではない。ある分位点でのリスクを再符号化で目標水準まで下げたとしよう。再符号化は全ての順位の一意数を下方へシフトさせるので、分位点の外でも十分にリスクが下がるかもしれない。特に一意数曲線が連続なら、分位点の近傍はリスク管理されているとみなせるだろう。実際に図 1 の左の方では、一意数曲線は連続に見える。このように分位点の外の攻撃者が必ずしも問題とは限らないが、そのような攻撃者には匿名化以外の手法で対処せざるを得ない。例えば攻撃者の分布の右裾を急峻に落とすような制度的工夫が効果的と考えられる。具体的には、攻撃者分布の右裾に加わりそうな者にデータを渡さなければよい。2.1 節で議論したように、これまでの議論は攻撃者がデータを入手可能ということが前提であった。

3.2　分位点の開示リスク評価

前節の議論から、個体識別の容易度 $\mathrm{Pr}(a,b,c)$ を評価する対象は大半の攻撃者である。本節ではその方針の下、具体的に $\mathrm{Pr}(a,b,c)$ の評価方法を定める。以下では $\mathrm{Pr}(a), \mathrm{Pr}(b|a), \mathrm{Pr}(c|a,b)$ の要因毎に評価方法を議論する。最終的にはそれらの積を $\mathrm{Pr}(a,b,c)$ として用いればよい。

3.2.1　$\mathrm{Pr}(a)$ の評価

Marsh らのモデル (1) を導入した時、条件 (a) は公開ファイルと攻撃用ファイルのキー変数で（時点や分類の）符号化基準が同じ状態、と説明された。Marsh らの議論に従えば、公開ファイルの属性値は誤記や誤分類により、攻撃用ファイルの（真の）属性値と異なる。それから属性値が経時変化するなら、公開ファイルの昔の属性値と攻撃ファイルの（今の）属性値は異なるだろう。故にそれらの程度を見れば、$\mathrm{Pr}(a)$ が評価できることになる。

まず Marsh らによると、キー変数が増えれば誤分類が混入する確率は増えるが、母集団一意数は増える。ただし具体的な確率評価ではこのような一般論は用いず、1981 年の英国センサスの事後調査 (Post Enumeration Survey) で求めた変数の誤分類率を参照している。結果として 1991 年の英国センサスについて、5 つのキー変数が全て正確に分類されている割合は 0.8 程度と見積もられている。そして誤分類が公開ファイルと攻撃用ファイルのキー変数で独立に起きていると仮定し、$\mathrm{Pr}(a) = 0.8^2 = 0.64$ と評価した。

それから経時変化について、1971 年の英国センサスの 1 年後に再調査した結果、同じ職業だった人の割合が 61%でしかない例を Marsh らは挙げている。1991 年の英国センサスについては、Dale

[21] ただし個体が調査されたという情報をキー変数として用い、リサンプリング率の妥当性を検討できる。リサンプリングは、公開標本と統計調査の標本を乖離させることが目的である。それにより被調査個体を知る攻撃者でも、個体を識別するには公開標本の一意が統計調査標本の一意でもあることを確証しなければならない。その難易度は、公開標本中の統計調査標本一意が多ければ下がるはずだ。このような構造は、公開標本中の母集団一意が多いほど確証しやすいと考える我々のモデルと平行している。従って被調査個体を知る攻撃者にとってリサンプリング率が十分か判断するには、公開標本中の統計調査標本一意の割合を異なる事例で比較すればよい。

and Elliot (2001) が各キー変数の経時変化する程度を調べている。ただ Dale and Elliot も述べているように、本気の攻撃者は標的の調査が数年後に公開されることを見込み、同時点に調査した攻撃用ファイルを準備しておくだろう。この場合は、調査時点や変数の定義の差に多くの保護効果を期待出来ない。

やはり $\Pr(a)$ の評価も攻撃方法の想定（シナリオ）に依存する。従って我々は大半の攻撃者を管理するという方針で、$\Pr(a)$ を評価するためのシナリオを定めなければならない。

多くの攻撃者は公開ファイルと同じ基準の攻撃用キー変数を持たないはずだが、どの程度同じと考えたらよいか。この問題は攻撃用情報の精度の評価であり、一意数を計算するためにキー変数の区分を定めた時と同様に考えることにする。我々は推定誤差と手間を天秤に掛けて、キー変数の選択において順位の内点を検討から外した。同様に我々のシナリオでは、公開ファイルと攻撃用ファイルで調査時点の差を設けない。両者の符号化の差は、匿名化によるもののみとする。このように考えると存在しない最強の攻撃者を想定していると見えるかもしれないが、キー変数が所与であることに注意しよう。

具体的に $\Pr(a)$ を評価するにはどうしたらよいか。我々のシナリオを言い換えると、所与のキー変数について、攻撃者が匿名化前のデータを見られる状態、ということになる。つまり所与のキー変数の匿名化前のデータが、攻撃用ファイルである。この場合、攪乱による匿名化が用いられていなければ、$\Pr(a) = 1$ である。攪乱がある場合、キー変数に攪乱が施されていないレコードの割合を $\Pr(a)$ とするのが一つの方法である。誤分類や誤記を意図せざる攪乱とみなせば、そのように考えるのが自然であろう。エディットや補定も攪乱と似た効果を持つが、例外的とみなして計量しないことにする。

攻撃者が攪乱前のデータを見られるなら、攪乱されていないレコードが分かることに注意しよう。そのようなレコードの割合を $\Pr(a)$ と考えたが、もし攪乱の逆変換が出来るなら、$\Pr(a)$ は攪乱が破れるレコードも含んだ割合と考えるべきかもしれない。

攪乱が攻撃者に破られる確率の評価については研究蓄積がある。よく用いられる方法では攻撃ファイルと公開ファイルのキー変数を比較し、近いレコードを同個体（のペア）と判定する。そこで正しく判定された割合[22] は、攪乱が破られる確率と解釈できる。例えば伊藤他 (2009) を見よ。

このように攪乱が破られる確率は評価可能だが、手間はかかる。そして攪乱は補助的に使われるのが普通で、その割合は一般に大きくない。事例間で $\Pr(a)$ の評価方針についてゆれを避けたいこともあり、攪乱が破られる可能性は無視してもいいのではないか。$\Pr(a)$ の評価値は、キー変数が攪乱されていないレコードの割合で十分と思われる。

なお実際の攻撃者は攪乱前のデータを見られないだろう。その場合、攻撃者は攪乱が施されているレコードがどれか分からない。従って全てのレコードの真値は不確定である。そして攪乱が使われていなくても、補定やエディットの可能性が有れば、全レコードの真値は不確定である。そもそも、実際の攻撃用情報は攪乱前のデータと同等の精度を持たず、攻撃者にとって全レコードの真値は不確定だろう。

しかしそれでも個体識別は起きるかもしれない。攻撃者が母集団一意の確証のために多くの追加情報を得てキー変数が増え、あるレコードの個体を識別したと確信するに至ったとしよう。そ

[22] 何を分母とするかは議論の余地がある。本当に評価したいのは、母集団一意レコードについてのキー変数の精度である。しかし全数調査でないと、母集団一意のレコードを決めるのは難しい。そして評価の歪みより方法の変動を避けたいので、標本一意数を分母とするのが一案である。近さの計算方法によるが、一意にペア相手が見つかるレコード数と標本一意数はほぼ同じである。なお分母が 0 の場合は $\Pr(a) = 0$ とみなして差し支えないだろう。

の攻撃者が個体識別を仮定して得られる情報を実用し、対象個体の被害が明るみに出れば、母集団一意が確証されたと言えるのではないか。

従って、少しでも攪乱を施せば真値が不確定なので安全という主張[23] は、楽観的過ぎるかもしれない。エビデンスに基づいた匿名化では観測によって安全性を判断するので、攪乱を施せば個体識別が起きないと決めつけるべきではない。攪乱が少ないほど母集団一意の確証は容易になるはずで、本節の方法で $\Pr(a)$ を評価するのは妥当と考える。

3.2.2 $\Pr(b|a)$ の評価

Marsh らの議論で確率 $\Pr(b|a)$ は、公開標本サイズと母集団サイズの比である。例えば 1991 年の英国センサス匿名化標本では 2%となる。

母集団から等確率でリサンプリングするなら個体は等確率で公開ファイルに含まれる。この場合は一律の $\Pr(b|a)$ でよい。しかし多くの標本調査では、個体は不等確率で抽出される。そこから等確率でリサンプリングすれば、不等確率の公開標本を得る。個体毎の乗率は分かるので、原理的には個体毎の $\Pr(b|a)$ を用いるべきかもしれない。

このようなレコードレベルでの開示リスク評価は理想的だが、以下で議論するように母集団一意数はファイルレベルで評価するのが妥当である。従って我々はファイルレベルで平均的に個体が公開される確率を評価すると考え、Marsh らのように一律の $\Pr(b|a)$ を用いる。

3.2.3 $\Pr(c|a,b)$ の評価

Marsh らの議論に従い、母集団一意数と母集団サイズの比を求めればよい。一意を数える前提については 3.1 節で詳しく議論した。問題は、全数調査でない限り、母集団一意数の推定である。星野 (2003) で説明したように、母集団一意数の推定は単純ではない。

Marsh らは英国センサスの全数データが使えなかったため、イタリアのセンサスデータで母集団一意を数えて外挿している。キー変数が 8 つで 10 万人レベルの地域区分を公開するとして、$\Pr(c|a,b)$ は 2.4%程度とされた。なおこの値は世帯単位ではなく個人単位で評価されている。1991 年の英国センサスデータについては、Dale and Elliot (2001) によるとキー変数が 7 つで 12 万人レベルの地域区分を公開する前提で、$\Pr(c|a,b)$ は 4.8%であった。

母集団一意数評価は Marsh らの時代に比べてかなり進歩している。母集団のデータが利用出来なくても、公開ファイル[24] から母集団一意数を推定できる。しかし推定手法による結果の違いは大きく、相対比較をすることを考えれば、各事例を同一手法によって評価しなければならない。

幅広い母集団について一意数の推定精度をルーチンワークとして確保するには、ピットマンモデル (Pitman, 1995) の使用を推奨する。この方法においてデータは、無限母集団すなわちピットマ

[23]例えば Cleveland et al. (2012) は、地理情報にスワッピングを施せば個体識別を絶対確実に言えないと主張する。しかしそもそもこの主張は間違っている。地理情報のみ攪乱の可能性があるなら、地理情報以外の情報で母集団一意が確証されてしまえば個体は識別される。例えば女で 99 歳の弁護士が母集団一意としよう。そのようなレコードは現実と地理情報が異なっていてもスワッピングが原因であり、母集団一意は確証される。

ただし地理情報のスワッピングは二次元のノイズを乗せるので、年齢などにスワッピングを施す際の一次元ノイズより破りにくい。従って匿名化手法としては推奨する。

[24]攪乱等の匿名化を施した後のデータで母集団一意数を推定する。そうしないとミクロアグリゲーションのような手法の安全性が一意数に反映されない。

ン分布からの標本とみなされる。そして母集団[25]も同一無限母集団からの標本とみなすので、データからピットマン分布の母数を最尤推定し、推定値の下で母集団一意数の挙動を求める。より具体的には、付録の手順書を参照されたい。

開示リスクを評価するファイルのレコード数は、母集団のせいぜい一割程度であろう。この場合に安定的な母集団一意数の推定量は、全てバイアス[26] を持つ。手順書の推定量も例外でなく、おそらく過大に一意数を推定する。しかし我々は $\mathrm{Pr}(a,b,c)$ の相対関係のみ利用するので、$\mathrm{Pr}(a,b,c)$ の大小関係に影響しないバイアスは無視できる。例えば、定率や定数のバイアスは問題にならない。推奨手法のバイアスが定率もしくは定数とは言えないが、同一手法で全ての事例を評価すれば、バイアスの問題は軽減されるはずだ。

なお特定のモデルと決めつけるよりも、モデル集合からデータに良く当てはまるモデルを選択し、そのモデルで一意数を推定する方が正確になろう。しかしモデル集合の空間をうまく張らないと、一意数の推定は不安定になる。また経験的に多くの場合、ピットマンモデルが選択[27] される。故に手間や精度及び結果の整合性を勘案すれば、母集団一意数は常にピットマンモデルによって推定するのが最善と思われる。

母集団一意数を評価する他の方法として、レコードレベルの属性情報を利用した対数線型モデルによる推定が研究されている。しかし実データは大規模かつ疎な分割表なので、うまくいかないだろう。またそのような方法はデータ毎に個別のモデリングが要求され、開示リスク評価の試行錯誤にも向かない。実務への採用は難しいはずだ。

3.3　具体例—住宅・土地統計調査

本節では、平成 15 年住宅・土地統計調査の匿名データについて、実際に個体識別の容易度を算出する。本調査は平成 15 年 10 月 1 日現在の状況を明らかにしている。

本匿名データのレコードは、住宅（住宅以外の建物も含む）、持ち家の属性と居住世帯の属性から構成されている。含まれている変数について、正確には独立行政法人統計センターのホームページ等に掲載されている「データレイアウト及び符号表」を参照のこと。

居住世帯のない住宅（空き家）のレコードも存在するので、個体識別の対象は住宅が自然かもしれない。しかし匿名データの定義（統計法第 2 条第 12 項）には「特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む。）ができないように加工」と書いてあり、住宅は個体識別の対象とは読めない。従って我々は世帯を個体識別の対象とし、空き家のレコードは除く。

結果として公開標本サイズは 31 万 266 となる。本調査の報告書（全国編第 2 表）によれば、同時点の総世帯数は約 4726 万 (47255300) である。従ってこれを母集団サイズとおく。結果として

[25]一部が観測されている現実の母集団について推定するのではなく、同サイズの母集団を新たに発生させる場合の挙動が推定される。

[26]有限母集団から非復元単純無作為抽出する場合、一意数の不偏推定量は一意に存在する。しかしこの不偏推定量は標準誤差が大きく、標本抽出率がかなり高く（8 割程度と思う）ないと実用に耐えない。そして一意な不偏推定量なので、推定を安定させるためのいかなる工夫もバイアスを生む。

[27]裾の長いモデルとして代表的な負の二項分布は、統計的開示制限の分野ではポアソン＝ガンマモデルとして知られている。このモデルは基本的に広義のピットマンモデルの特殊ケース ($\alpha \leq 0$ に対応）である。故にピットマンモデルのデータへのあてはまりは、基本的にポアソン＝ガンマモデルを下回らない。そしてポアソン＝ガンマモデルによる母集団一意数の推定値は、必ずピットマンモデルの推定値より（かなり）小さくなると考えて良い。

標本抽出率、すなわち $\Pr(b|a)$ は 0.66 パーセントと評価される。本調査は地域毎に抽出率が異なる層化 2 段抽出法で行われたが、3.2.2 節で議論したように一律の $\Pr(b|a)$ を用いる。それから本匿名データは撹乱されていないので、$\Pr(a) = 1$ とみなす。

残るは $\Pr(c|a,b)$ の評価である。3.2.3 節で議論したように、ピットマンモデルを用いて母集団一意数の推定値 $\hat{S}_1$ を求める。そして $\hat{S}_1$ を母集団サイズ (4726 万) で割った値を $\Pr(c|a,b)$ の評価値とする。

母集団一意数の推定で、問題はキー変数の選択である。我々は公知と考えられる変数をキーとして用いるが、具体的な判断は分かれるだろう。従って 8 種類のシナリオについて結果を報告する。それらのシナリオを Case 1 から 8 で表し、選択したキー変数の種類を以下に記す。Case 2 は Case 1 から都道府県コードをキー変数から落としたという意味であり、Case 4 は Case 3 にキー変数として世帯主情報（性別、年齢、従業上の地位）を加えたということである。これらの中で一つだけシナリオを選ぶなら、Case 6 と著者は判断する。

- Case 1：都道府県、世帯の種類、同居世帯の有無、夫婦の組数、家族類型、世帯の型、65 歳以上の世帯員の有無、75 歳以上の世帯員の有無、65 歳以上の世帯員のみか、75 歳以上の世帯員のみか、高齢夫婦の有無、世帯内の最高年齢
- Case 2：Case 1-都道府県
- Case 3：都道府県、世帯員各員について性別・年齢（15 歳未満は各歳）・配偶者の有無・続柄
- Case 4：Case 3+世帯主情報（性別、年齢、従業上の地位）
- Case 5：Case 4+現在の居住形態、所有の形態
- Case 6：Case 5+建物に関する事項、むねに関する事項、住宅の種類、所有関係、民営借家の所有区分、住宅の建て方、建築の時期
- Case 7：Case 6+地下室有無、自動車所有の有無、駐車スペースが敷地内、敷地外、住宅の購入・新築・建て替え等の別、H11 年以降の増改築有無
- Case 8：Case 7+台所、トイレ、浴室の設備状況

表 1 は、これら 8 種類のシナリオについて、母集団一意数の推定値 $\hat{S}_1$、$\Pr(c|a,b)$ の評価値、算出された個体識別の容易度 $\Pr(a,b,c)$ をまとめたものである。

4　個票開示リスクの総合的管理

これまでの議論で積み残した重大な問題として、エビデンスとして使える事例の範囲の決定が挙げられる。すなわち個体識別の容易度の閾値 β が共通する範囲を定めなければならない。我々はこの問題を 4.1 節で議論することにする。

それからこれまで提示した理論モデルは様々な仮定を用いているが、それが現実と近ければ提示された手法は実効性を持つはずだ。そのため現実を仮定に近づけるような制度的工夫を各所で議論してきたが、4.2 節でまとめることにする。

	$\hat{S}_1$	$\Pr(c\|a,b)$	$\Pr(a,b,c)$
Case 1	4 918 819	.104	.00068
Case 2	1 683 983	.036	.00023
Case 3	5 038 968	.107	.00070
Case 4	6 871 365	.145	.00096
Case 5	9 374 185	.198	.00130
Case 6	29 082 561	.615	.00404
Case 7	35 610 454	.753	.00495
Case 8	42 962 590	.909	.00597

表 1: 匿名データ（H15 住宅・土地統計調査）の開示リスク

最終的に、理論モデルに基づいた制度運用による、総合的な個票開示リスクの管理が提案されている。

4.1 個体識別が可能となる閾値 β の共通範囲

我々は個体識別の容易度の許容限界 β が共通する事例の範囲を、$\Pr(a,b,c)$ の計量方法を定めた後に決めるということであった。2.3 節の最後で述べたように、個体識別の成功に影響する要因が $\Pr(a,b,c)$ に反映されていなければ、β を変動させる。従ってそのような要因を挙げて、効果を検討しなければならない。

まず、攻撃者が母集団一意を確証する能力が未知の β で表されていることを思い出そう。そして母集団一意は指紋に関する全数情報により確証されるので、母集団一意の確証能力は集団に関する情報収集能力と言える。

そのように考えれば、母集団が異なるとその集団に関する情報収集の容易さも異なるはずで、β も異なるのではないか。例えば、世帯データと事業所データの匿名化事例は同列には扱えないことになる。なお世帯よりも事業所の方が一般的に個体識別をしやすいと言われている。

では母集団の経時変化は、集団に関する情報収集の容易さを変化させるだろうか。Marsh らは $\Pr(a)$ の評価で経時変化を考慮に入れたが、キー変数はデータ計測時点の既存情報に基づいて選択するのが筋である。だからこそ我々のシナリオでは、公開ファイルと攻撃ファイルで調査時点の差がない。そして母集団についての既存情報は変化する。例えば高額納税者の番付は、過去のデータのみ公知である。このように考えると、既存情報の経時変化について β への影響を検討しなければならない。

既存情報の変化を、変数の種類の変化、変数の精度の変化、変数が既知な個体量の変化、の三側面から検討する。まず（公知の）変数の種類の変化はキー変数の変化として扱うので、$\Pr(a,b,c)$ の評価に反映される。それから公知の変数の精度の変化は、$\Pr(a)$ の評価では無視できるということであった。しかし変数が公知な個体の量は、$\Pr(a,b,c)$ の評価において議論されていない。

変数が部分集団についてのみ公知という場合はあり得る。その部分集団について、本来はキー変数を追加して母集団一意数を評価するのが正しい。特に確証用ファイルを作りやすい部分集団

については、この手間を惜しんではならない。ただ 2.1 節で議論したように、匿名化によりそのような部分集団が埋もれていることが前提であった。

変数が公知な個体数量増加を、単にマクロで照合可能な個体数量の増加と考えればどうなるか。これは既存情報が増加したということであり、確証用の全数情報が変化しなければ、確証用ファイルの作成労力が減る。従って母集団一意の確証は容易になると考えられる。公知な個体数量が減少すれば、その逆になるだろう。しかし β への影響を定量的に評価することは難しい。実際には公知な個体量を継続観察し、あまり変化が無ければ β が共通とみなすのが妥当と思われる。

公知な個体量の継続観察は、匿名データのチェックリストで行えばよい。そこには「外部の情報」を記入する欄が設けられており、この欄に公知の個体情報が記入されるべきである。公知な情報としては、民間データベースなどが挙げられよう。記入事項は変数の種類と含有集団の規模（個体数量）であると、明確化するべきである。

これまでの議論では確証用情報と既存情報の量の差が β に影響すると考えたが、その差が一定でも攻撃者の情報収集能力が変化すれば β は変わるだろう。従ってデータ公開事例において攻撃者の情報収集能力が一定でなければ、β が共通するとはみなせないはずだ。特に問題になるのは情報収集能力の向上であり、統計当局はそれを監視しなければならない。匿名データについては利用審査があり、それにより攻撃者の情報収集能力をある程度管理できるはずだ。次節で攻撃者分布の右裾の管理とからめて議論する。

他にこれまで評価していない匿名化の安全要因である「曖昧さ」を検討しておこう。ここでは匿名化に用いたデータ変換の形を攻撃者が完全には知らない場合を曖昧と呼ぶ。例えば国勢調査の匿名データのように、スワッピングが施されているがその割合やスワップ相手の選択方法などが未公開な状態は曖昧である。他方、労働力調査等の匿名データでは、符号表を読むことで匿名化が施されている変数や程度が分かる。この状態は曖昧ではない。なお補定やエディットは攪乱と同等なので、それらの詳細が明らかでない状態も曖昧と考えられる。従って匿名化について曖昧さがないデータでも、統計当局は補定やエディットによる曖昧さの存在を主張できる。

曖昧さが母集団一意の確証に影響する例を挙げよう。年齢と性別の 2 キー変数について、元ファイルが $\{(110,\mathrm{M}),(120,\mathrm{F})\}$、公開ファイルが $\{(120,\mathrm{M}),(110,\mathrm{F})\}$ だとする。年齢をスワップしたと考えれば第一レコード同士が同一個体であり、性別をスワップしたと考えれば元ファイルの第一レコードと公開ファイルの第二レコードが同一個体となる。もし匿名化手法について何も情報が無ければ、公開ファイルの個体と元ファイルを通した母集団の個体との対応は分からない。ところが「年齢変数に適当なノイズを付加した」という程度の情報が有れば、元ファイルと公開ファイルで同一個体が分かる。このように母集団一意の確証可能性は、曖昧さの程度に依存する場合がある。

ただ、一般に曖昧さは余り研究されておらず、その効果[28]に定説はない。一つの理由として、計算機科学では曖昧さによる安全性を認めないことが挙げられる。曖昧さは統計当局が情報を隠している状態であり、匿名化を曖昧さに依存して設計すると、隠した情報が漏れた場合[29] に設計意図が達成されない。従って計算機科学では保守的に考え、曖昧さが失われた状態で議論をする。

[28]特定の曖昧さの効果は、例えば以下のように評価できる。保守的な攻撃者なら、曖昧な部分に自分に不利な事前分布（主観）を入れる。このように評価される攻撃の難易度と真の難易度（客観）の差が、曖昧さの効果である。

[29]関係者による情報漏洩だけ考慮すれば良いわけではない。攻撃者が攪乱の母数を推定できる可能性がある。例えば匿名データと匿名化されていないデータの分析結果を比較することで、攪乱の率の見当をつけられるかもしれない。攻撃者本人が 33 条申請による目的外使用でデータを手に入れなくても、他人が書いた論文や公の集計表が比較対象になり得る。

曖昧さの効果は無しと決めつければ、明らかに閾値 β の変動要因にならない。しかし国勢調査の匿名データのように、曖昧さを使っている事例が存在する。使われれば、効果はゼロかもしれないが、個体識別に関する観測情報は曖昧さ所与での結果である。従って具体的に影響を検討しておくべきだろう。

国勢調査の匿名データでは、スワッピングが施されているレコードの割合が曖昧にされている。このような曖昧さはよく用いられるが、都道府県コードのみ攪乱の可能性があるという情報は無視すれば、その効果は $\Pr(a)$ が攻撃者にとって不確実になることと考えられる。そして母集団一意の確証は、$\Pr(a)$ が既知なら成功だが未知なら失敗ということがない[30]。つまり母集団一意の確証成功という事象は $\Pr(a)$ という情報と独立である。言い換えれば、母集団一意の確証可能性は $\Pr(a)$ の真値に基づいて定まる。だとすれば、スワッピングの割合の曖昧さは β に影響を与えない。

ただし $\Pr(a)$ の不確実性は、攻撃者が識別を試みる確率には影響するはずだ。何故なら母集団一意の確証には情報収集等のコストがかかるが、真の $\Pr(a,b,c)$ が閾値を上回っていなければ無駄足である。従って不確実な $\Pr(a,b,c)$ でも閾値を上回ると判断しなければ、識別を試みないだろう。結果として個体識別発生の社会的認知が起きる確率 $p(\cdot)$ を小さくするが、β の点推定値は変わらない。ただ、$\hat{\beta}$ が過大推定の確率は増大する。

スワッピングの割合の曖昧さについてはこのように評価されるので、閾値 β の変動要因とみなさないことにする。なお他のタイプの曖昧さを用いるとしても、匿名化の一部として明示的に設計するべきである。

4.2　開示リスク管理のための制度運用—チェックリストの改善

本稿の議論の中心は、大半の攻撃者の開示リスクを定量的に管理することである。そしてそのような考え方が妥当であるためには、制度の運用を工夫しなければならない。これまでの議論で指摘されたそのような工夫を、以下にまとめる。

まず 2.1 節では、母集団一意とその確証を分離することの重要性を指摘した。そのための工夫として以下の三点を挙げた。

1. リサンプリングは必須である。
2. リサンプリング処理の担当者も、結果が確率的にしか分からないように内部統制をする。
3. 部分集団に関する全数名簿の存在には特に注意し、それを無力化するように匿名化する。

また母集団一意の確証を妨げる工夫として、一点指摘した。

1. 指紋に関する全数調査が出来ないように、広義の地理情報の精度を粗く保つ。

これらの工夫は、4.1 節で議論したように、閾値 β が一定とみなす前提でもある。

3.1 節では攻撃者分布の分位点で開示リスクを管理することを議論した。もし分布の右裾が長いと、管理されない攻撃者の問題が顕在化しかねない。故に攻撃者分布の右裾を落とすために、分布の右裾に加わりそうな者にデータを渡さない、と述べた。

[30] 3.2.1 節で議論したように、$\Pr(a) = 1$ が既知か否かは確証可能性に影響を与えるかもしれない。しかしエディットや補定の可能性を考えれば、真に $\Pr(a) = 1$ ということはあり得ない。

この問題は追加で議論しよう。攻撃者分布の右裾に位置するということは、既存情報を多く利用できるということである。そのような攻撃者は名簿業者かもしれないし、名簿業者の利用意思がある者かもしれない。おそらくそのような攻撃者は「本気」であり、左裾の興味本位かつのぞき見趣味的な攻撃者とは違う。定量的な開示リスク管理をすれば、興味本位の攻撃者はカバーされる。従ってデータ利用審査では、個体情報を多く持つ者へデータが渡る可能性を特に検討するべきである。

4.1 節では攻撃者の情報収集能力の向上を注視せよと述べたが、それは名簿業者や個人情報収集組織の実態を監視するということでもある。これらは攻撃者分布の右裾の管理に必要な情報である。

4.1 節では他に、閾値 β が一定か判定する為に必要な情報を、チェックリストに記入することも述べた。チェックリストは匿名データのリスク管理において重要な役割を果たすので、あり方を再論する。

まずチェックリストは、蓄積して参照するものだということをはっきりさせておきたい。EBA は過去の経験をエビデンスとして用いる。故に過去のチェックリストは、経験の要約として位置づけられる。従ってチェックリストには、個体識別が可能か否かの判断に用いる情報が過不足なく記入されるのが望ましい。このような観点から、世帯調査のチェックリスト（H23/3/28 改正版）についてのみ、改善できる点を指摘したい。

まず部分集団の全数名簿は母集団一意の確証について重大なので、質問項目として特に欄を設けるべきである。そのような名簿が存在するなら、名称、部分集団の種類、含有個体数量、キー変数の種類を記述させるとよい。他に狭義の地理情報については記入欄が存在するが、広義の地理情報の有無を確認するべきだ。

それからチェックリスト記入の焦点をしぼるべき箇所がある。まず「ミクロデータを特定できる可能性のある外部ファイル」の存在を記入することになっているが、どのようなファイルが該当するのか明確化するべきである。具体的には、照合可能な変数の情報をもつ公知な外部ファイルの有無を問うべきだ。そしてそのファイルの名称、キー変数の種類、及び個体数を分けて記述してもらう方がよい。また「秘密の情報」（センシティブ変数）のうち、「特に秘匿する必要性の高い調査項目」の有無を聞いているが、必要性の意味を明白にしたほうがよい。個体識別の可能性を制限するための必要性ではなく、実用性を限るための匿名化の必要性を聞かなければならない。またチェックリストには「誤差（ノイズ）」を聞く項目が存在する。誤差の付加は「撹乱」手法の例なのだが、用語の問題は別にして、撹乱手法のパラメータと、その公開方針を分けて書かせるべきである。これは匿名化の曖昧さを明示的に設計することに関わる。

またチェックリストに記入される情報の使われ方は、説明書を用意するべきであろう。現行のチェックリストも冒頭で匿名化の考え方などが書かれているが、やや説明不足に見える。個体識別可能性の判定方式を明示すれば、焦点がずれたチェックリスト記入の恐れは減る。

EBA を制度化するなら、個体識別の容易度 $\Pr(a,b,c)$ はチェックリストに記載するべきである。それにより、事例間での比較が容易になる。付録の手順書に従えば、$\Pr(a,b,c)$ の計算はそれほど手間はかからない。匿名化表現の要約として、費用対効果が高いと考える。

それから個体識別の容易度を記載するなら、採用したキー変数も併記するべきである。そのようなチェックリストの蓄積により、Elliot らの主張するキー変数のデータベースが擬似的に構築される。それは個体識別の容易度を相対比較する際の一貫性にとって重要だ。

5　おわりに

これまでの議論により、匿名化の程度は個体識別の容易度で測られ、それは統計的証拠に基づいて定めることができる。ただし個体識別の容易度を管理するだけでなく、それが問題にならないように制度的工夫を併用することが重要である。

なお複数の匿名化手法で同じ個体識別の容易度が達成できるかもしれないが、その場合はデータの有用性が高いものを選べばよい。本稿で有用性の評価は議論しなかったが、例えば星野 (2010) は基本的な考え方を説明している。

本稿では誌面の都合もあり、平成 15 年の住宅・土地統計調査の匿名データのみ、個体識別の容易度を評価した。他の匿名データについての評価は、別途報告する予定である。

Acknowledgements

本研究で利用した匿名データは、統計法（平成 19 年法律第 53 号）に基づいて独立行政法人統計センターから提供を受けた。また本研究は科学研究費及び統計数理研究所の共同研究経費の補助を受けている。

付録

A　世帯調査のチェックリスト（H23/3/28 改正版）要約

1. 地理的情報
   (a) 地理情報のレベル、加工の有無
   (b) 地理情報以外の地理的情報の有無
   (c) 地域分析用の地理情報提供の有無
   (d) 特定の種類の施設の情報の有無
2. 世帯の識別情報
   (a) 世帯のキー変数
   (b) キー変数への匿名化及び分布
   (c) 世帯のまとまりへの匿名化の有無
3. 個人の識別情報
   (a) 個人のキー変数
   (b) キー変数への匿名化及び分布
4. 攪乱の有無
5. サブサンプリングの有無

6. 外部の情報
   (a) 個人・世帯の特定に使える外部情報の存在
   (b) 母集団情報として利用している情報
7. その他
   (a) データの並び順についての匿名化措置
   (b) サンプリング情報により特定の地域や集団であることが明らかになる可能性
   (c) センシティブ変数への匿名化
   (d) 提供時期と調査時点との差
   (e) その他の匿名化処理の有無

B 母集団一意数の推定手順

以下では標本サイズを n、母集団サイズを $\tilde{n}$ と記す。

1. 評価するキー変数とその精度を決める。
2. 決められたキー変数全てについてクロス集計する。つまり（高次元の）分割表を作り、各セルに所属するレコード数（度数）を数える。
   - 第 j セルの度数を f_j と書く。セル総数が J として $j = 1, 2, \ldots, J$ だが、インデクス j の付け方に結果は依存しない。
   - セル総数 J は、全てのキー変数のカテゴリー数の積である。連続変数であっても、開示リスク評価においては（攻撃者の持つ情報精度にあわせて）離散として扱うのが妥当である。
3. 空でないセルの度数の度数（寸法指標）を数える。
   - $i = 1, 2, \ldots, n$ について度数 i のセルの数を s_i と表す。つまり指示関数 $1(\cdot)$ を使えば、寸法指標は $s_i = \sum_{j=1}^{J} 1(f_j = i)$ である。
   - 寸法指標の和を $u := \sum_{i=1}^{n} s_i$ で表す。
   - 寸法指標のベクトルを $\boldsymbol{s}_n := (s_1, s_2, \ldots, s_n)$ で表す。
   - 最大のセルの度数が m ならば、$m < i$ について $s_i = 0$ である。
4. データを生成した構造（確率分布）を推定する。
   - 現実の母集団を無限母集団（超母集団）からの標本とみなす。この場合、手元の標本から超母集団の分布を推定すれば、母集団の挙動も推定される。
   - 超母集団の分布として広義の Pitman モデルを仮定し、その母数を最尤推定する。

- Pitman モデルの確率関数は、母数 $0 \le \alpha < 1, \theta > -\alpha$ について以下のように定まる。

$$p(s_1, s_2, \ldots, s_n) = n! \frac{\theta^{[u:\alpha]}}{\theta^{[n]}} \prod_{j=1}^{n} \left(\frac{(1-\alpha)^{[j-1]}}{j!} \right)^{s_j} \frac{1}{s_j!}, \tag{12}$$

$$\boldsymbol{s}_n \in \mathcal{S}_{|n} := \{\boldsymbol{s}_n : s_i \in \mathbb{N}_0, i \in \{1, 2, \ldots, n\}, \sum_{i=1}^{n} i s_i = n\},$$

ただし $\theta^{[u:\alpha]} = \theta(\theta+\alpha)\cdots(\theta+(u-1)\alpha)$, $\theta^{[n]} = \theta(\theta+1)\cdots(\theta+n-1)$ である。

- Pitman モデルは母数 α が負の場合と正の場合で分けて考えた方が良い。どちらの場合も $u = \sum_{i=1}^{n} s_i$, $n = \sum_{i=1}^{n} i s_i$ である。
  (a) $0 \le \alpha < 1, \theta > -\alpha$ について Pitman モデルの確率関数は (12) 式で表される。
  (b) $\alpha < 0$ の場合は (12) 式で $\theta = -J\alpha$ とおき、さらに $-\alpha = \gamma$ とおく。すると一母数の確率関数を得る：

$$p(s_1, s_2, \ldots, s_n) = \frac{n! J! \Gamma(J\gamma)}{\Gamma(J\gamma + n)} \prod_{i=0}^{n} \left(\frac{\Gamma(\gamma+i)}{\Gamma(\gamma) i!} \right)^{s_i} \frac{1}{s_i!}. \tag{13}$$

  ここで $\gamma > 0$ であり、$s_0 = J - u$ である。

- モデル (12) を「(狭義の)Pitman モデル」と呼ぶ。モデル (13) を「多項ディリクレモデル」と呼ぶ。本来は AIC 等によりデータ依存でいずれかをモデル選択するのが良いが、ここでは簡易的な選択基準を示す：
  (a) 母集団サイズ $\tilde{n}$ が総セル数 J より大の場合、多項ディリクレモデルを用いる。
  (b) その他の場合は Pitman モデルを用いるが、尤度の最大化に失敗する（繰り返し計算が収束しない）場合、多項ディリクレモデルを用いる。

- 狭義の Pitman モデルの最尤推定は以下のように行えば良い。
  (a) 対数尤度関数は定数を除いて

$$L(\alpha, \theta) = \sum_{i=1}^{u-1} \log(\theta + i\alpha) - \sum_{i=1}^{n-1} \log(\theta + i) + s_1 + \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1} \log(j - \alpha).$$

  (b) 最尤推定量は以下の同時方程式の解である.

$$\frac{\partial L(\alpha, \theta)}{\partial \theta} = \sum_{i=1}^{u-1} \frac{1}{\theta + i\alpha} - \sum_{i=1}^{n-1} \frac{1}{\theta + i} = 0,$$

$$\frac{\partial L(\alpha, \theta)}{\partial \alpha} = \sum_{i=1}^{u-1} \frac{i}{\theta + i\alpha} - \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1} \frac{1}{j - \alpha} = 0.$$

  (c) $L(\alpha, \theta)$ の最大化は汎用最大化ルーチン (R の optim() 関数等) に任せても良いだろう。

(d) 二次の微分係数 (14),(15),(16) 式を用いればニュートン＝ラフソン法で最尤推定値を計算できる。

$$\frac{\partial^2 L(\alpha,\theta)}{\partial\theta^2} = -\sum_{i=1}^{u-1}\frac{1}{(\theta+i\alpha)^2} + \sum_{i=1}^{n-1}\frac{1}{(\theta+i)^2}, \tag{14}$$

$$\frac{\partial^2 L(\alpha,\theta)}{\partial\alpha^2} = -\sum_{i=1}^{u-1}\frac{i^2}{(\theta+i\alpha)^2} - \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1}\frac{1}{(j-\alpha)^2} < 0, \tag{15}$$

$$\frac{\partial^2 L(\alpha,\theta)}{\partial\theta\partial\alpha} = -\sum_{i=1}^{u-1}\frac{i}{(i\alpha+\theta)^2} < 0. \tag{16}$$

(e) 近似的モメント推定量：

$$\hat\theta = \frac{nuc - s_1(n-1)(2u+c)}{2s_1u + s_1c - nc}, \quad \hat\alpha = \frac{\hat\theta(s_1-n)+(n-1)s_1}{nu},$$

ただし $c = s_1(s_1-1)/s_2$ は、ニュートン＝ラフソン法の初期値として使える。

(f) θ の推定は不安定なので、初期値をランダムに変えて繰り返し計算が同じ値に収束するか確認するのが望ましい。

- 多項ディリクレモデルの最尤推定は以下のように行えば良い。

(a) 対数尤度関数は定数を除いて

$$L(\gamma) = -\sum_{i=0}^{n-1}\log(J\gamma+i) + \sum_{i=1}^{n} s_i \sum_{j=0}^{i-1}\log(\gamma+j).$$

(b) 最尤推定値は尤度方程式

$$\frac{dL(\gamma)}{d\gamma} = -\sum_{i=0}^{n-1}\frac{J}{J\gamma+i} + \sum_{i=1}^{n} s_i \sum_{j=0}^{i-1}\frac{1}{\gamma+j} = 0$$

の解である。

(c) $L(\gamma)$ の最大化は汎用最大化ルーチン (R の optimize() 関数等) に任せても良いだろう。

(d) 二次の微分係数

$$\frac{d^2L(\gamma)}{d\gamma^2} = \sum_{i=0}^{n-1}\frac{J^2}{(J\gamma+i)^2} - \sum_{i=1}^{n} s_i \sum_{j=0}^{i-1}\frac{1}{(\gamma+j)^2}$$

を用いればニュートン＝ラフソン法で最尤推定値が計算できる。

(e) 尤度関数は単峰であり、それほど初期値に依存せず最大化が可能である。ただ最尤推定値が無限大に発散する事はあり得て、それは確率関数が等確率 J 項分布である事を意味する。また最尤推定値が 0 の場合、狭義の Pitman モデルの方が適切と思われる。

- 狭義の Pitman モデルと多項ディリクレモデルの境界 ($\alpha = 0$) のモデルを Ewens モデルという。Ewens モデルの確率関数は以下の通り：

$$p(s_1, s_2, \ldots, s_n) = n! \frac{\theta^u}{\theta^{[n]}} \prod_{j=1}^{n} \left(\frac{1}{j}\right)^{s_j} \frac{1}{s_j!}. \tag{17}$$

  – 同じデータについて Ewens モデルの最尤推定値を $\hat{\theta}_E$ と書き、Pitman モデルの最尤推定値を $\hat{\alpha}, \hat{\theta}_P$ と書く。もし $\hat{\alpha} > 0$ ならば

$$\hat{\theta}_E > \hat{\theta}_P.$$

  – 上の式は Pitman モデルのチェックに使える。また最尤推定の繰り返し計算の範囲を限定できる。
  – Ewens モデルの尤度関数は単峰であり、最大化は容易である。

5. 同定されたデータ構造の下で母集団一意数の推定値 $\hat{S}_1$ を求める。

   (a) 狭義の Pitman モデルの場合、母数の最尤推定値を $\hat{\alpha}, \hat{\theta}$ と書けば (18) で推定される。

$$\hat{S}_1 = \tilde{n} \frac{(\hat{\theta} + \hat{\alpha})(\hat{\theta} + \hat{\alpha} + 1) \cdots (\hat{\theta} + \hat{\alpha} + \tilde{n} - 2)}{(\hat{\theta} + 1)(\hat{\theta} + 2) \cdots (\hat{\theta} + \tilde{n} - 1)}. \tag{18}$$

   (b) 多項ディリクレモデルの場合、母数の最尤推定値を $\hat{\gamma}$ と書けば

$$\hat{S}_1 = \tilde{n}(J-1)\hat{\gamma} \frac{((J-1)\hat{\gamma} + 1)((J-1)\hat{\gamma} + 2) \cdots ((J-1)\hat{\gamma} + \tilde{n} - 2)}{(J\hat{\gamma} + 1)(J\hat{\gamma} + 2) \cdots (J\hat{\gamma} + \tilde{n} - 1)}.$$

- これらの推定値は、度数 1 のセル数の期待値の母数に推定値を代入して得られる。
- 注 1）Ewens モデルの母集団一意数推定式は、Pitman モデルの推定式に $\hat{\alpha} = 0$ を代入して得られる。
- 注 2) 等確率 J 項分布の母集団一意数推定値は $\tilde{n}(1 - 1/J)^{\tilde{n}-1}$ である。

参考文献

[1] Chaudhuri, K. and Mishra, N. (2006) When Random Sampling Preserves Privacy. sl Proceedings of the 26th Annual International Conference on Advances in Cryptology (CRYPTO 2006), 198–213, Springer, Berlin.

[2] Cleveland, L., McCaa,R., Ruggles, S. and Sobek, M. (2012). When Excessive Perturbation Goes Wrong and Why IPUMS-International Relies Instead on Sampling, Suppression, Swapping, and Other Minimally Harmful Methods to Protect Privacy of Census Microdata. *Privacy in Statistical Databases*, Domingo-Ferrer, J and Tinnirello, I. (Eds.). LNCS 7556, 179–187, Springer-Verlag, Berlin Heidelberg.

[3] Dale, A. and Elliot, M. (2001) Proposal for 2001 Samples of Anonymized Records: An Assessment of Disclosure Risk. *Journal of the Royal Statistical Society*, Series A, **164**, 427–447.

[4] Dalenius, T. (1986). Finding a needle in a haystack — or identifying anonymous census records. *Journal of Official Statistics*, **2**, 329–336.

[5] De Waal, A.G. and Willenborg, L.C.R.J. (1994) Minimizing the number of local suppressions in a microdata set, Internal Report, Statistics Netherlands, 1–16.

[6] Domingo-Ferrer, J. and Torra, V. (2001) A Quantitative Comparison of Disclosure Control Methods for Microdata. *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, Doyle et al. (Eds.), 111-133, Elsevier, Amsterdam.

[7] Duncan, G., Keller-McNulty, S.A. and Stokes, S.L. (2001) Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report 121, National Institute of Statistical Sciences, Durham, North Carolina.

[8] Elliot, M. J., Skinner, C. J., and Dale, A. (1998) Special Uniques, Random Uniques, and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk. *Research in Official Statistics*, **1**, 53–67.

[9] Elliot, M., Lomax, S., Mackey, E. and Purdam, K. (2010) Data Environment Analysis and the Key Variable Mapping System. *Privacy in Statistical Databases*, Domingo-Ferrer, J. and Magkos, E. (Eds.), LNCS 6344, 138–147, Springer-Verlag, Berlin Heidelberg.

[10] Elliot, M., Mackey, E. and Purdam, K. (2011) Formalizing the Selection of Key Variables in Disclosure Risk. *Int. Statistical Inst.: Proceedings of the 58th World Statistical Congress*, 2777–2784.

[11] Fung, B.C.M., Wang, K., Fu, A.W.C and Yu, P.S. (2010) *Introduction to Privacy-Preserving Data Publishing*, CRC Press, New York.

[12] 星野伸明 (2003) 「超母集団モデルによる個票開示リスク評価」, 統計数理, **51**, 297–319.

[13] Hoshino, N. (2009) The Quasi-multinomial Distribution as a Tool for Disclosure Risk Assessment, *Journal of Official Statistics*, **25**, 269–291.

[14] 星野伸明 (2010)「公的統計ミクロデータ提供制度の課題」, 日本統計学会誌, **40**, 23–45.

[15] 伊藤伸介 (2012) 「政府統計ミクロデータの提供における匿名化措置—イギリス統計法における法制度的措置と攪乱的手法の適用可能性を中心に—」, 明海大学経済学論集, **24**, 1–14.

[16] 伊藤伸介・磯部祥子・秋山裕美 (2009) 「秘匿性の評価方法に関する実証研究—全国消費実態調査のミクロアグリゲートデータを用いて—」, 統計センター製表技術参考資料, **11**, 12–14.

[17] Marsh, C., Skinner, C., Arber, S., Penhale, P., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991) The Case for a Sample of Anonymized Records from the 1991 Census. *Journal of the Royal Statistical Society*, Series A, **154**, 305–340.

[18] Paass, G. (1988) Disclosure Risk and Disclosure Avoidance for Microdata. *Journal of Business and Economic Statistics*, **6**, 487–500.

[19] Pitman, J. (1995) Exchangeable and Partially Exchangeable Random Partitions. *Probability Theory and Related Fields*, **102**, 145–158.

[20] 総務省政策統括官（統計基準担当）(2011). 「匿名データの作成・提供に係るガイドライン（平成 23 年 3 月 28 日改正版）」

[21] Sweeney, L. (2002) k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, **10**, 557–570.

[22] 竹村彰通 (1997) 「個票データ開示の理論」, 科学研究費補助金（課題番号 08209102）報告書, 2–25.

[23] U.S. Office of Federal Statistical Policy and Standards (1978). *Report on Statistical Disclosure and Disclosure Avoidance Techniques.* Statistical Policy Working Paper 2, U.S. Department of Commerce, Washington DC.

[24] Willenborg, L.C.R.J., De Waal, A.G. and Keller, W.J. (1995) Some methodological issues in statistical disclosure control, Internal Report, Statistics Netherlands, 1–13.

Estimating Daily Inflation Using Scanner Data: A Progress Report

Kota Watanabe
Meiji University
& Univ of Tokyo

Tsutomu Watanabe
University of Tokyo

January 30, 2015

Outline of the paper

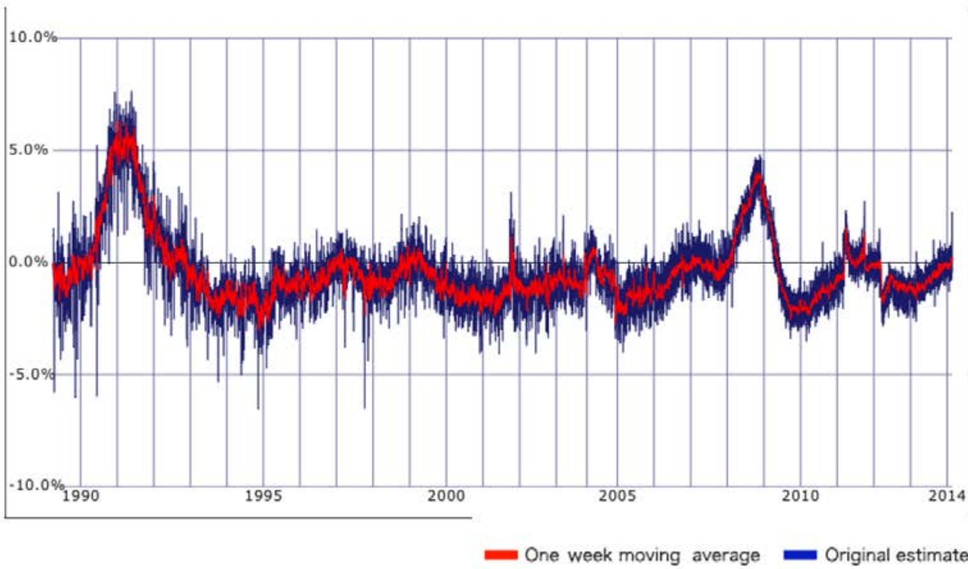
1. Introduction
2. Data
3. Methodology
   - Three-stage aggregation of price changes
   - Comparison of the POS index with the consumer price index
4. Results
   - Daily and monthly indexes
   - Item-level inflation
5. Some additional experiments
   - Core inflation
     - Time series decomposition
     - Trimmed mean estimators
   - Chained Tornqvist index
6. Conclusion

UTokyo Daily Price Project

The Daily Index for the most recent period has been updated on 2014-02-24

Daily Index | Monthly Index | FAQ

Daily Index > Nationwide



Original estimate:	-0.23%
One week moving average:	0.11%
Year-to-year inflation rate for 2014-02-21	

- Prices are collected every day from **300 supermarkets** sampled from all over Japan.
- The number of products is about 300K.
- Updated every day with a **two day lag**.
- Covers **17 percent** of the official CPI.
- Historical daily data is available over the last 25 years.

The current and historical data are downloadable. **Free!**

Daily Index

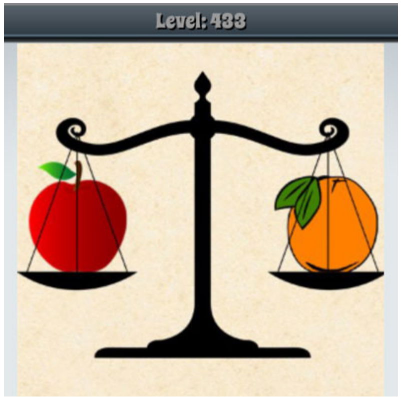
http://www.cmdlab.co.jp/price_u-tokyo/dailys_e

How is the UTokyo Daily Price Index calculated?

- The UTokyo Daily Price Index is **a daily version of the Törnqvist index**, which is known as one of the superlative price indexes.
  - CPI Manual released by ILO: "Many different kinds of mathematical formulae have been proposed over the past two centuries. While there may be no single formula that would be preferred in all circumstances, **most economists and compilers of CPIs seem to be agreed that, in principle, the index formula should belong to a small class of indices called superlative indices**." (Consumer price index manual: Theory and practice, 2004, p.2)
- **Price relatives**: The daily inflation rate is calculated as the weighted geometric mean of price relatives across products, which are defined as the price ratios between today (*t*) and some day in the past (*t-dt*). For example, *dt*=365.
- **Weights**: The weight for a product is given by the average of the sales shares of the product today and the sales share of the same product on the same day of the previous year.

$$\pi_{t,t-dt} = \sum_i \frac{s_{i,t} + s_{i,t-dt}}{2} \ln\left(\frac{p_{i,t}}{p_{i,t-dt}}\right)$$

東大指数とCPIの集計方法の比較



	CPI	東大指数
どの銘柄を対象とするか？	品目（例えばバター）に属する銘柄の中で特定の属性を満たすものだけが対象銘柄となる（これを「基本銘柄」とよぶ）。例えば，バターの基本銘柄は「カルトン入り，200g」。この条件を満たすのは30銘柄。	予め対象銘柄を絞り込むことはせず，全ての銘柄を対象とする。バターで言えば，全銘柄の数は369であり，CPIの基本銘柄の12倍。
同一の商品をどのように定義するか？	バターという品目に属する銘柄は全て同じ商品とみなす。異なる店舗で売っているバターも同じ商品とみなす。つまり，バターという品目に含まれる限り，銘柄間・店舗間の差は無視できるほどに小さいと仮定。	バターという品目に属する銘柄であっても完全に同一ではない。また，同じ銘柄であっても，店舗によって価格帯が違うので同一商品とは言えない。銘柄間・店舗間の異質性を重視する。同一商品を「ある店舗におけるある銘柄」と狭く定義。
価格の平均か価格比の平均か？	全国から約600のバターの価格を収集する。異なる銘柄，異なる店舗が混在することは無視して，600の価格を平均する。その上で前年同月との比を計算する。	異なる銘柄，異なる店舗の価格を平均することはしない。ある店舗のあるバターの銘柄の前年同日比をとり，それを銘柄間・店舗間で平均する。
単純平均か加重兵平均か？	総務省はどの銘柄のバターが売れているかの情報は集めていない。そのため，600のバターの価格をウエイトなしで平均する（単純平均）。	バターの各銘柄の販売シェアをウエイトとして（これを「トルンクビスト・ウエイト」とよぶ），各銘柄の前年同日比を加重平均する。
幾何平均か算術平均か？	算術平均	幾何平均

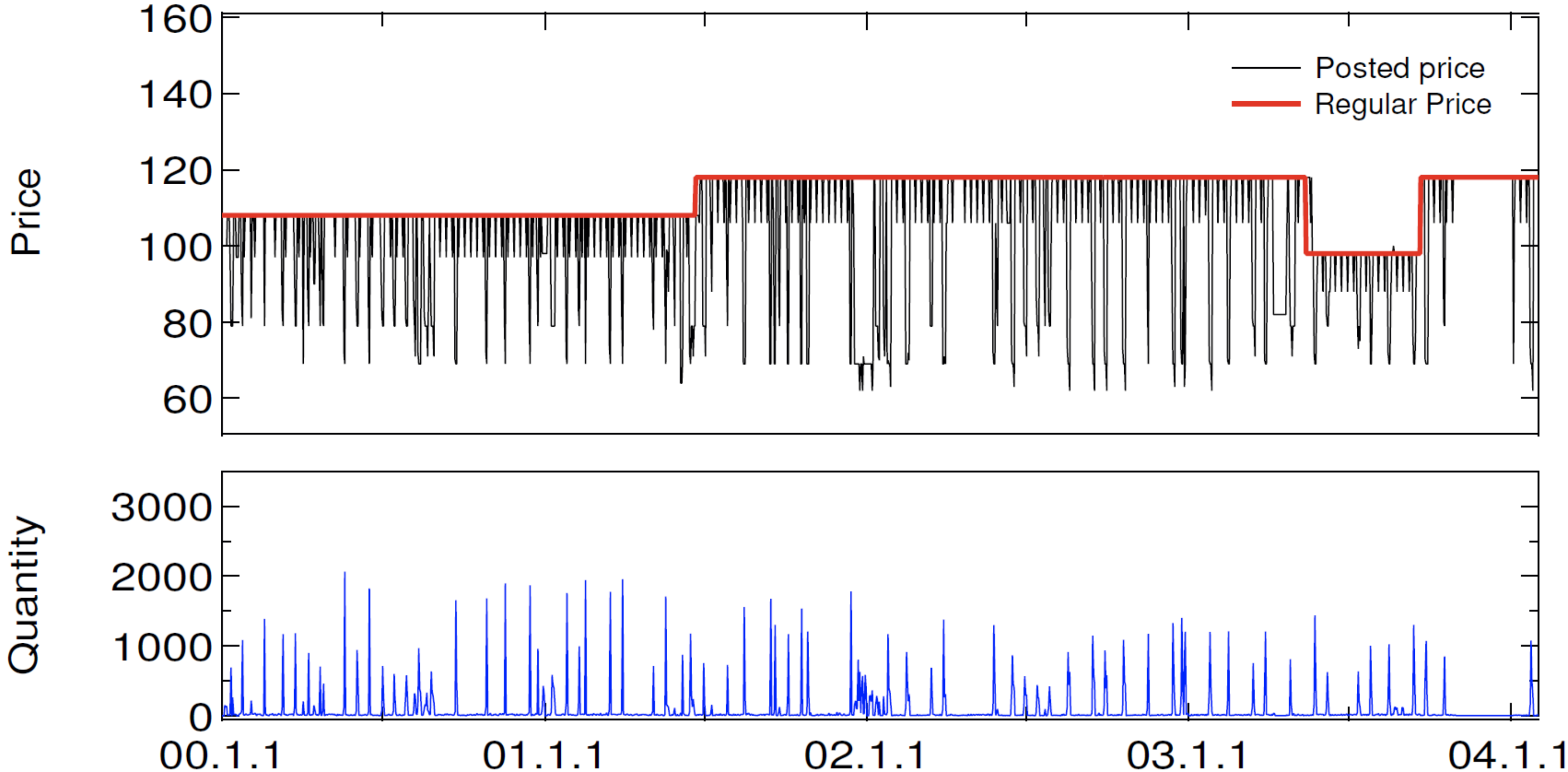
Original data is provided by Nikkei Digital Media, Inc.

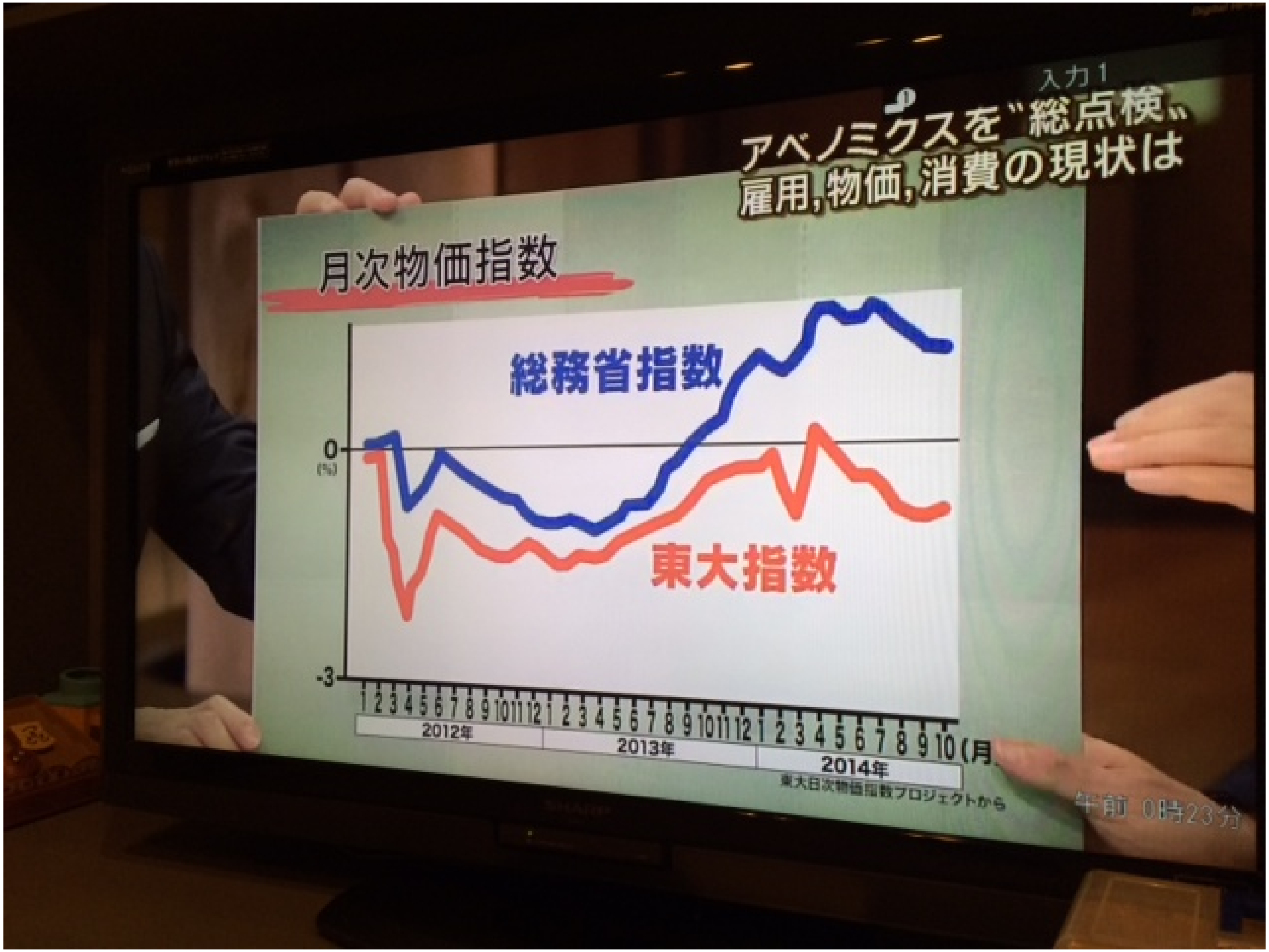
Year	No. of retailers	No. of products	Sales amount (yen)	No. of records
1988	29	88,207	24,967,387,530	25,397,753
1989	45	118,459	38,848,140,951	39,967,625
1990	50	131,217	47,914,018,985	46,449,145
1991	53	133,201	56,554,113,519	50,762,796
1992	62	135,862	67,325,003,923	56,069,411
1993	65	139,929	75,403,002,651	61,371,512
1994	103	157,148	115,779,158,308	91,670,103
1995	124	169,366	149,242,076,718	119,894,820
1996	132	177,116	180,557,355,210	150,298,311
1997	150	194,522	205,874,958,531	171,939,036

....

2004	202	278,894	306,121,269,565	281,899,515
2005	187	287,680	328,939,470,128	309,625,996
2006	189	305,223	334,615,509,093	323,381,091
2007	274	347,185	373,166,817,586	378,924,802
2008	261	367,064	407,677,569,675	412,836,053
2009	264	357,928	404,988,058,786	416,290,153
2010	259	358,282	395,223,198,995	415,348,828
2011	249	358,813	380,908,900,263	403,645,269
2012	261	356,587	399,628,611,703	445,046,118

Sales price and quantity sold for a particular product at a particular retailer





テレビ朝日 報道ステーション「アベノミクスを”総点検” 雇用、物価、消費の現状は」(2014年11月28日放送）

メディア報道の例：Financial Times “Why Shinzo Abe could nix tax hike two”（November 16, 2014）

ft.com > comment > blogs >

FTAlphaville

Why Shinzo Abe could nix tax hike two

Ben McLannahan Author alerts Nov 16 16:12 2 comments

For clues as to why Japan's prime minister seems very keen to avoid another consumption tax increase so soon after the last, you could look at a whole host of economic indicators – third-quarter GDP, consumer confidence surveys, industrial production or housing starts.

Or you could just examine charts put together by Tsutomu Watanabe, a Tokyo University professor who has spent much of the past six years poring over point-of-sales data from supermarkets.

A glance at his UTokyo Daily Price Indices suggests that the April 2014 tax hike – from 5 per cent to 8 per cent – has had just as chilling an effect on consumer demand as the last one, in April 1997.

Back then, when Japan lifted the tax from 3 per cent to 5 per cent, the economy was in recession within a year, with weakness exacerbated by implosions in Japan's banking sector and the effects of the regional currency crisis. Deflation took a grip not long after that.

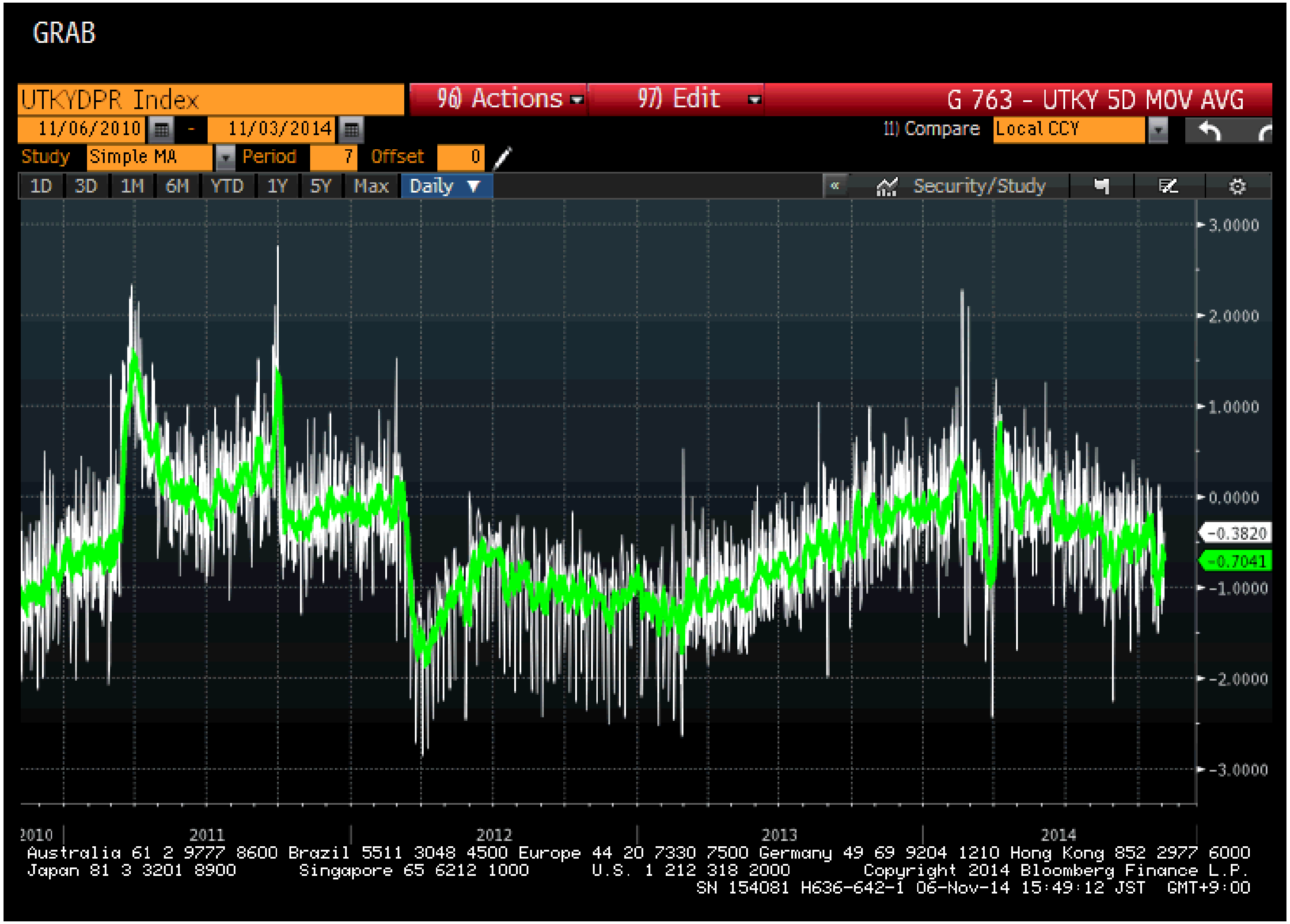
As government officials pushed for this year's increase, they swore that the same thing could not happen again. Not only was Asia in a much better state, they said, but the Bank of Japan was pumping unprecedented amounts of liquidity and the finance ministry was easing the extra burden on households through the biggest fiscal stimulus package it could muster.

But according to Prof Watanabe's crunching of daily price data – gathered from scans of hundreds of thousands of taxable items sold at about 300 supermarkets around the country,and collected by a unit of Nikkei Inc, the publishing firm – there are still strong similarities in the apparent effects of the hikes of '97 and '14.

The two indices show the same basic pattern: a confident pass-through of the full tax (and then some) in early April, an instant snap-back, then a patchy recovery which peters out by early June. Thereafter, prices are falling, year-on-year.

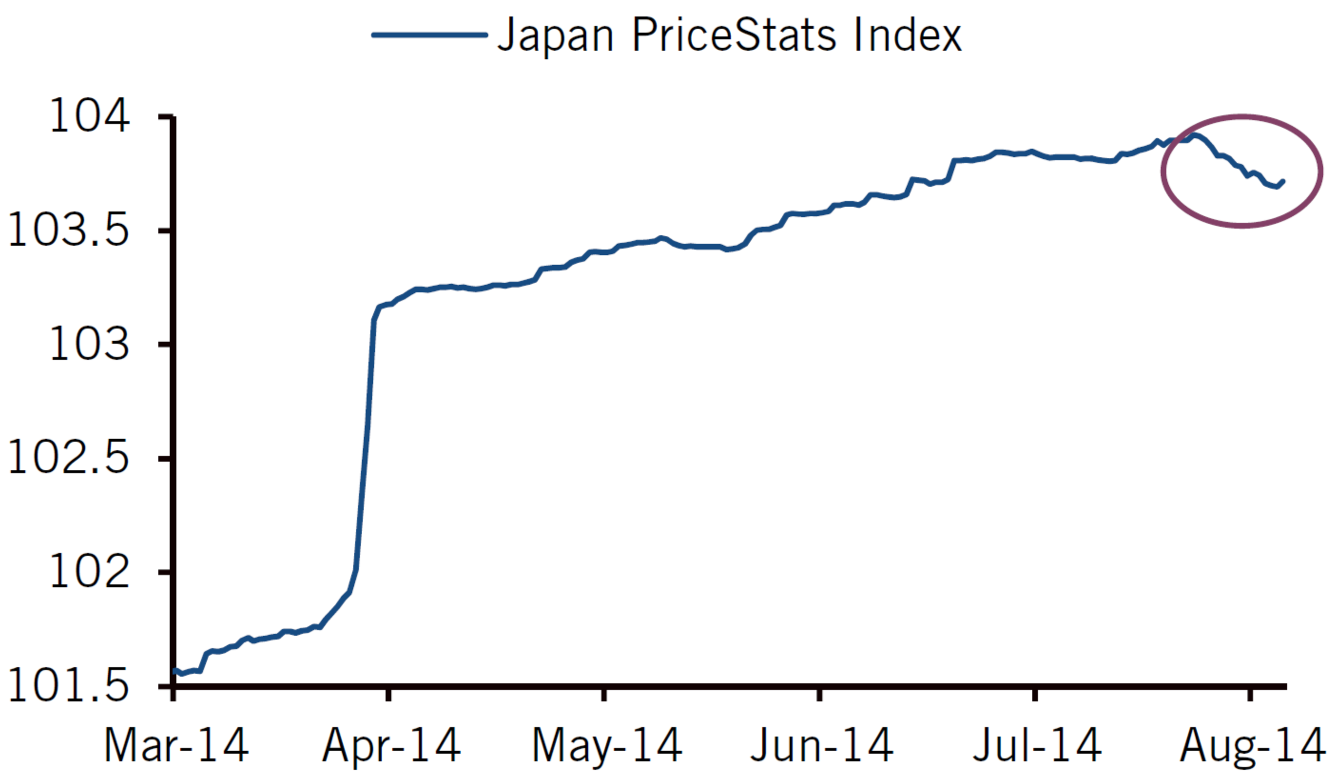
The '97 t
less the s
doctorat
Within a
There are some b
basket, which sh
cent in April to 1
particular brand
brands, weighed
the official index
trading down to
Another importa
durables. As suc
expectations, wh
karaoke booth. T
to influence, he r
Even so, as supe
snapshot of real-
And it could also
which tracks gro
continued, they r
there's been no b
Watanabe.

- GDP、GDPデフレーター、産業生産統計、家庭統計等を見るのではなく、東大日次物価指数を見れば現在の状況を把握することは容易
- 97年の消費増税と2014年の消費増税はほぼ全く同じ動きを辿っており、このままだと物価は下がり続けると予想
- 現在の需給ギャップをリアルタイムで把握するためには、現状よりも優れた手法


GRAB
UTKYDPR Index
96) Actions
97) Edit
G 763 - UTKY 5D MOV AVG
11/06/2010 - 11/03/2014
11) Compare Local CCY
Study Simple MA Period 7 Offset 0
1D 3D 1M 6M YTD 1Y 5Y Max Daily
Security/Study
3.0000
2.0000
1.0000
0.0000
-0.3820
-0.7041
-1.0000
-2.0000
-3.0000
2010 2011 2012 2013 2014
Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2014 Bloomberg Finance L.P.
SN 154081 H636-642-1 06-Nov-14 15:49:12 JST GMT+9:00

同業他社:PriceStats

Figure 1: Running out of steam



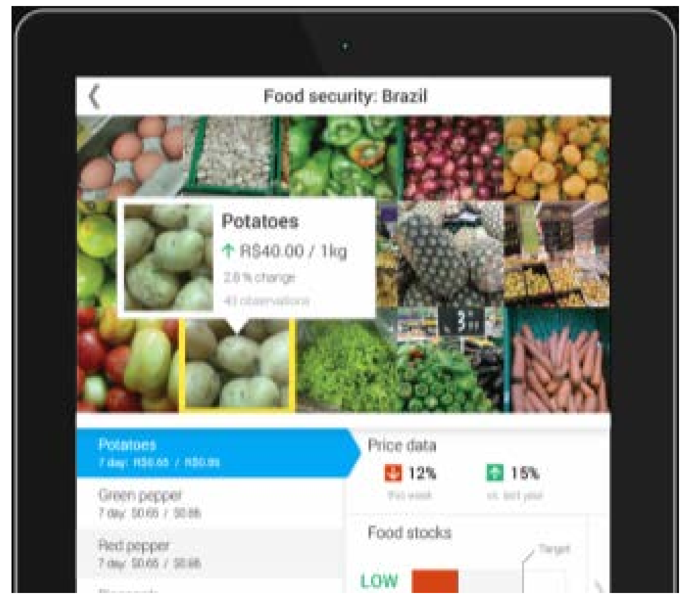
Source: State Street Global Markets

PREMISE

Premise data corp

Overview

- Business: Provider of a platform that offers real–time macroeconomic data for a streaming global economy.
- Service
  - Business for financial services
    - Real time inflation tracking
    - Street level activity gauge
    - Local demand indicators
  - Business for Government, brand and Telecoms
- Founded 2012 January
- Employee: 21
- Financing
  - 1st round in Aug 2012
    - Andreesen Horowits
    - Google Ventures
    - Harrison Metal Capital
    - New Venture Capital
  - 2nd round in Mar 2014 ($11M)
    - All 1st round investors
    - Bowery capital
    - The social capital platform
- Management
  - David Soloff: CEO, Metamarkets Group; Director, Information Products, Rapt
  - Joseph Reisinger: CTO, Fellow, Google; Chief Scientist, Metamarkets; Researcher, IBM
  - Chamath Palihapitiya: Institutional investor, the social capital platform



Use of scanner data by national statistical agencies

Statistics Netherland

Swiss Federal Statistical Office

Statistics Sweden

Australian Bureau of Statistics

Statistics Austria

Statistics Denmark

Statistics Iceland

STATEC, Luxembourg

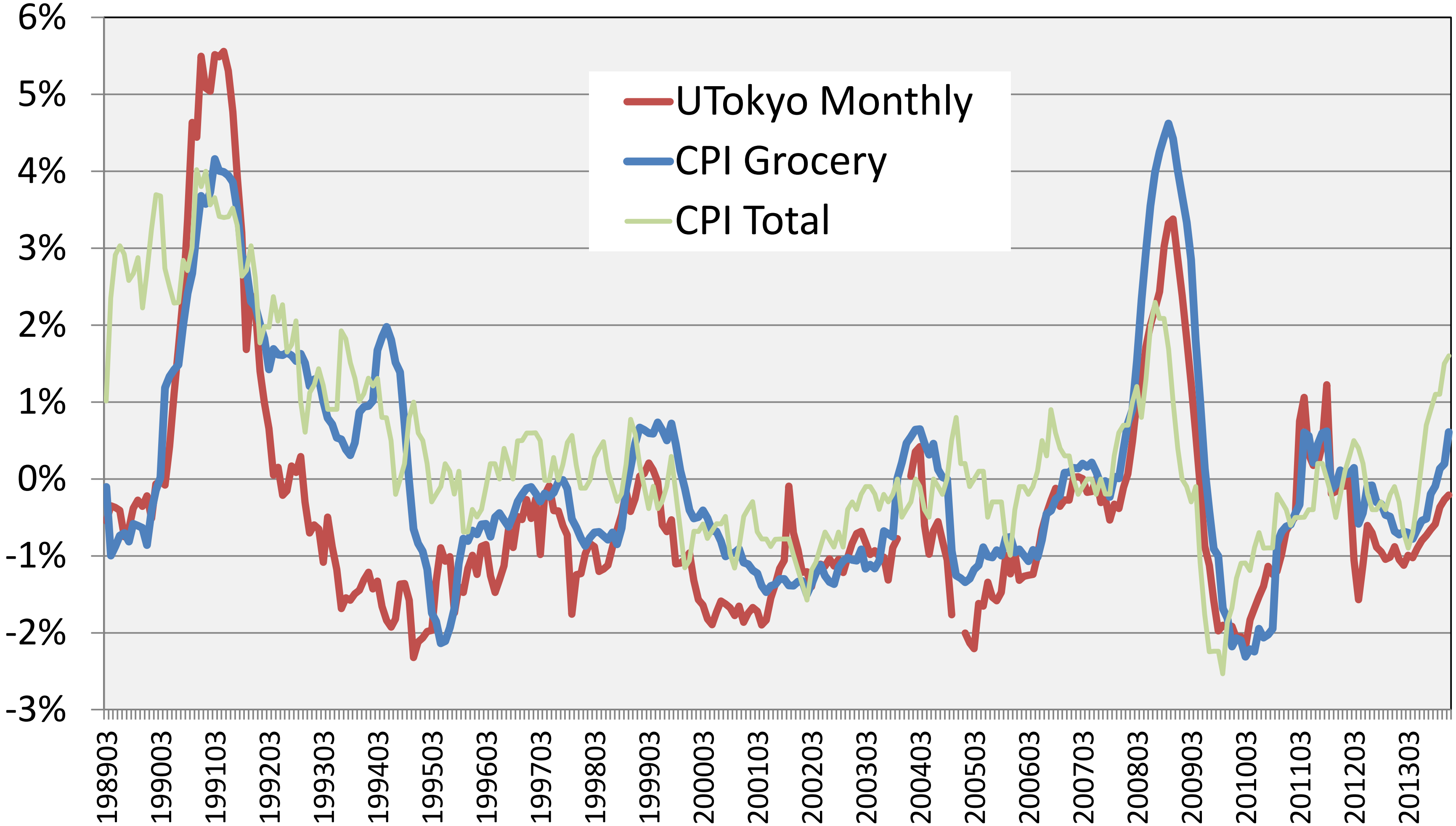
Statistics New Zealand

Federal Statistical Office, Germany

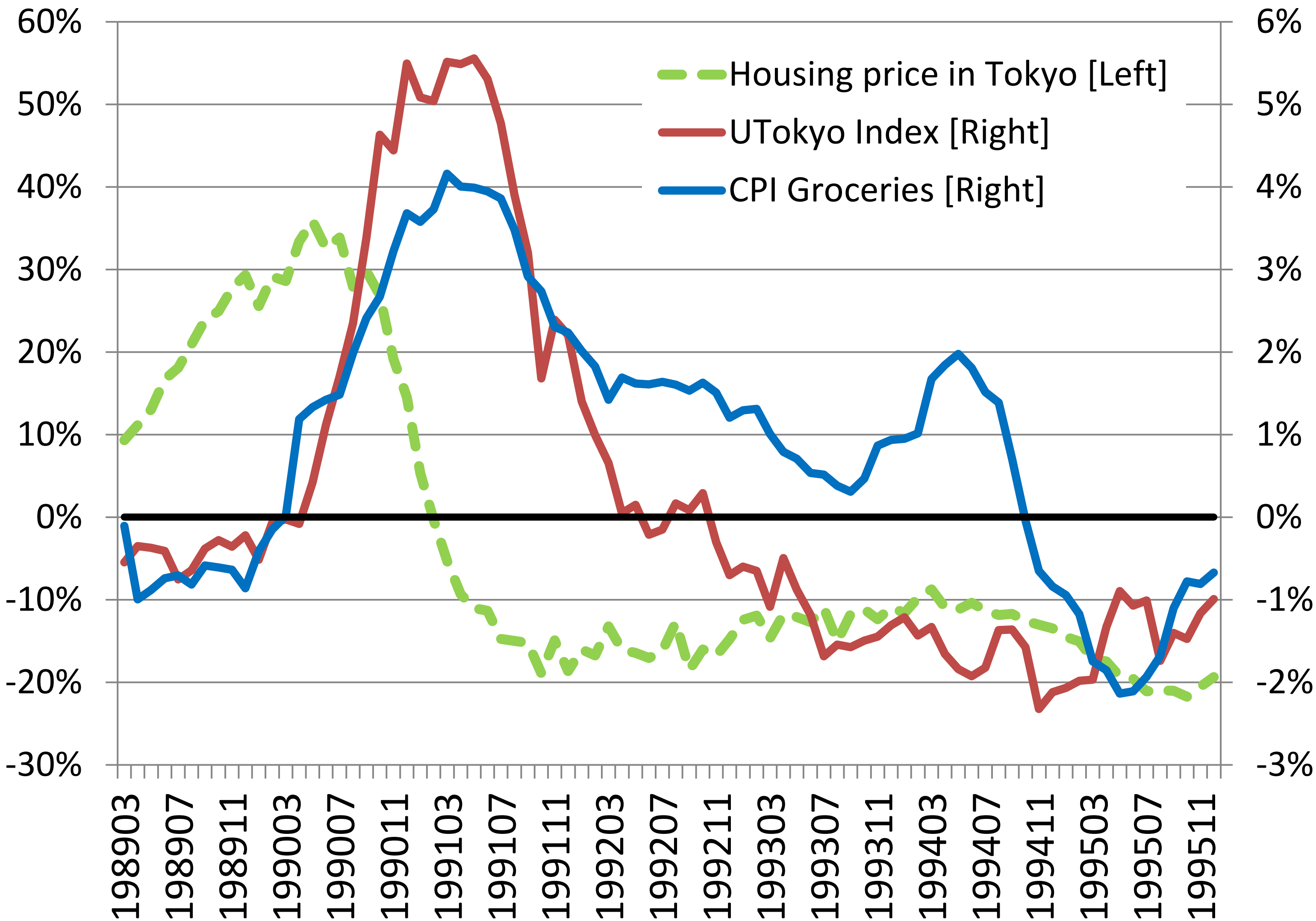
Hungarian Central Statistical Office

Central Bureau of Statistics, Israel

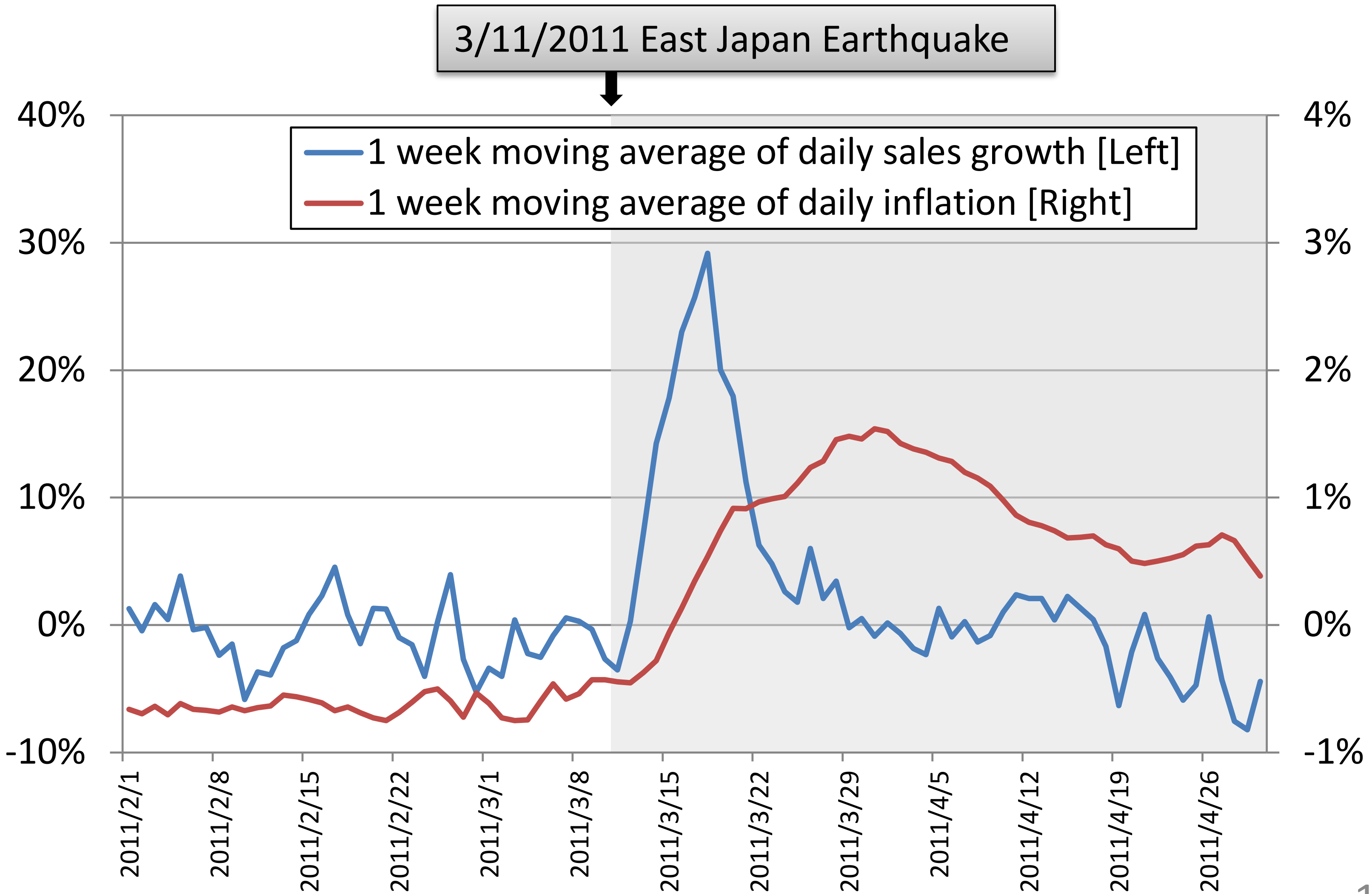
東大指数とCPIの比較



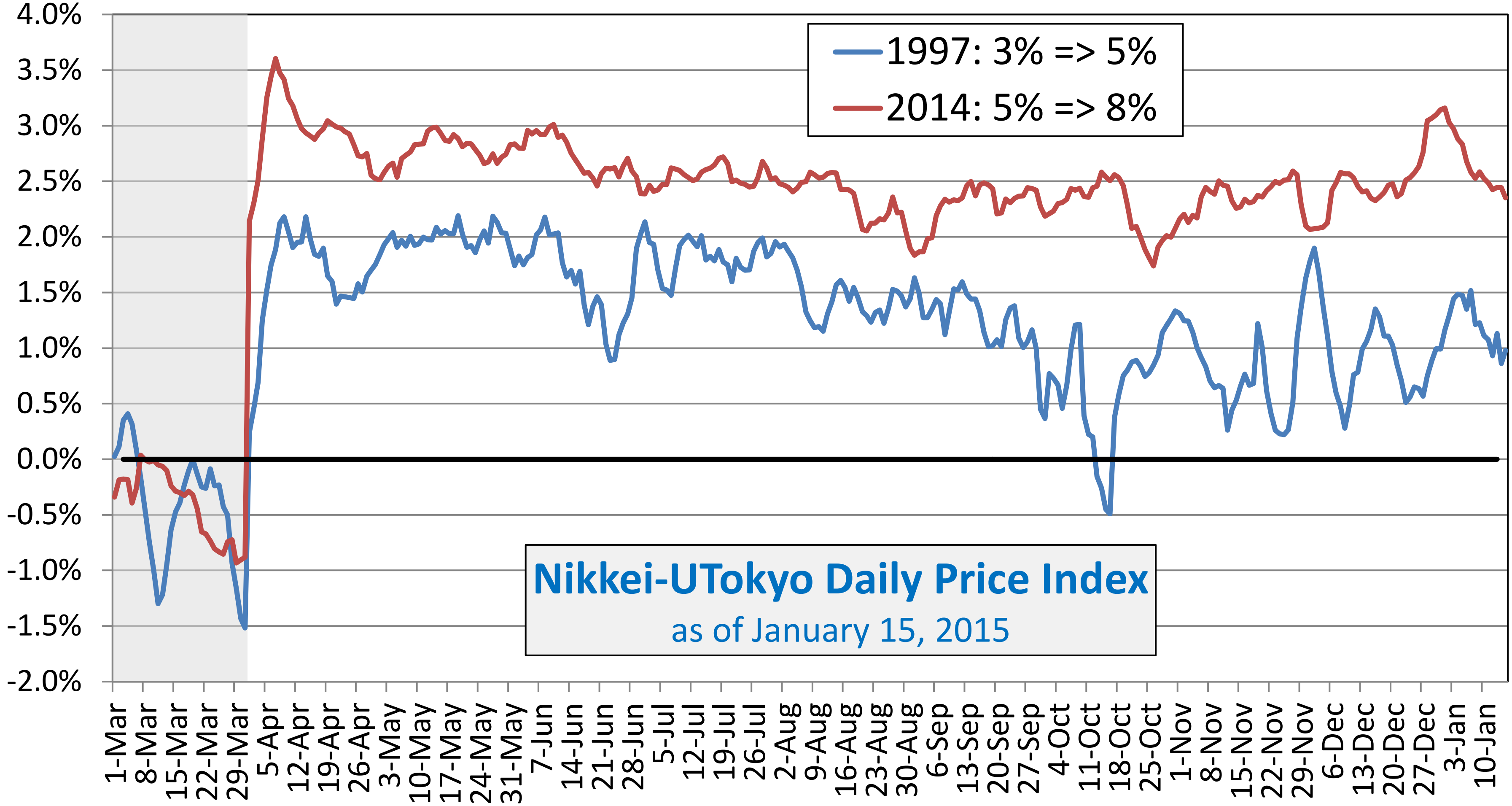
UTokyo Index vs. CPI for 1989-1995



東日本大震災直後の物価と売上

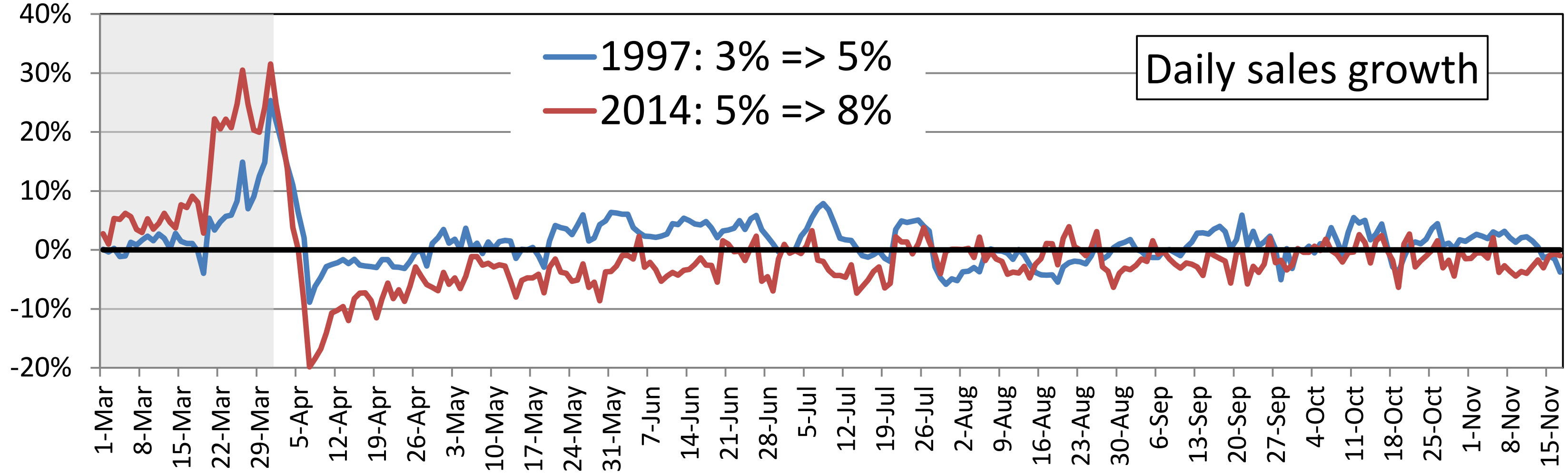


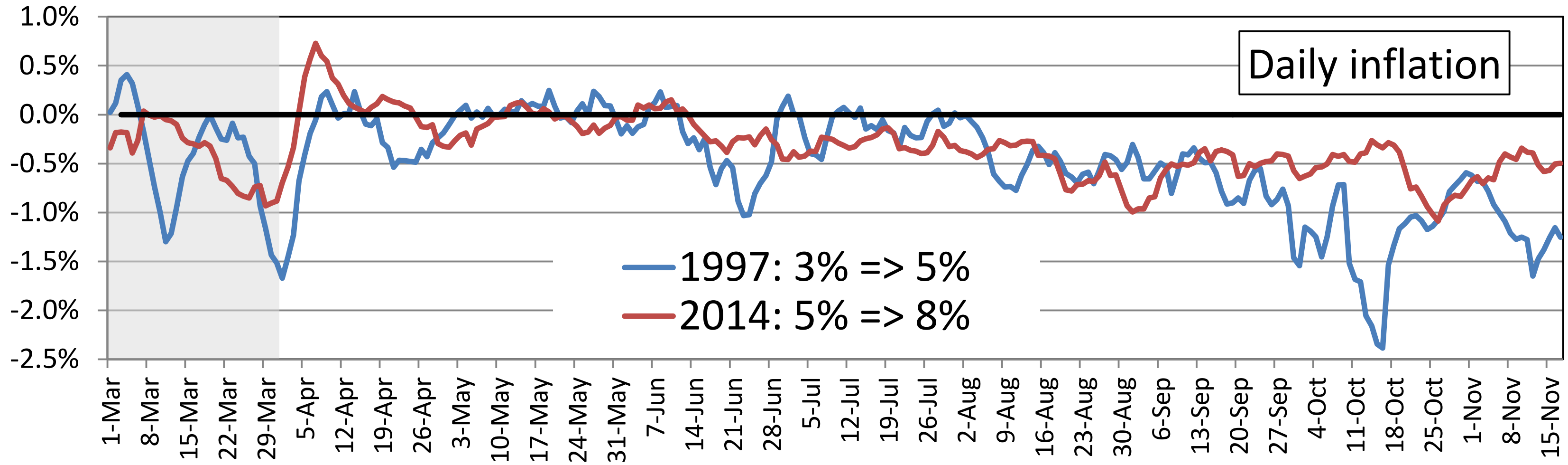
消費税増税後の物価



1997: 3% => 5%
2014: 5% => 8%
Nikkei-UTokyo Daily Price Index
as of January 15, 2015
4.0%
3.5%
3.0%
2.5%
2.0%
1.5%
1.0%
0.5%
0.0%
-0.5%
-1.0%
-1.5%
-2.0%
1-Mar
8-Mar
15-Mar
22-Mar
29-Mar
5-Apr
12-Apr
19-Apr
26-Apr
3-May
10-May
17-May
24-May
31-May
7-Jun
14-Jun
21-Jun
28-Jun
5-Jul
12-Jul
19-Jul
26-Jul
2-Aug
9-Aug
16-Aug
23-Aug
30-Aug
6-Sep
13-Sep
20-Sep
27-Sep
4-Oct
11-Oct
18-Oct
25-Oct
1-Nov
8-Nov
15-Nov
22-Nov
29-Nov
6-Dec
13-Dec
20-Dec
27-Dec
3-Jan
10-Jan

消費税率引き上げ後の売上と物価

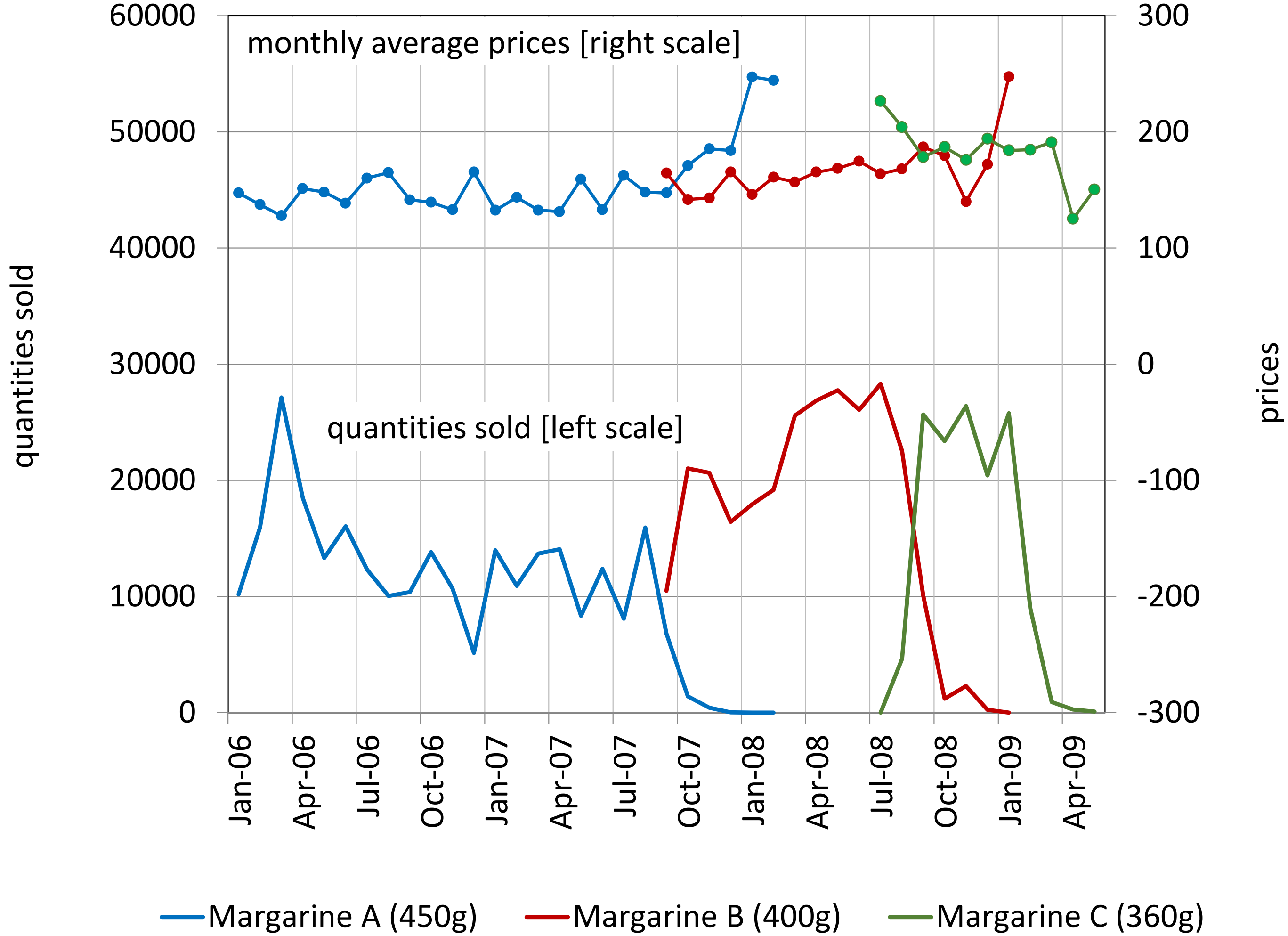




東大指数に関するFAQ

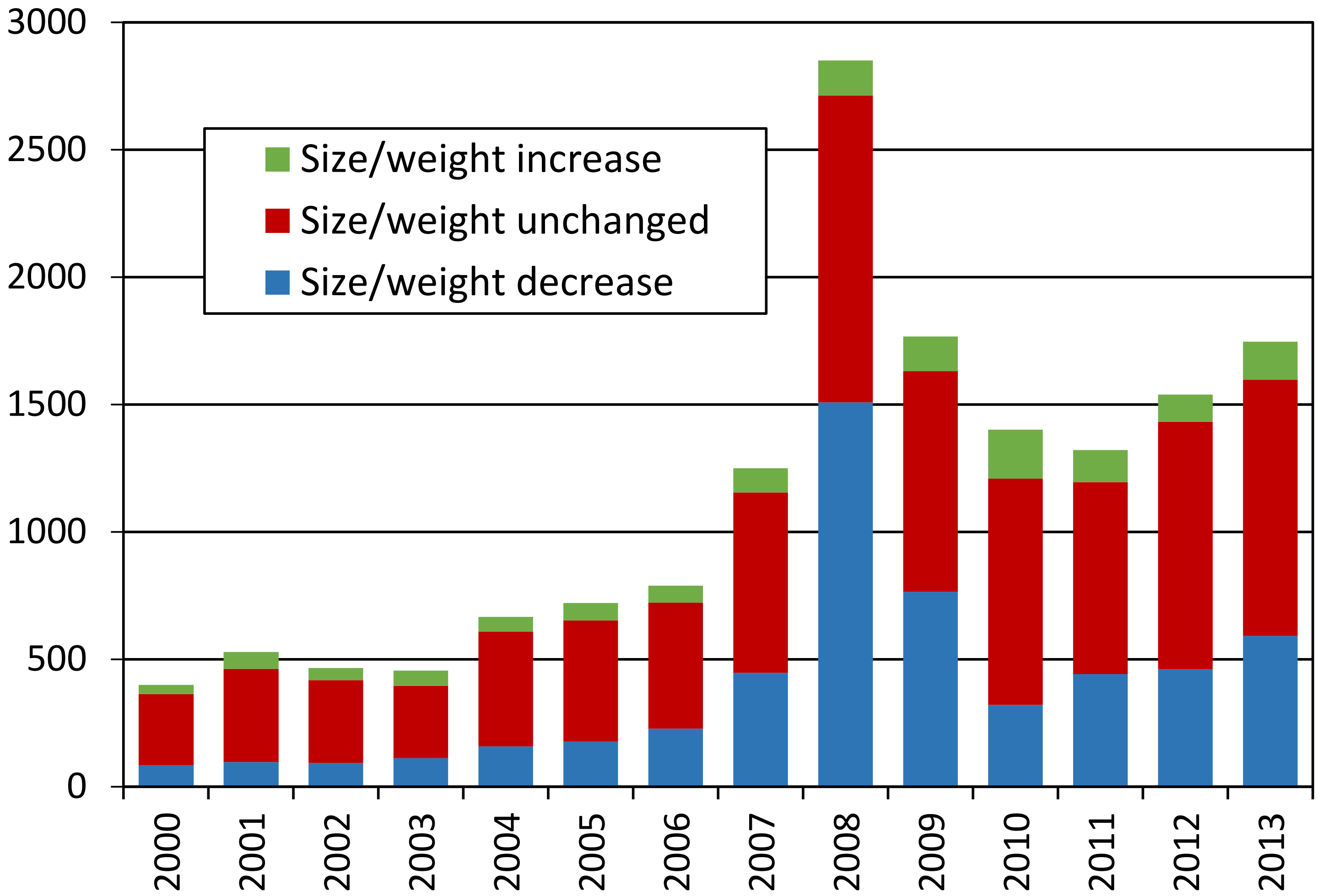
- 東大指数は商品を減量する実質値上げを考慮できているのか。
- 東大指数は新商品をどう扱っているのか。
- 東大指数はカバレッジが低い。問題ではないか。
- 東大指数はCPIをどの程度正確に予測できるのか。

マーガリンの世代交代



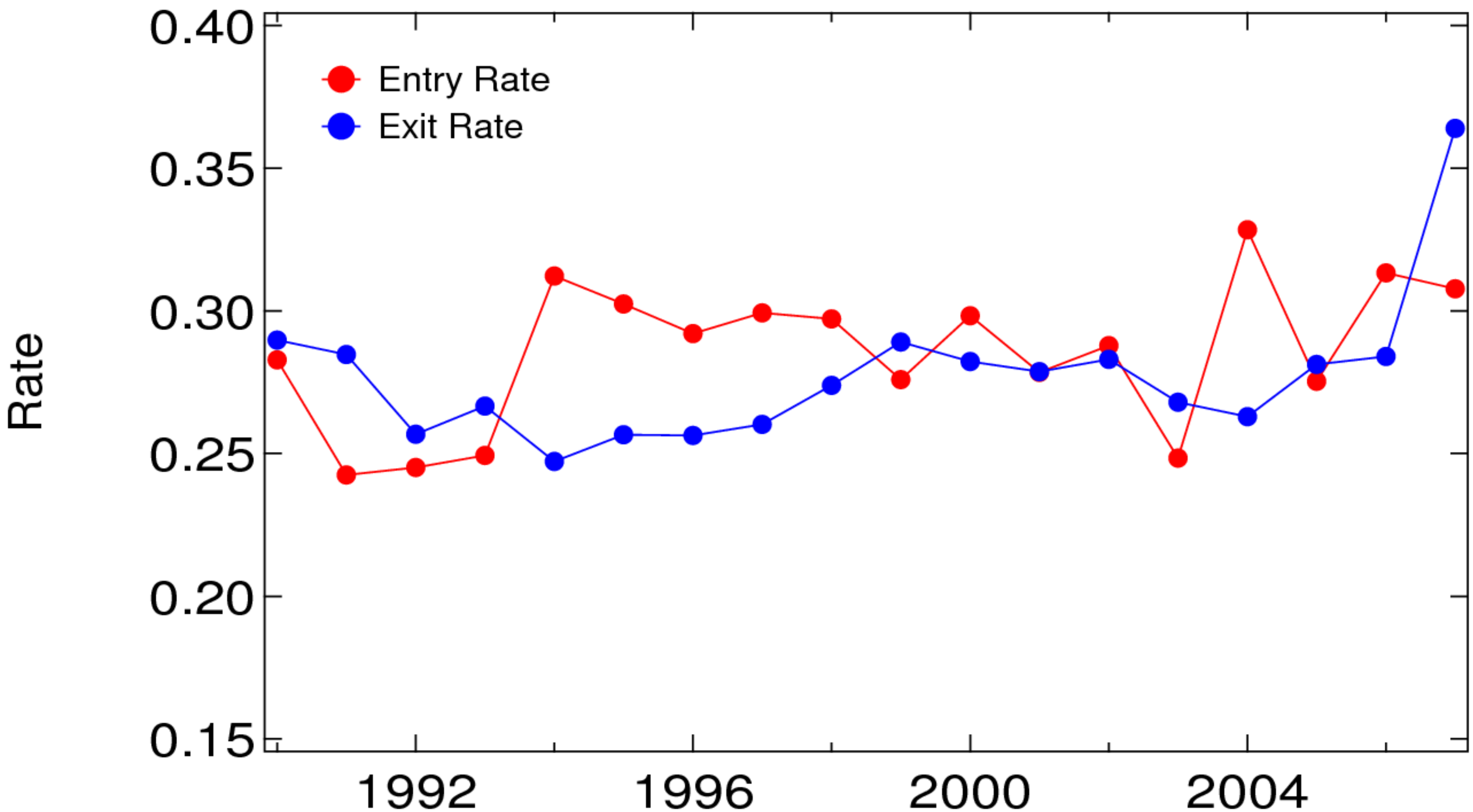
Source: "Product Downsizing and Hidden Price Increases: Evidence from Japan's Deflationary Period." S. Imai, T. Watanabe, Asian Economic Policy Review, Volume 9, Issue 1, 2014, 69-89.

商品の減量・増量の数



Source: “Product Downsizing and Hidden Price Increases: Evidence from Japan's Deflationary Period” (with S. Imai), Asian Economic Policy Review, Volume 9, Issue 1, 2014, 69-89.

商品の参入・退出



Entry Rate
Exit Rate
Rate
0.40
0.35
0.30
0.25
0.20
0.15
1992
1996
2000
2004

東大指数に関するFAQ

- 東大指数は商品を減量する実質値上げを考慮できているのか。
- 東大指数は新商品をどう扱っているのか。
- 東大指数はカバレッジが低い。問題ではないか。
- 東大指数はCPIをどの程度正確に予測できるのか。

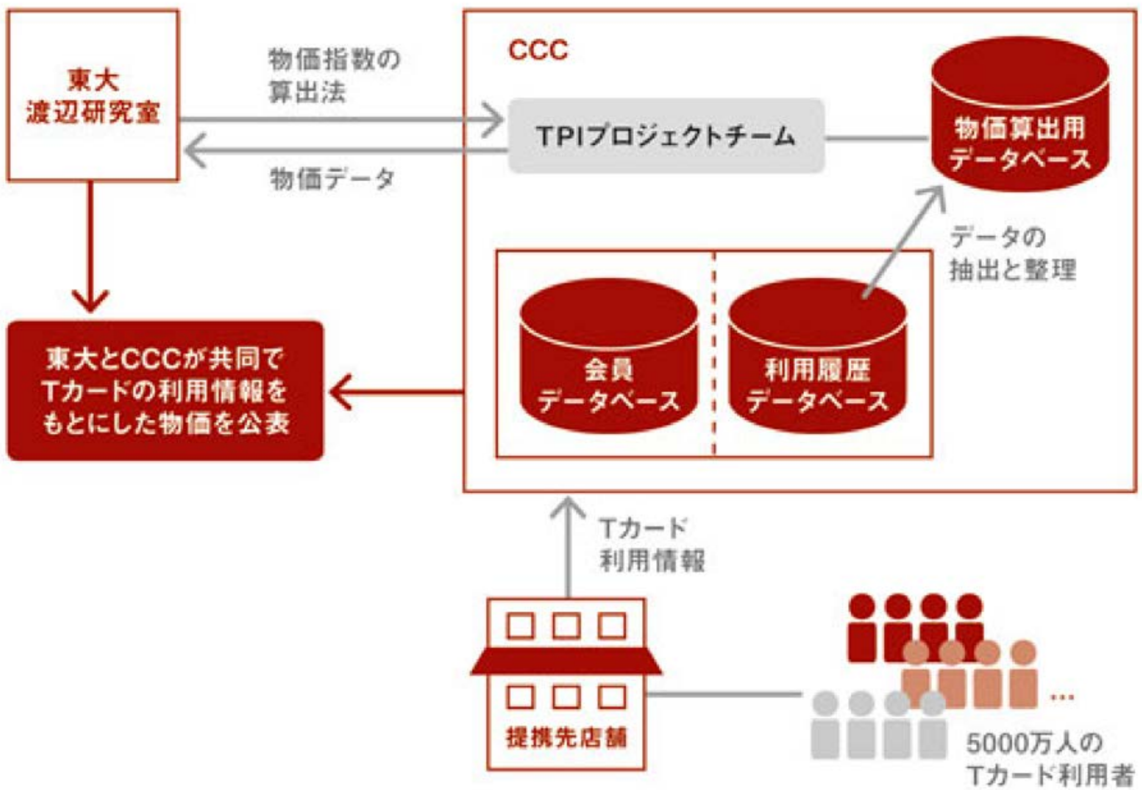
データ活用で成長戦略を描く
日経 NIKKEI BigData ビッグデータ

Data Market | データ市場動向

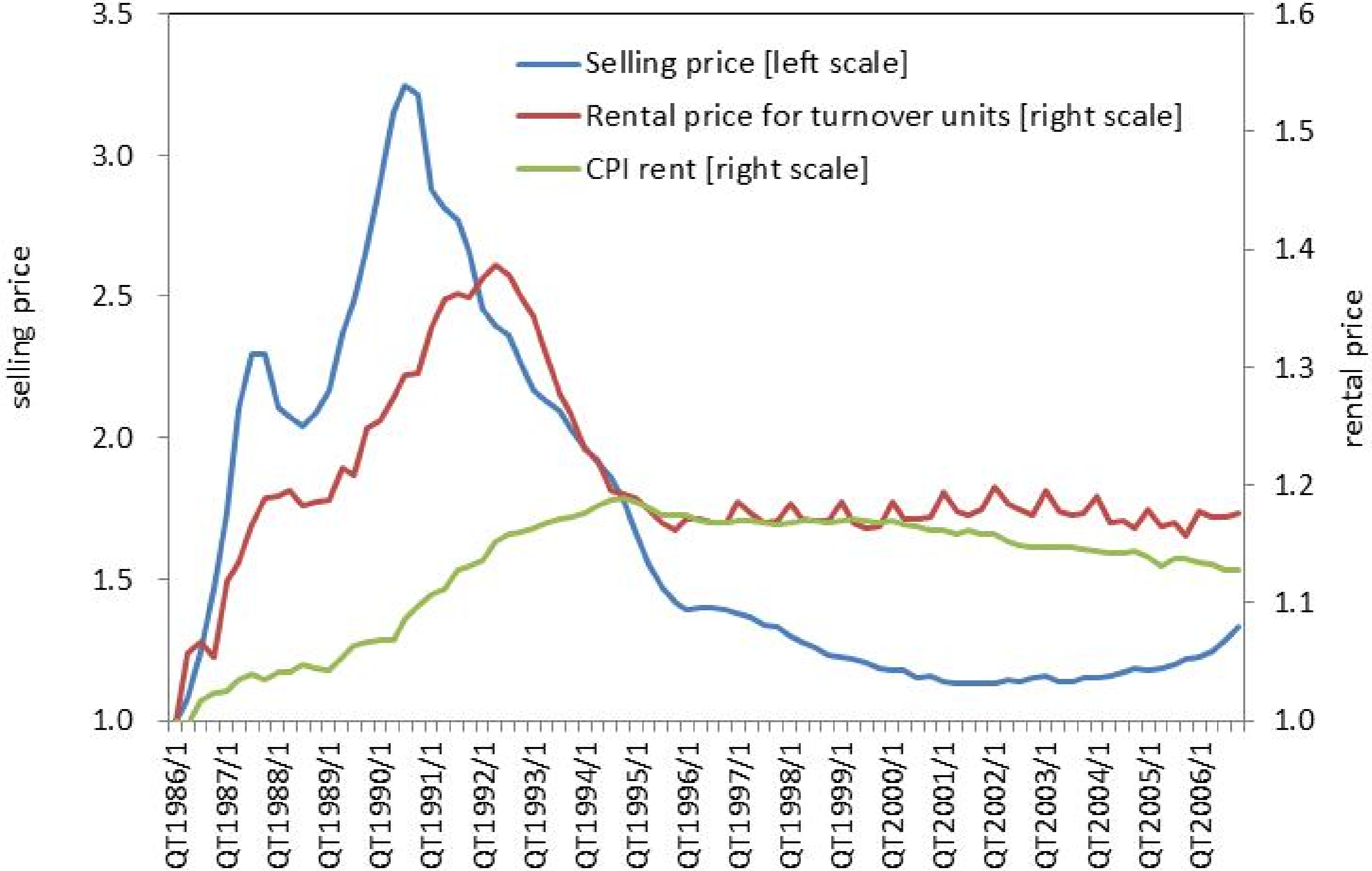
データ市場

Tポイントから物価指数を算出、東大とCCCがカード利用情報を使い共同開発

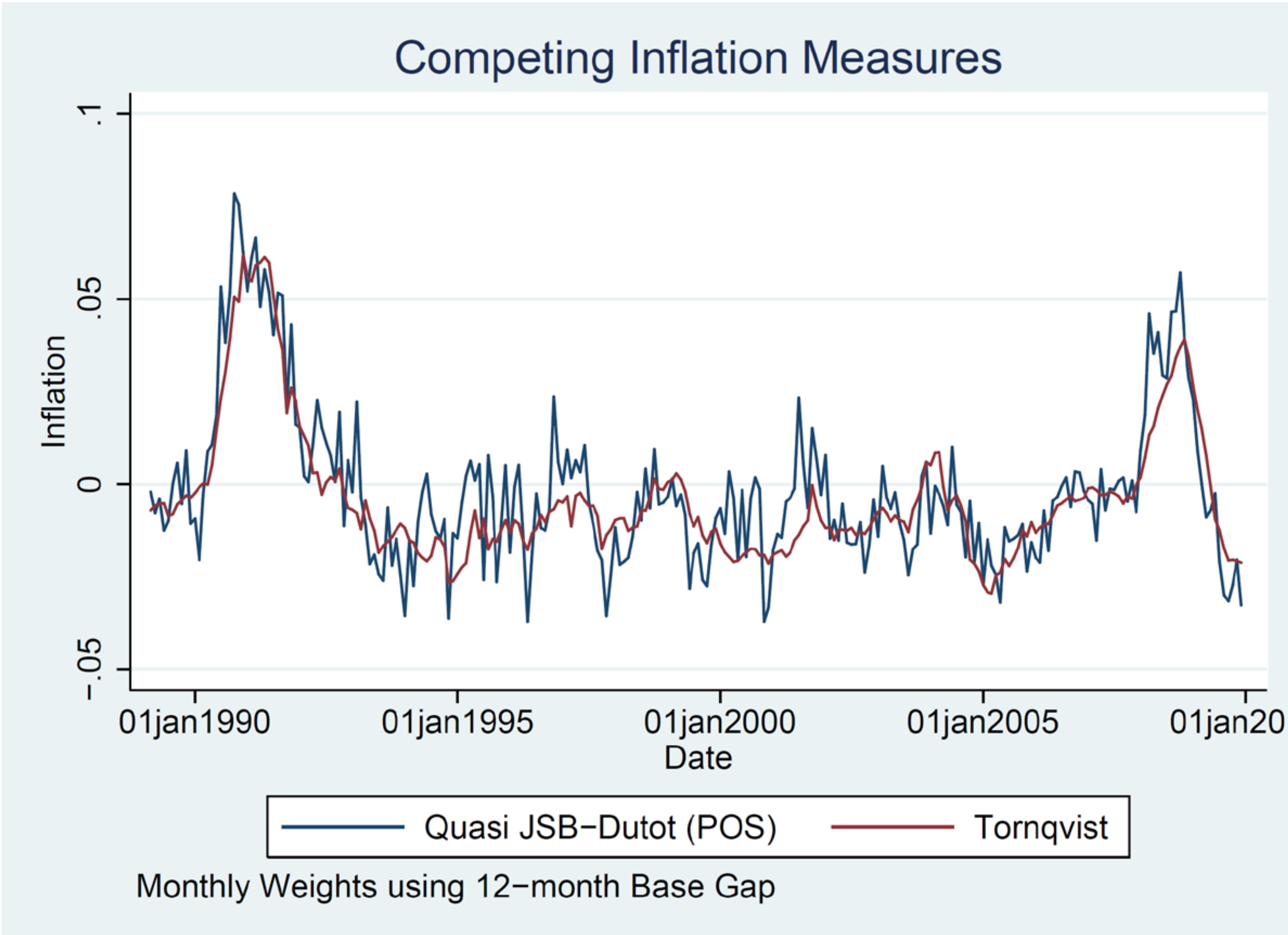
2014.12.19　加古川 群司



Hedonic estimates of selling and rental prices of residential properties



Source: "Residential Rents and Price Rigidity: Micro Structure and Macro Consequences." C. Shimizu, K.G. Nishimura, T. Watanabe, Journal of the Japanese and International Economies, Volume 24, Issue 2, June 2010, 282-299.



Source: “How Much Do Official Price Indexes Tell Us about Inflation?” J. Handbury, T. Watanabe, D. E. Weinstein, NBER Working Paper No.19504, October 2013.

経済ナウキャスト指標としての東大物価指数

- 「今この瞬間の経済の体温を測る」
  - これが東大指数の原点
  - 株価や為替と同じ頻度で物価を観察したい。
  - 高頻度で物価を知ることにより消費者や企業の行動も変わるはず。

- 将来予測（Forecasting）ではなく足元の予測（Nowcasting）
  - 将来予測には誤差がつきもの。いくら手法を洗練させても無視できない誤差が残る。
  - 足元の予測には失敗がない。ビッグデータを用いて緻密に計算すればするほど精度が上がる。
  - 参照価格（reference price）を提供したい。消費者，企業，政府の迅速で正確な意思決定に資する指標を提供していきたい。

消費関連統計の比較

宇南山　卓
（財務総合政策研究所）

目的

- 消費の動向は日本経済にとって重要
  - 貯蓄の動向を明らかにするため
  - 消費刺激策の政策効果を把握するため
- ミクロデータで消費を把握する必要
  - 集計された消費では家計の行動が把握できない
  - 消費の格差などの分布の情報が必要
- ミクロ統計間での整合性の確認
  - 分析結果の信頼性の評価
  - 統計ごとのクセを把握して適正利用の促進

消費関連統計の概要

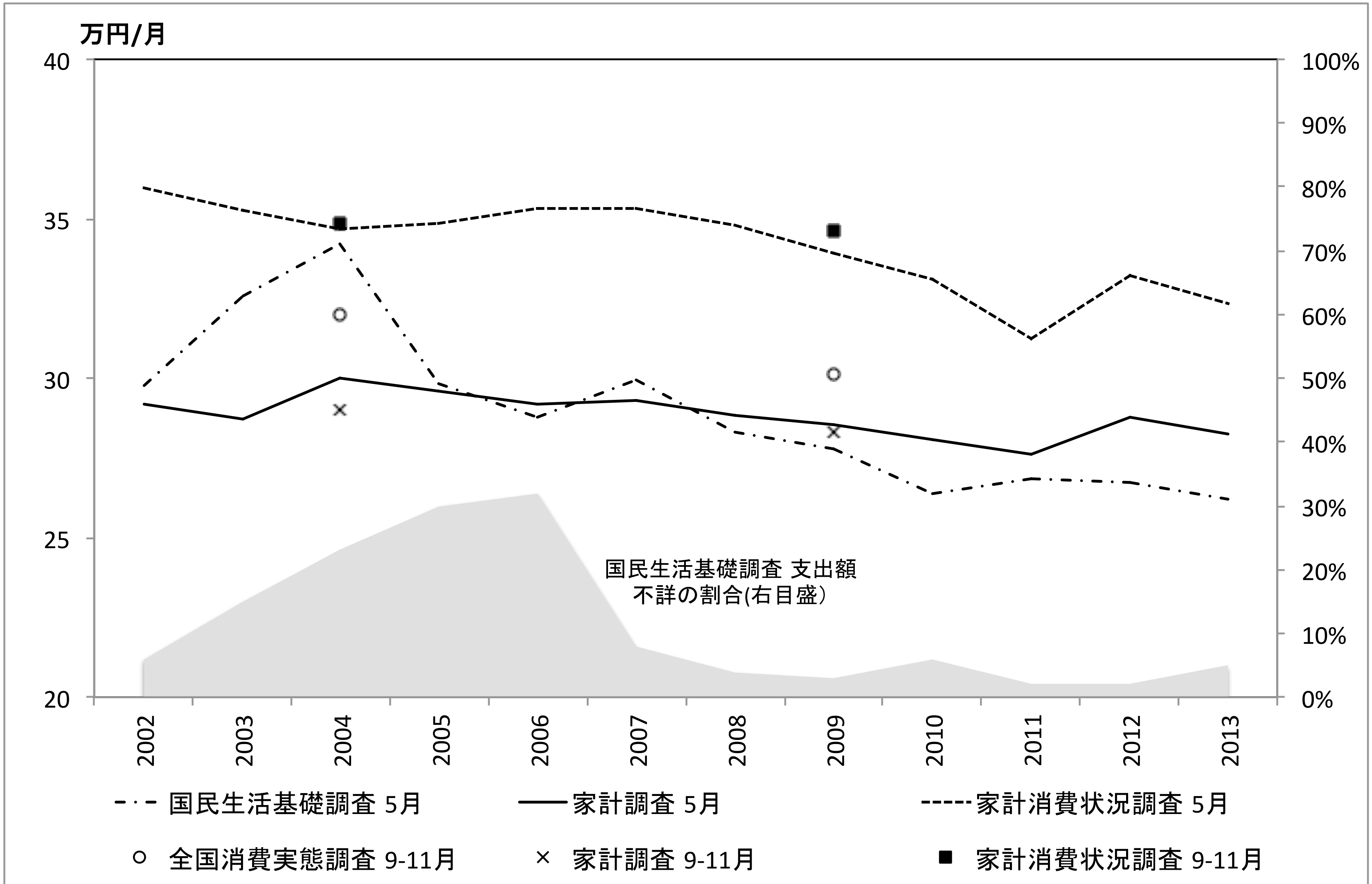
	家計調査	全国消費実態調査	国民生活基礎調査	家計消費状況調査	定額給付金の調査*
頻度	月次	5年ごと	年次（3年毎大規模）	月次（2002年以降）	特別調査（2009年）
調査月	1–12月	9–11月（3ヶ月平均）	6月（調査対象は5月）	1–12月	3–9月
サンプル数	約9千世帯/月 うち単身700	約5.7万世帯 うち単身4000	約5.5万世帯(大規模年は約30万世帯)	3万世帯/月 うち単身3000 (有効約1.9万)	15,000世帯(9,194世帯)
調査方法	家計簿	家計簿	総額記入(万円)	総額記入(円)	階級値選択(5万円毎)
調査項目	全個別支出（消費総額を500品目程度に分類）	全個別支出（消費総額を500品目程度に分類）	消費総額 うち仕送り金	消費総額 高額品目 仕送り 贈与金	総額 定額給付金での購入

*「定額給付金に関連した消費等に関する調査」（内閣府）

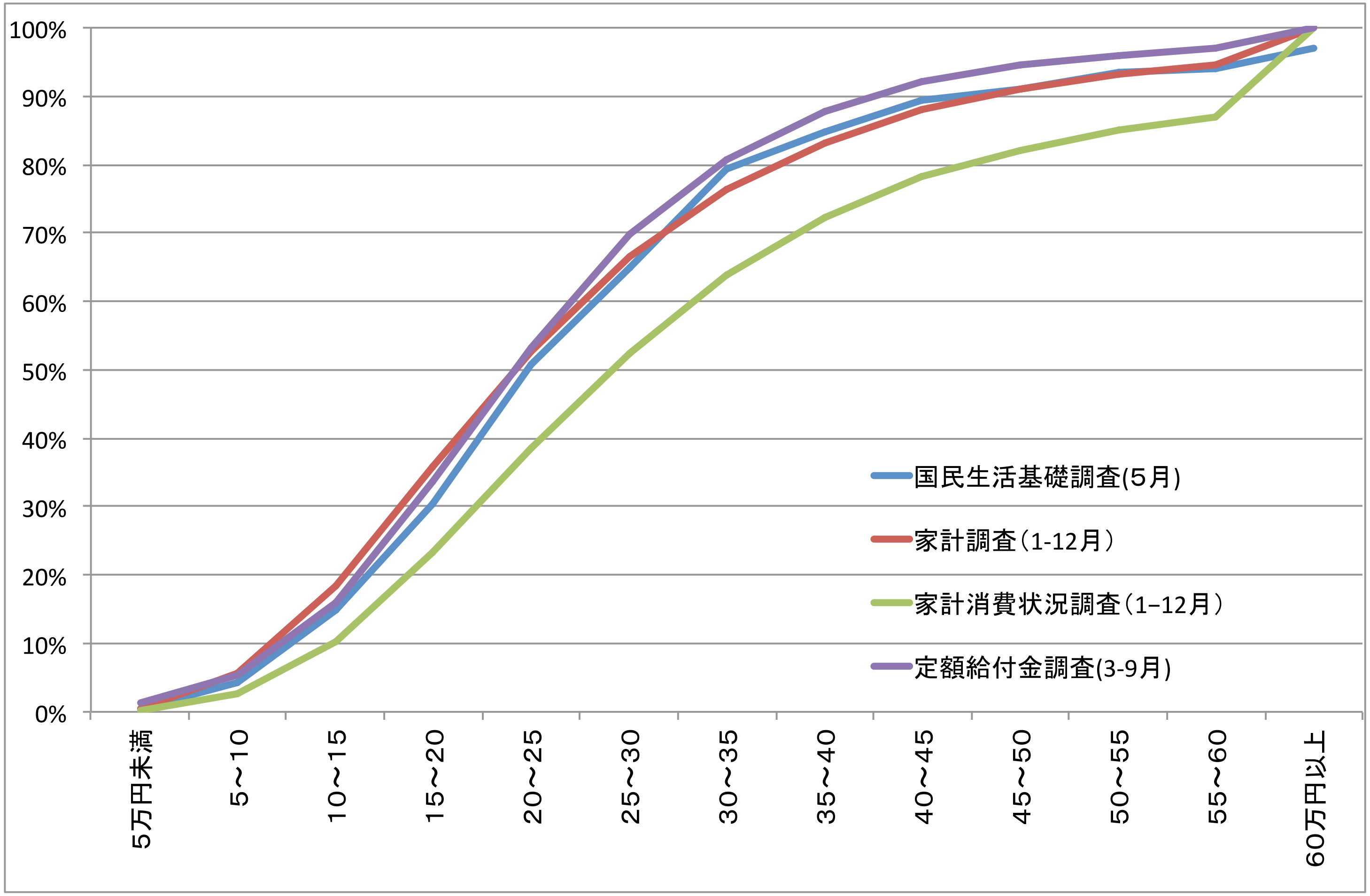
「消費総額」の範囲

- 家計消費状況調査
  - 「毎日の世帯全体の支出金額を合計した、今月1か月間(1日～月末)の支出総額(消費税込み)」
  - 仕送り金・贈与金は別途調査
- 国民生活基礎調査
  - 定義:世帯の方全員の支出金額の合計額
  - (含まないもの)税金、社会保険料、事業上の支払い(農家における肥料や農具、商店における商品の仕入れに使った金等)、貯蓄、借金や住宅 ローンなどの返済、掛け捨て型以外の生命保険料・損害保険料
- 定額給付金に関する調査
  - この総支出金額は、食料品・日用品購入、被服費、光熱水道代、交際費、塾・習い事の月謝、家賃など、家計のために支出した金額すべてを含みます(ただし住宅ローンの返済分は除き ます)。

各統計での消費の推移

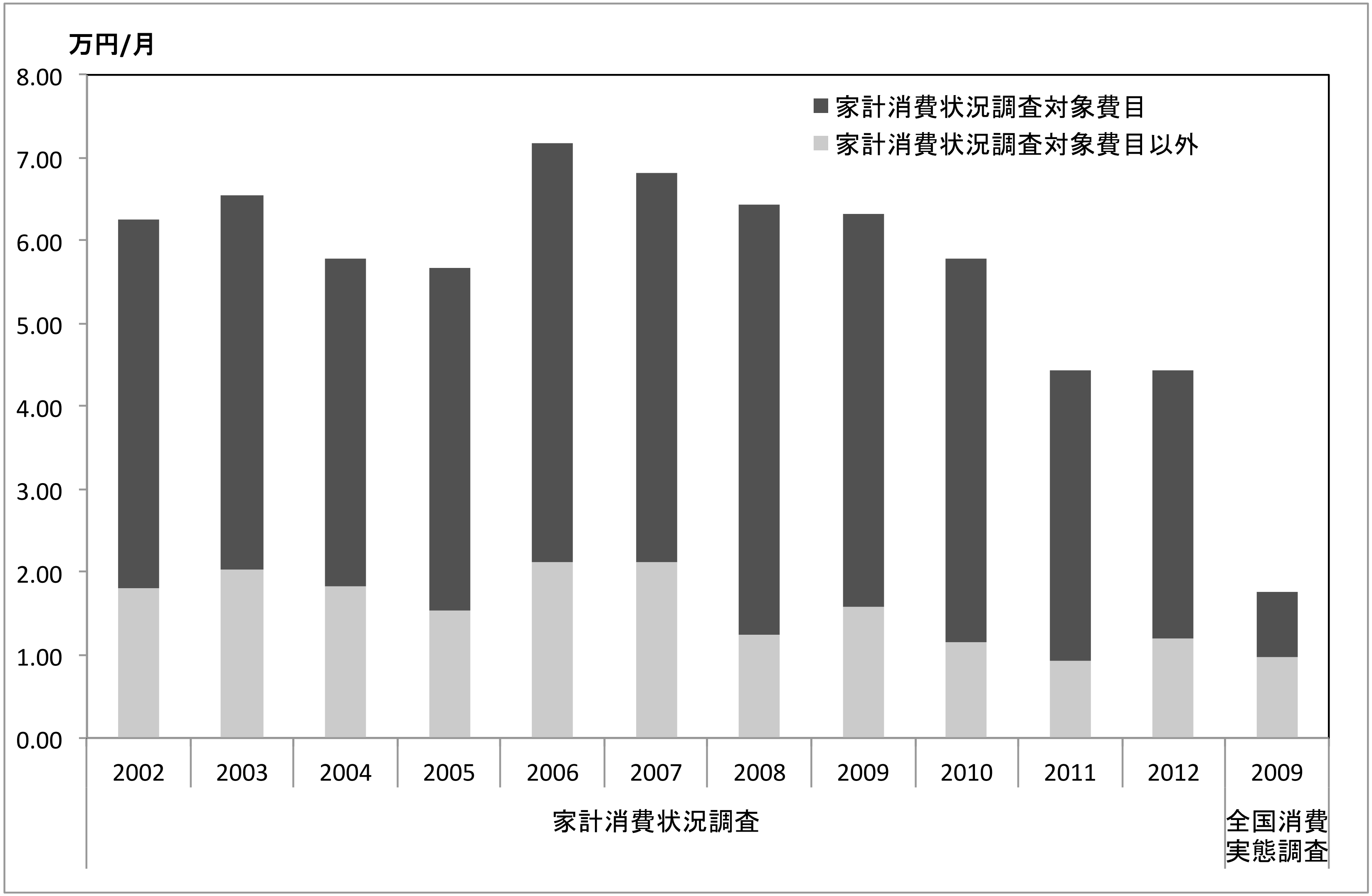

万円/月
40
35
30
25
20
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
国民生活基礎調査 支出額
不詳の割合(右目盛）
国民生活基礎調査 5月
家計調査 5月
家計消費状況調査 5月
全国消費実態調査 9-11月
家計調査 9-11月
家計消費状況調査 9-11月

消費水準の分布

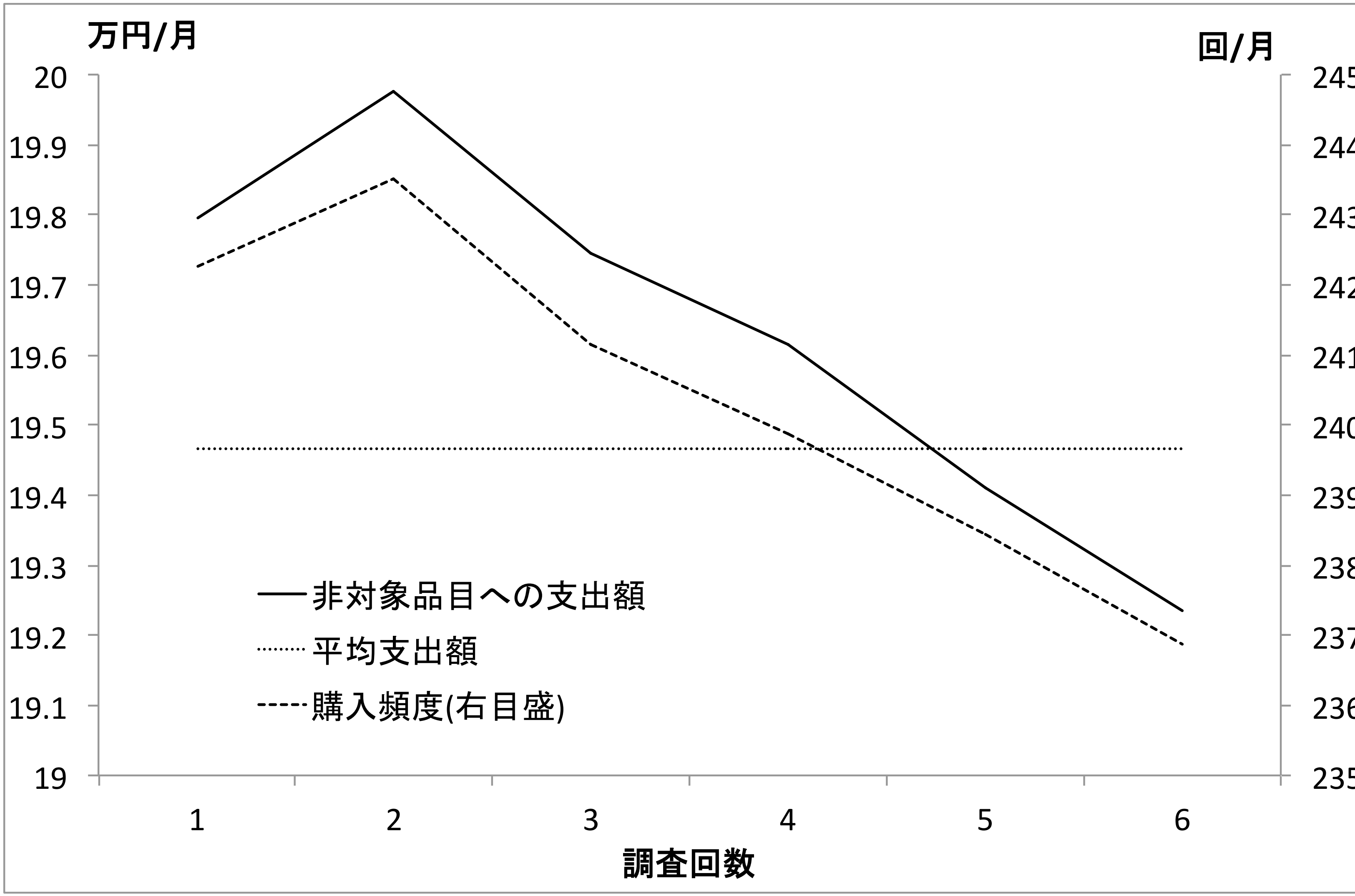


*各調査2009年の消費支出総額

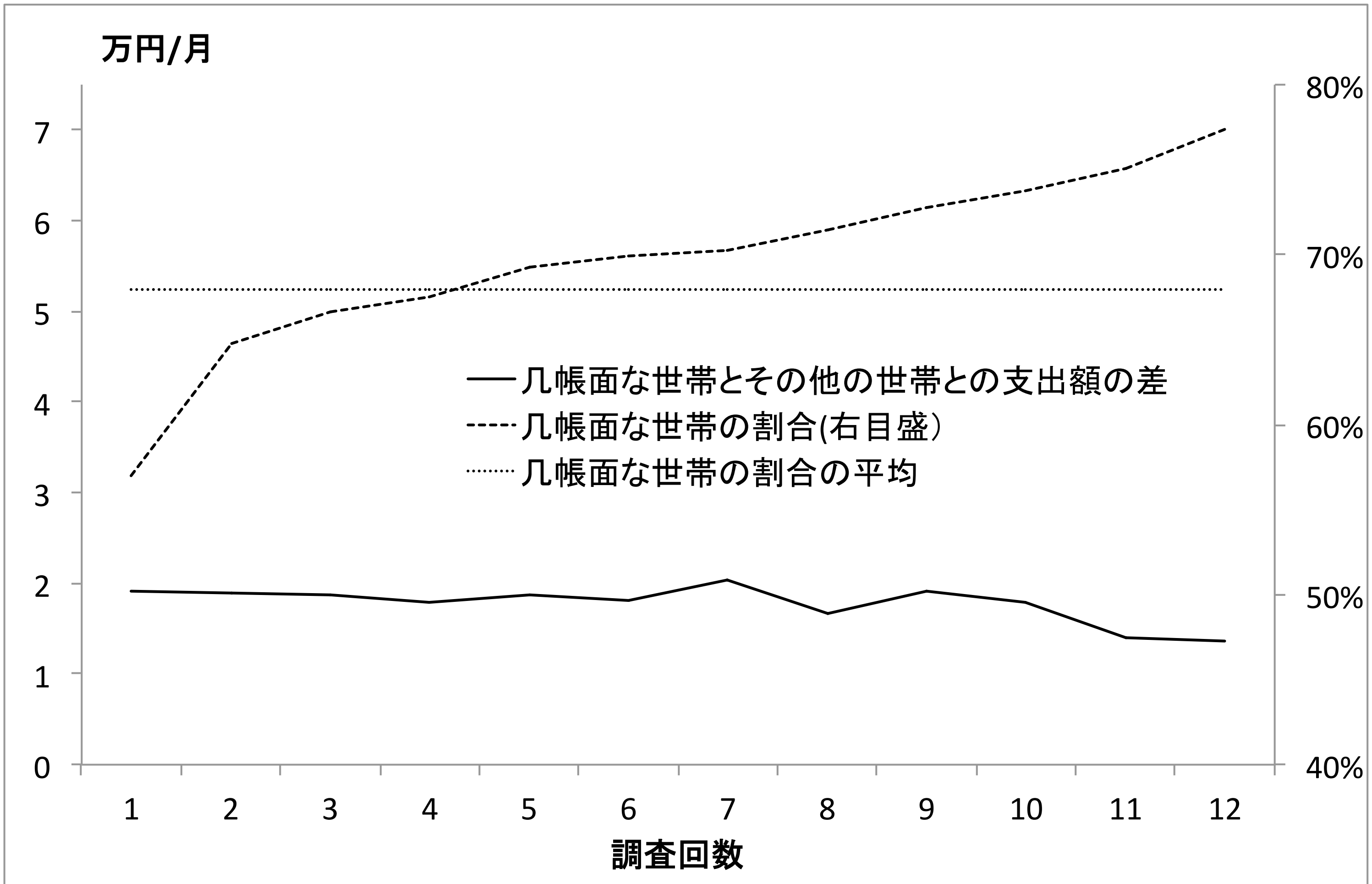
家計調査と家計消費状況調査の差



家計簿の記入負担と消費支出

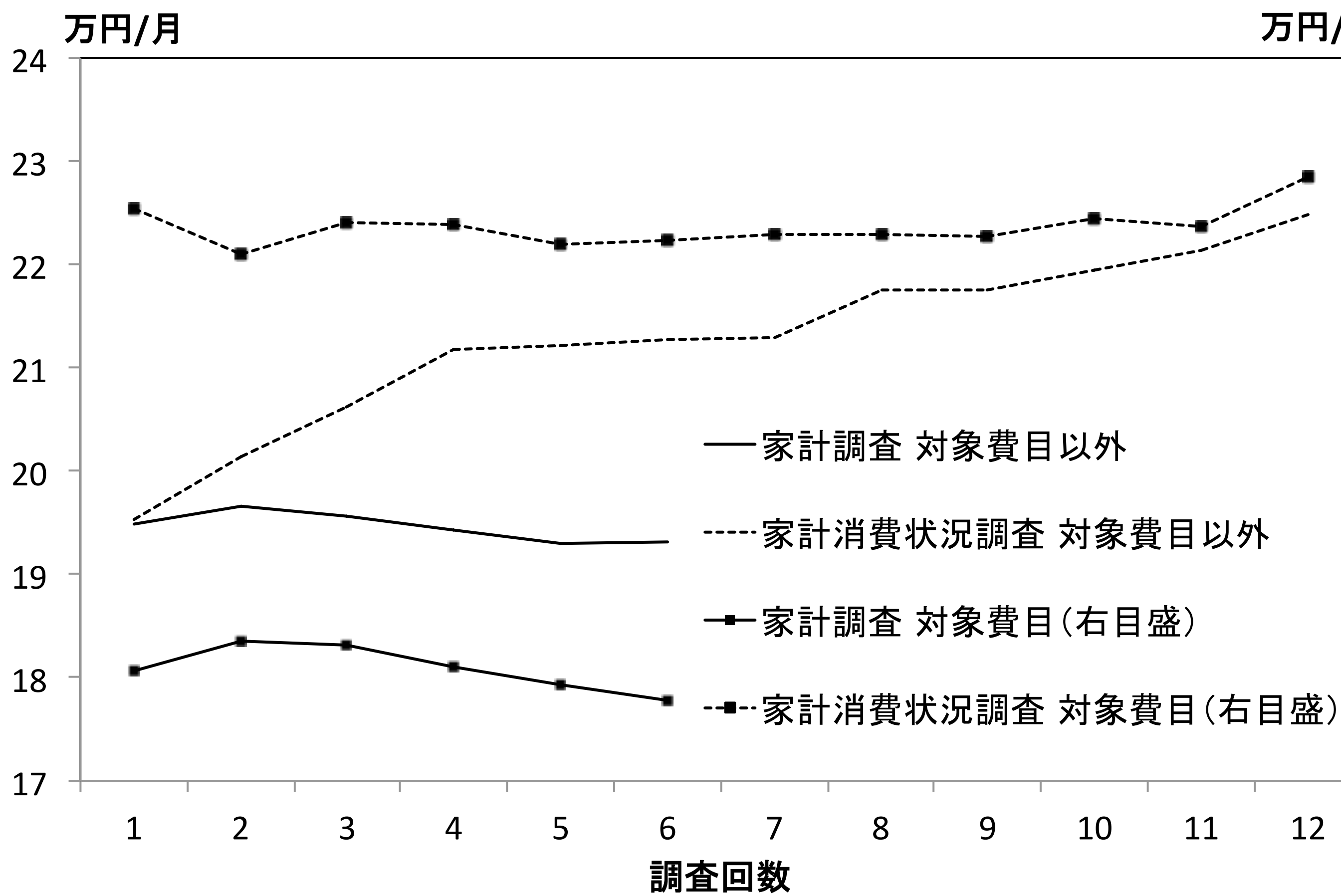


家計消費状況調査の上昇理由:世帯の几帳面さと消費

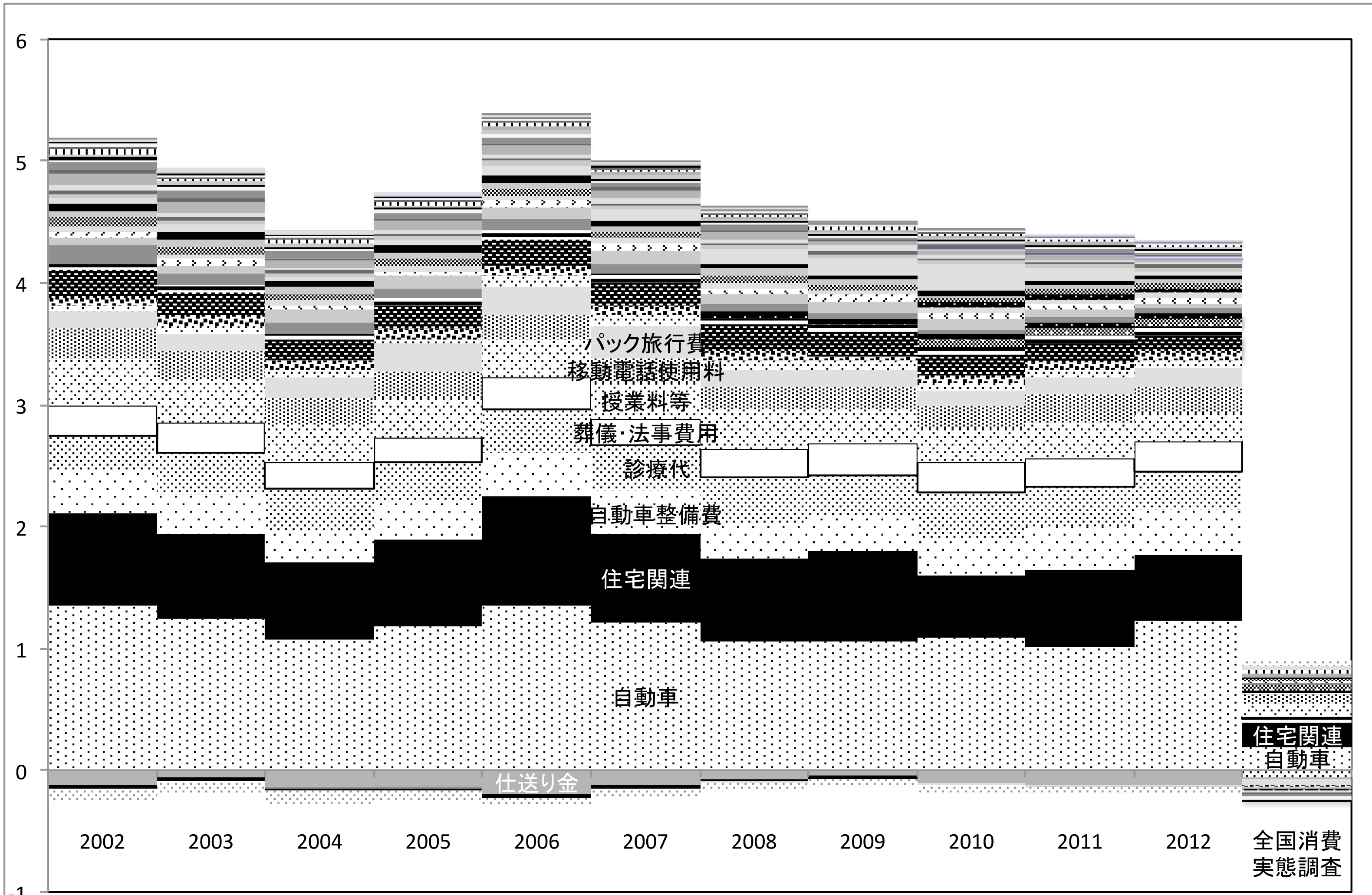


*几帳面な世帯とは、消費総額を万円以下の単位まで回答している世帯。

調査の継続と消費支出の差



家計調査と家計消費状況調査：差を生む品目

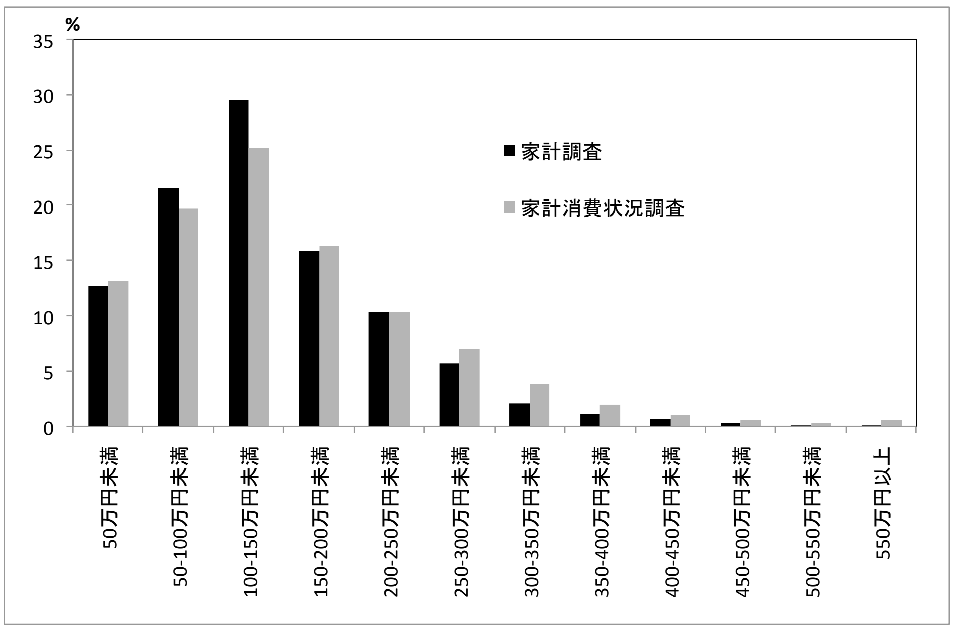


家計調査と家計消費状況調査：購入頻度と購入単価

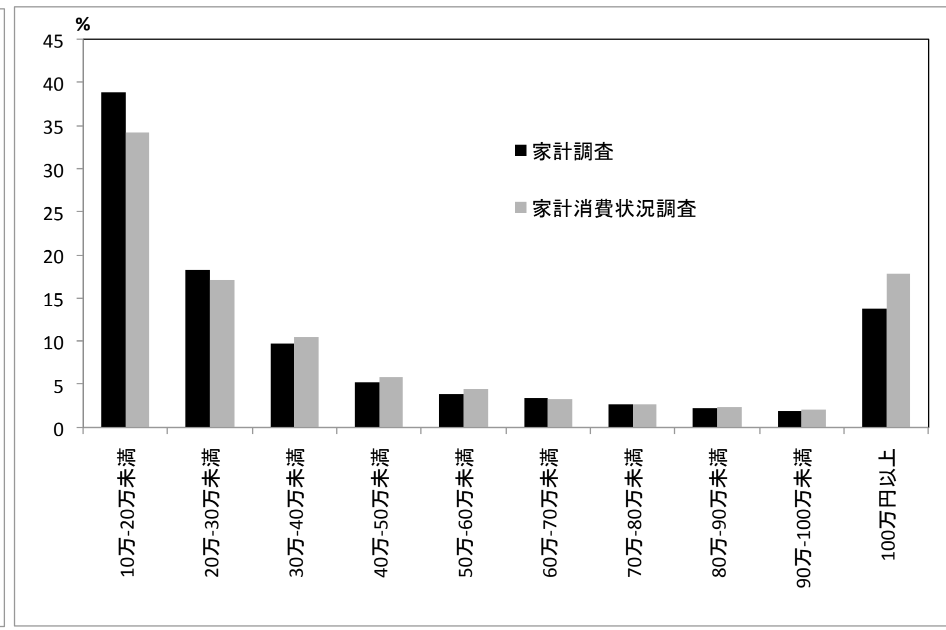
	家計調査			家計消費状況調査			比率		
	支出額	単価	支出者の割合	支出額	単価	支出者の割合	支出額	単価	支出者の割合
移動電話使用料	6,978	11,278	61.87%	9,178	12,203	75.22%	1.32	1.08	1.22
インターネット・放送受信料	3,596	6,936	51.84%	2,852	5,715	49.90%	0.79	0.82	0.96
自動車	5,394	1,380,000	0.39%	17,233	1,520,000	1.13%	3.19	1.10	2.90
自動車整備費	1,456	24,688	5.90%	4,721	56,050	8.42%	3.24	2.27	1.43
住宅関連	6,741	110,000	6.13%	13,504	181,000	7.48%	2.00	1.65	1.22
家賃	10,423	52,947	19.69%	9,372	53,912	17.38%	0.90	1.02	0.88
診療代	7,366	11,111	66.29%	10,656	15,743	67.69%	1.45	1.42	1.02
授業料等	9,525	40,505	23.52%	12,614	78,637	16.04%	1.32	1.94	0.68
パック旅行費	4,676	37,850	12.35%	6,363	88,042	7.23%	1.36	2.33	0.59
挙式・披露宴費用	592	938,000	0.06%	2,424	973,000	0.25%	4.10	1.04	4.17
葬儀・法事費用	1,506	241,000	0.63%	3,886	395,000	0.98%	2.58	1.64	1.56
信仰関係費	1,879	8,109	23.17%	2,657	30,833	8.62%	1.41	3.80	0.37

購入単価の妥当性

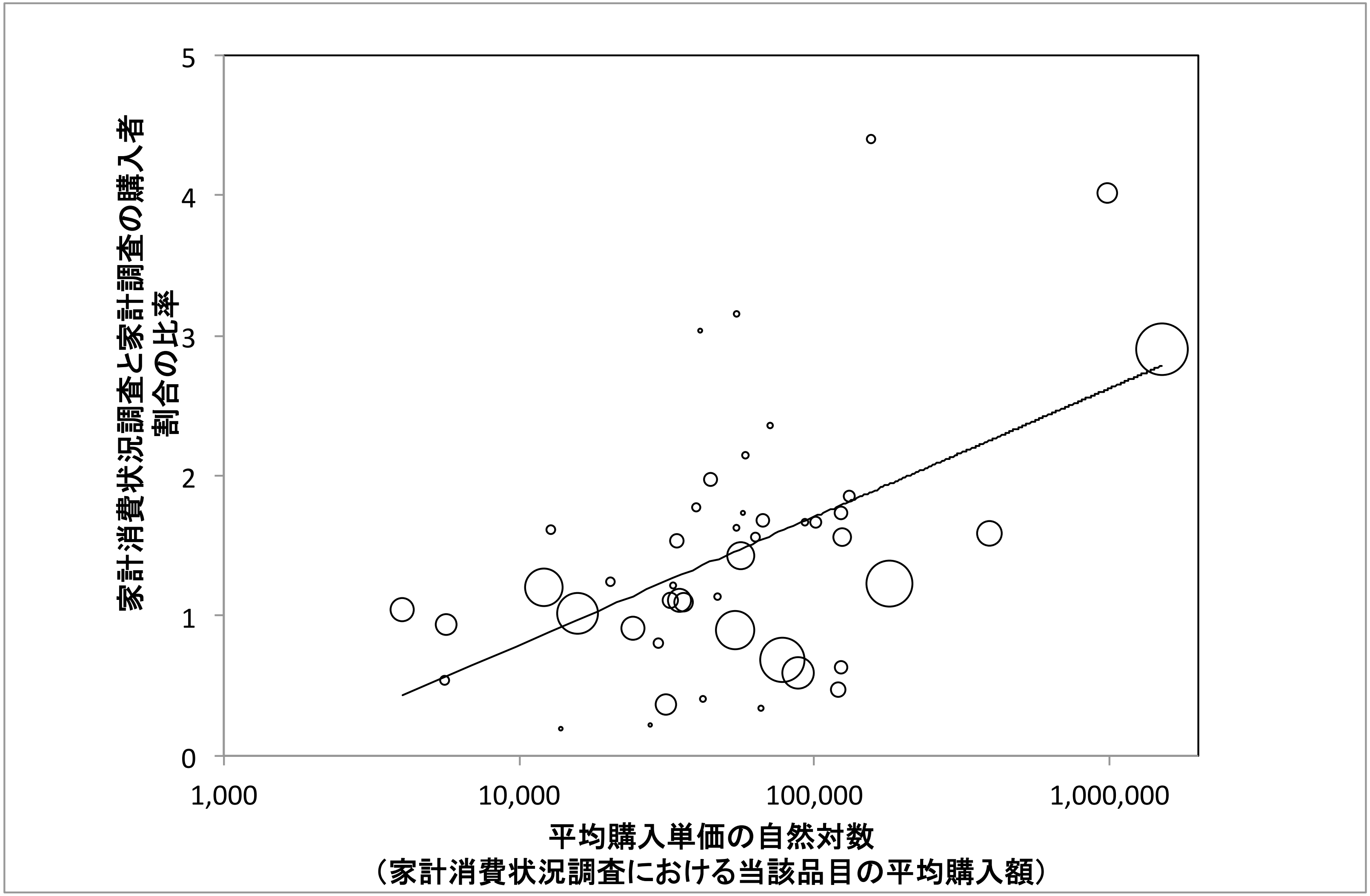
自動車等購入費



住宅修繕関連



平均購入単価と記入漏れ



記入漏れの原因：可能性

- 単価の大きな財・サービスが通常の「家計簿」という概念となじまないため
  - 調査世帯が記入するべきではないと判断している可能性
  - 調査方法（自由記入の家計簿方式）の問題
- 結婚式や葬式などの儀礼的な行事への支出額を明らかにすることへの心理的抵抗
  - 調査実務の問題（調査員との関係など）
- 海外旅行・結婚式・葬式などのために多忙で調査に十分に協力できていない可能性
  - サンプルセレクションの問題

まとめ

- 消費関連統計には、家計調査・全国消費実態調査・国民生活基礎調査・家計消費状況調査などがある
- 国民生活基礎調査は、おおむね家計調査と整合的な時系列推移
  - かつては家計支出額不詳が多かったため不規則な変動
  - 現在は、支出総額はおおむね同水準だが総額しか利用できない
- 家計調査よりも家計消費状況調査の方が支出が多い
  - 差の1/3は、調査対象品目以外の差は「調査疲れ」と「几帳面な世帯の割合増」で説明できる
  - 差の2/3は、家計消費状況調査の調査対象品目で説明できる
    - 自動車購入・住宅工事関連で半分程度説明できる
    - 自動車購入などが少ない理由については検討が必要
- 家計調査を消費の主要な統計として利用するためには
  - 公表データでは、高額消費を家計消費状況調査等で補正する必要
    - 家計消費指数の活用
  - ミクロデータでは、調査回数を説明変数に加えるなどの対応が必要

Prediction in Heteroscedastic Nested Error Regression Models with Random Dispersions

Tatsuya Kubokawa[*], Shonosuke Sugasawa[†], Malay Ghosh[‡], and Sanjay Chaudhuri[§]

University of Tokyo, University of Florida and National University of Singapore

January 5, 2015

Abstract

The paper concerns small-area estimation in the heteroscedastic nested error regression (HNER) model which assumes that the within-area variances are different among areas. Although HNER is useful for analyzing data where the within-area variation changes from area to area, it is difficult to provide good estimates for the error variances because of small samples sizes for small-areas. To fix this difficulty, we suggest a random dispersion HNER model which assumes a prior distribution for the error variances. The resulting Bayes estimates of small area means provide stable shrinkage estimates even for small sample sizes. Next we propose an empirical Bayes procedure for estimating the small area means. For measuring uncertainty of the empirical Bayes estimators, we use the conditional and unconditional mean squared errors (MSE) and derive their second-order approximations. It is interesting to note that the difference between the two MSEs appears in the first-order terms while the difference appears in the second-order terms for classical normal linear mixed models. Second-order unbiased estimators of the two MSEs are given with an application to the posted land price data.

Key words and phrases: Asymptotic approximation, conditional mean squared error, empirical Bayes, parametric bootstrap, second-order approximation, second-order unbiased estimate, small area estimation.

1 Introduction

Linear mixed (LM) models and the model-based estimators including empirical Bayes estimator (EB) or empirical best linear unbiased predictor (EBLUP) have been studied quite extensively in the literature from both theoretical and applied points of view. For a good review and account on this topic, see Ghosh and Rao (1994), Pfeffermann (2002), Rao (2003) and Datta (2009). Of these, the nested error regression (NER) model introduced by Battese, Harter and Fuller (1988) has been used as a unit-level model. The NER model with m small-areas assumes that the data $\boldsymbol{y}_i = (y_{i1}, \dots, y_{in_i})^T$ are taken from the i-th small-area, $i = 1, \dots, m$, where $\boldsymbol{y}_1, \dots, \boldsymbol{y}_m$ are mutually

[*]Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN. E-Mail: tatsuya@e.u-tokyo.ac.jp

[†]Graduate School of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, E-Mail: shonosuke622@gmail.com

[‡]Department of Statistics, University of Florida, 102 Griffin-Floyd Hall, Gainesville, Florida 32611. E-Mail: ghoshm@stat.ufl.edu

[§]Department of Statistics and Applied Probability, National University of Singapore, Block S16, Level 7, 6 Science Drive 2, SINGAPORE 117546. E-Mail: stasc@nus.edu.sg

independent. It is further assumed that y_{ij} is normally distributed with $E[y_{ij}] = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}$, $Var(y_{ij}) = \sigma_y^2$ and $Corr(y_{ij}, y_{ik}) = \rho$, $j \neq k$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, σ_y^2 and ρ are unknown parameters, $\boldsymbol{x}_{ij}$'s are known vectors of covariates, and $Corr(y_{ij}, y_{ik})$ denotes the correlation coefficient of y_{ij} and y_{ik}.

The NER model can be expressed as a random effects model with

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad i = 1, \ldots, m;\ j = 1, \ldots, n_i, \tag{1.1}$$

where v_i's and ε_{ij}'s are mutually independent with $v_i \sim \mathcal{N}(0, \lambda\sigma^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Then $Var(v_i)/Var(\varepsilon_{ij}) = \lambda$, and that σ_y^2 and ρ correspond to

$$\sigma_y^2 = (1 + \lambda)\sigma^2 \quad \text{and} \quad \rho = \lambda/(1 + \lambda).$$

Jiang and Nguyen (2012) illustrated that the within-area sample variances change dramatically from small-area to small-area for the data given in Battese, *et al.* (1988). Figure 1, given in Section 5 in this paper, also indicates variability of the within-area variances. Jiang and Nguyen (2012) assumed that the variance $E[(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta})^2]$ is proportional to σ_i^2, which depends on the area i. Since $E[(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta})^2] = E[(v_i + \varepsilon_{ij})^2] = Var(v_i) + Var(\varepsilon_{ij})$, this assumption implies that $Var(v_i) + Var(\varepsilon_{ij}) = C\sigma_i^2$ for some constant C. If we let $Var(\varepsilon_{ij}) = \sigma_i^2$, we can see that $\{Var(v_i)/Var(\varepsilon_{ij}) + 1\}Var(\varepsilon_{ij}) = \{Var(v_i)/Var(\varepsilon_{ij}) + 1\}\sigma_i^2 = C\sigma_i^2$, which means that

$$Var(v_i)/Var(\varepsilon_{ij}) = C - 1.$$

That is, $Var(v_i)/Var(\varepsilon_{ij})$ is a constant.

Using the same notation as in the NER model, we write $Var(v_i)/Var(\varepsilon_{ij}) = \lambda$. Thus, the heteroscedastic nested error regression (HNER) model suggested by Jiang and Nguyen (2012) is the model given in (1.1) with

$$Var(v_i) = \lambda\sigma_i^2 \quad \text{and} \quad Var(\varepsilon_{ij}) = \sigma_i^2. \tag{1.2}$$

In the HNER model, Jiang and Nguyen (2012) showed that the maximum likelihood (ML) estimators of $\boldsymbol{\beta}$ and λ are consistent for large m and that the resulting empirical Bayes (EB) estimator of $\xi_i = \overline{\boldsymbol{x}}_i^T\boldsymbol{\beta} + v_i$ ($\overline{\boldsymbol{x}}_i = n_i^{-1}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}$) estimates the Bayes estimator consistently since the Bayes estimator does not depend on $\sigma_1^2, \ldots, \sigma_m^2$, but on $\boldsymbol{\beta}$ and λ. This is quite interesting, because the number of unknown variances σ_i^2's increases as m tends to infinity. However, the posterior variance of v_i given $(y_{i1}, \ldots, y_{i,n_i})$ is

$$Var(v_i|y_{i1}, \ldots, y_{i,n_i}) = \sigma_i^2\lambda/(1 + n_i\lambda),$$

which implies that the mean squared error (MSE) of the EB estimator of ξ_i depends on σ_i^2. Then, we need to estimate σ_i^2 for estimating the MSE of the EB estimator of ξ_i. Since the sample sizes n_i is small in the small-area estimation, we cannot provide good estimates for σ_i^2 with reasonable precision.

In this paper, we propose a random dispersion HNER (RHNER) model which assumes that the areawise precisions $(\sigma_i^2)^{-1}$, $i = 1, \ldots, m$, are mutually independent gamma random variables. The resulting Bayes estimator of σ_i^2 gives stable shrinkage estimates of small area means even for $n_i - p = 0$.

For measuring uncertainty of the empirical Bayes estimator $\hat{\xi}_i^{EB}$ of ξ_i, we use the conditional and unconditional mean squared errors (MSE) defined by

$$\begin{aligned} cMSE(\omega; \hat{\xi}_i^{EB}|\boldsymbol{y}_i) =& E[(\hat{\xi}_i^{EB} - \xi_i)^2|\boldsymbol{y}_i], \\ MSE(\omega; \hat{\xi}_i^{EB}) =& E[(\hat{\xi}_i^{EB} - \xi_i)^2]. \end{aligned}$$

When data of the small area of interest are observed as $\boldsymbol{y}_i$ and one wants to know the prediction error of the EB estimators based on these data, the conditional mean squared error (cMSE) given $\boldsymbol{y}_i$ is used instead of the conventional unconditional MSE. Booth and Hobert (1998) showed that the difference between the cMSE and MSE is quite small and appears in the second-order terms in classical normal linear mixed models. In this article, however, we show that the difference appears in the leading or the first-order terms for the RHNER model.

2 HNER Models with Random Dispersions

2.1 Setup of models and predictors

We begin with the model given in (1.1) and (1.2). For stable estimators of σ_i^2's, we need sufficient amount of data from each area. Since n_i's are typically small, σ_i^2 cannot usually be estimated with reasonable precision. To give more stable estimators for σ_i^2, we assume a prior distribution for σ_i^2. Let $\eta_i = 1/\sigma_i^2$. It is assumed that $\eta_1, \ldots, \eta_m$ are independent and identically distributed with common pdf

$$\pi(\eta_i|\tau_1, \tau_2) \sim \mathcal{G}a(\tau_1/2, 2/\tau_2), \tag{2.1}$$

a gamma distribution with mean τ_1/τ_2 and variance $2\tau_1/\tau_2^2$. The HNER model given in (1.1) and (1.2) with the random dispersion (2.1) is called a *Random Heteroscedastic Nested Error Regression* (RHNER) model.

Let $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_m^T)^T$, $\boldsymbol{X}_i^T = (\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,n_i})$ and $\boldsymbol{X} = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_m^T)$. All the unknown parameters are denoted by $\omega = (\boldsymbol{\beta}, \lambda, \boldsymbol{\tau})$ for $\boldsymbol{\tau} = (\tau_1, \tau_2)$. Then, the RHNER model is given by

$$\begin{aligned} \boldsymbol{y}_i|v_i, \eta_i \sim& \mathcal{N}_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{j}_{n_i}v_i, \eta_i^{-1}\boldsymbol{I}_{n_i}), \\ v_i|\eta_i \sim& \mathcal{N}(0, \lambda\eta_i^{-1}), \\ \eta_i \sim& \mathcal{G}a(\tau_1/2, 2/\tau_2). \end{aligned} \tag{2.2}$$

The conditional distribution of v_i given $\boldsymbol{y}_i$ and η_i is $\mathcal{N}(\hat{v}_i, \lambda\eta_i^{-1}/(n_i\lambda + 1))$, where

$$\hat{v}_i = \hat{v}_i(\boldsymbol{\beta}, \lambda) = \frac{n_i\lambda}{n_i\lambda + 1}(\overline{y}_i - \overline{\boldsymbol{x}}_i^T\boldsymbol{\beta}). \tag{2.3}$$

It is noted that $\hat{v}_i = E[v_i|\boldsymbol{y}_i]$ does not depend on η_i or σ_i^2.

In this paper, we consider the problem of predicting the mixed quantity

$$\xi_i = \overline{\boldsymbol{x}}_i^T\boldsymbol{\beta} + v_i, \quad i = 1, \ldots, m.$$

The conditional expectation of ξ_i given $\boldsymbol{y}_i$ and η_i is

$$\hat{\xi}_i^B(\boldsymbol{\beta}, \lambda) = E[\xi_i|\boldsymbol{y}_i, \sigma_i^2] = \overline{\boldsymbol{x}}_i^T\boldsymbol{\beta} + \hat{v}_i(\boldsymbol{\beta}, \lambda).$$

This is interpreted as the Bayes estimator of ξ under squared error loss. Since it does not depend on η_i, the estimator $\hat{\xi}_i^B(\boldsymbol{\beta}, \lambda)$ continues to be the conditional expectation of ξ_i given $\boldsymbol{y}_i$ after integrating out the η_i, that is the Bayes estimator of ξ_i is the same in the two situations. However, the empirical Bayes estimators, which substitute estimators of $\boldsymbol{\beta}$ and λ into $\hat{\xi}_i^B(\boldsymbol{\beta}, \lambda)$, are different between the HNER and RHNER models.

In the HNER model, we need to estimate $(m + p + 1)$ parameters $\boldsymbol{\beta}$, λ and $\sigma_1^2, \ldots, \sigma_m^2$. Noting that the number of parameters increases as m increases and that n_is are bounded in small-area estimation, we are faced with the problem of consistency and instability of the estimators of σ_i^2. In this situation, Jiang and Nguyen (2012) established the surprising result that the MLEs of $\boldsymbol{\beta}$ and λ are consistent, which lead to the consistency of the EB estimator $\hat{\xi}_i^B(\widehat{\boldsymbol{\beta}}, \hat{\lambda})$. However, there are no consistent estimators of the σ_i^2. This problem can be fixed if instead the RHNER model is used. In fact, the parameters we need to estimate in the RHNER model are $\boldsymbol{\beta}$, λ, τ_1 and τ_2, and we can provide their consistent estimators.

2.2 A motivation from estimation of dispersions

We give more detailed motivation from the estimation of the dispersion parameters σ_i^2 in the HNER model. We first treat the simple case that $\boldsymbol{\beta} = \mathbf{0}$ and $n_1 = \cdots = n_m = n$ in (1.1). Let $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_m^2)^T$ and $\gamma = 1/(1+n\lambda)$. It follows from the equation (4) of Jiang and Nguyen (2012) that the log-likelihood is then

$$L^H(\gamma, \boldsymbol{\sigma}^2) = \sum_{i=1}^m \Big[-n \log \sigma_i^2 + \log \gamma - \Big\{ \sum_{j=1}^n (y_{ij} - \overline{y}_i)^2 + n\gamma \overline{y}_i^2 \Big\} / \sigma_i^2 \Big] + K,$$

where K is a generic constant. Differentiating $L^H(\gamma, \boldsymbol{\sigma}^2)$ with respect to γ and σ_i^2's, we see that the maximum likelihood (ML) estimators, $\hat{\gamma}^H$ and $\hat{\sigma}^2_{(H)i}$, of γ and σ_i^2's are solutions of the equations

$$\begin{aligned} \hat{\gamma}^H =& \frac{m}{\sum_{i=1}^m n\overline{y}_i^2 / \hat{\sigma}^2_{(H)i}}, \\ \hat{\sigma}^2_{(H)i} =& \frac{1}{n} \Big\{ \sum_{j=1}^n (y_{ij} - \overline{y}_i)^2 + n\hat{\gamma}^H \overline{y}_i^2 \Big\}. \end{aligned} \tag{2.4}$$

It is interesting to point out that $\hat{\sigma}^2_{(H)i}$ is an asymptotically unbiased estimator of σ_i^2. For the proof, we can use the fact that $n\overline{y}_i^2$ and $\sum_{j=1}^n (y_{ij} - \overline{y}_i)^2$ are mutually independent with $n\overline{y}_i^2/(1+n\lambda) \sim \sigma_i^2 \chi_1^2$ and $\sum_{j=1}^n (y_{ij} - \overline{y}_i)^2 \sim \sigma_i^2 \chi_{n-1}^2$. Then, from the consistency of $\hat{\gamma}^H$, it follows that $E[\hat{\sigma}^2_{(H)i}]$ converges to

$$\frac{1}{n} E\Big[\sum_{j=1}^n (y_{ij} - \overline{y}_i)^2 + n\gamma \overline{y}_i^2 \Big] = \frac{\sigma_i^2}{n} E[\chi_{n-1}^2 + \gamma(1+n\lambda)\chi_1^2] = \sigma_i^2,$$

which shows that $\hat{\sigma}^2_{(H)i}$ is an asymptotically unbiased estimator of σ_i^2.

Although $\hat{\sigma}^2_{(H)i}$ is asymptotically unbiased, it is clear that $\hat{\sigma}^2_{(H)i}$ is not consistent when $m \to \infty$, but n is bounded. Thus, we need to modify $\hat{\sigma}^2_{(H)i}$ when n is small. For example, we look at the empirical Bayes estimator of ξ_i. In the simple case we treat here, we have $\xi_i = v_i$, and the EB estimator of ξ_i is given by

$$\hat{\xi}_i^H = (1 - \hat{\gamma}^H)\overline{y}_i = \Big\{ 1 - \frac{m}{\sum_{i=1}^m n\overline{y}_i^2 / \hat{\sigma}^2_{(H)i}} \Big\} \overline{y}_i,$$

from (2.4). This is a natural shrinkage estimator, and it is reasonable for large m since $\hat{\gamma}^H$ is consistent. When m is not large, however, there is a concern about the precision of the estimator $\hat{\sigma}^2_{(H)i}$. Since $\sum_{j=1}^n (y_{ij} - \overline{y}_i)^2 \le n\hat{\sigma}^2_{(H)i} \le \sum_{j=1}^n y_{ij}^2$, it is seen that

$$\frac{\overline{y}_i^2}{\sum_{j=1}^n y_{ij}^2/n} \le \frac{\overline{y}_i^2}{\hat{\sigma}^2_{(H)i}} \le \frac{\overline{y}_i^2}{T_i/n},$$

for $T_i = \sum_{j=1}^n (y_{ij} - \overline{y}_i)^2$. When n is small, clearly $\overline{y}_i^2/\hat{\sigma}^2_{(H)i}$ has a large variation, which leads to the instability of the empirical Bayes estimator $\hat{\xi}_i^H$. Although the simple case of equal replications n is considered here, in the survey data we need to handle the case of small sample sizes n_i's for some small-areas, and the estimators $\hat{\sigma}^2_{(H)i}$'s would not have enough degrees of freedom.

To overcome this drawback, we need to stabilize $\hat{\sigma}^2_{(H)i}$ by shrinking it to a point. The random dispersion model is helpful for this purpose. Assume that $\eta_i = 1/\sigma_i^2$ has a distribution $\mathcal{G}a(\tau_1/2, 2/\tau_2)$. Since $T_i\eta_i = T_i/\sigma_i^2 \sim \chi_{n-1}^2$, from the joint distribution of (T_i, η_i), the posterior mean of σ_i^2 given T_i is

$$E[\sigma_i^2 | T_i] = (T_i + \tau_2)/(n - 1 + \tau_1).$$

When $\hat{\tau}_1$ and $\hat{\tau}_2$ are estimators of τ_1 and τ_2 based on the statistics $T_1, \ldots, T_m$, it is reasonable to estimate σ_i^2 by

$$\hat{\sigma}^2_{(RH)i} = (T_i + \hat{\tau}_2)/(n - 1 + \hat{\tau}_1).$$

Clearly, $\hat{\sigma}^2_{(RH)i}$ is more stable than the unbiased estimator $T_i/(n-1)$ when n is small. Replacing $\hat{\sigma}^2_{(H)i}$ in $\hat{\xi}_i^H$ with a shrinkage estimator like $\hat{\sigma}^2_{(RH)i}$, one can get the more stabilized empirical Bayes estimator

$$\hat{\xi}_i^{RH} = \Big\{1 - \frac{m}{\sum_{i=1}^m n\overline{y}_i^2/\hat{\sigma}^2_{(RH)i}}\Big\}\overline{y}_i.$$

Another need for a consistent estimator of σ_i^2 appears in evaluation of uncertainty of the empirical Bayes estimator $\hat{\xi}_i^H$. When the mean squared error is used for measuring the uncertainty, the MSE of $\hat{\xi}_i^H$, denoted by $E[(\hat{\xi}_i^H - \xi_i)^2]$ converges to

$$E[\mathrm{Var}(v_i|\boldsymbol{y}_i)] = \sigma_i^2\lambda/(1+n\lambda) = \sigma_i^2(1-\gamma)/n$$

for large m. In order to estimate the uncertainty of $\hat{\xi}_i^H$, we want to estimate the leading term of the MSE consistently. Since $\hat{\sigma}^2_{(H)i}$ is not consistent, however, we cannot provide any consistent estimator of the leading term in the MSE of $\hat{\xi}_i^H$ in the HNER model. This drawback is overcome in the RHNER model.

3 Predictors and Asymptotic Properties of MSE

3.1 MLE of parameters and the empirical Bayes estimator

We now return back to the RHNER model given in (2.2). When λ and $\boldsymbol{\beta}$ are known, the best predictor or the Bayes estimator of $\xi_i = \overline{\boldsymbol{x}}_i^T\boldsymbol{\beta} + v_i$ is given by

$$\begin{aligned}\hat{\xi}_i^B =& \hat{\xi}_i^B(\boldsymbol{\beta},\lambda) = E[\xi_i|\boldsymbol{y}_i]\\ =& \overline{\boldsymbol{x}}_i^T\boldsymbol{\beta} + (1-\gamma_i)(\overline{y}_i - \overline{\boldsymbol{x}}_i^T\boldsymbol{\beta}),\end{aligned} \tag{3.1}$$

where $\gamma_i = \gamma_i(\lambda) = 1/(n_i\lambda + 1)$. In our case since λ and $\boldsymbol{\beta}$ are unknown, we need to estimate them from the marginal distributions of the $\boldsymbol{y}_i$. We provide the maximum likelihood (ML) estimators for unknown parameters $\omega = (\boldsymbol{\beta}, \lambda, \tau_1, \tau_2)$.

The marginal likelihood of $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)^T$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)^T$ after integrating out the full joint likelihood with respect to v_i's can be expressed as

$$\begin{aligned}f(\boldsymbol{y},\boldsymbol{\eta}|\omega) =& \prod_{i=1}^m \Big\{\frac{\eta_i^{n_i/2}}{(2\pi)^{n_i/2}\sqrt{n_i\lambda+1}}\exp\Big[-\frac{\eta_i}{2}\{\sum_{j=1}^{n_i}(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta})^2 - \frac{n_i^2\lambda}{n_i\lambda+1}(\overline{y}_i - \overline{\boldsymbol{x}}_i^T\boldsymbol{\beta})^2\}\Big]\\ & \times \pi(\eta_i|\tau_1,\tau_2)\Big\}\\ =& \prod_{i=1}^m\Big\{\frac{\tau_2^{\tau_1/2}\eta_i^{(n_i+\tau_1)/2-1}2^{-(n_i+\tau_1)/2}}{\pi^{n_i/2}\Gamma(\tau_1/2)\sqrt{n_i\lambda+1}}\exp\Big[-\frac{\eta_i}{2}\{Q_i(\boldsymbol{y}_i,\boldsymbol{\beta},\lambda) + \tau_2\}\Big]\Big\},\end{aligned} \tag{3.2}$$

where

$$\begin{aligned}Q_i = Q_i(\boldsymbol{y}_i,\boldsymbol{\beta},\lambda) =& \sum_{j=1}^{n_i}(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta})^2 - \frac{n_i^2\lambda}{n_i\lambda+1}(\overline{y}_i - \overline{\boldsymbol{x}}_i^T\boldsymbol{\beta})^2\\ =& \sum_{j=1}^{n_i}\{(y_{ij} - \overline{y}_i) - (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)^T\boldsymbol{\beta}\}^2 + n_i\gamma_i(\lambda)(\overline{y}_i - \overline{\boldsymbol{x}}_i^T\boldsymbol{\beta})^2,\end{aligned} \tag{3.3}$$

where $\gamma_i = \gamma_i(\lambda) = 1/(n_i\lambda + 1)$. Integrating out the joint density $f(\boldsymbol{y}, \boldsymbol{\eta}|\omega)$ in (3.2) with respect to $\boldsymbol{\eta}$ yields the marginal likelihood of $\boldsymbol{y}$ given by

$$f(\boldsymbol{y}|\omega) = \prod_{i=1}^{m} \Big\{ \frac{\tau_2^{\tau_1/2}\Gamma((n_i+\tau_1)/2)}{\pi^{n_i/2}\sqrt{n_i\lambda+1}\Gamma(\tau_1/2)} \{Q_i(\boldsymbol{y}_i, \boldsymbol{\beta}, \lambda) + \tau_2\}^{-(n_i+\tau_1)/2} \Big\}. \tag{3.4}$$

Let $L = L(\boldsymbol{\beta}, \lambda, \tau_1, \tau_2) = \log f(\boldsymbol{y}|\omega)$. Then,

$$\begin{aligned}2L =& - n_i \log \pi + m\tau_1 \log \tau_2 + 2\sum_{i=1}^{m} \psi(\frac{n_i+\tau_1}{2}) - 2m\psi(\frac{\tau_1}{2}) \\ & - \sum_{i=1}^{m} \log(n_i\lambda+1) - \sum_{i=1}^{m} (n_i+\tau_1)\log(Q_i+\tau_2),\end{aligned}$$

where $\psi(a) = \log(\Gamma(a))$. Let $L_{\boldsymbol{\beta}}$, L_λ, L_{τ_1} and L_{τ_2} be the derivatives of L with respect to $\boldsymbol{\beta}$, λ, τ_1 and τ_2. Then,

$$\begin{aligned}2L_{\boldsymbol{\beta}} =& - \sum_{i=1}^{m} \frac{n_i+\tau_1}{Q_i+\tau_2}\frac{\partial Q_i}{\partial \boldsymbol{\beta}}, \\ 2L_\lambda =& - \sum_{i=1}^{m} \frac{n_i+\tau_1}{Q_i+\tau_2}\frac{\partial Q_i}{\partial \lambda} - \sum_{i=1}^{m} n_i\gamma_i, \\ 2L_{\tau_1} =& \sum_{i=1}^{m} \log\Big(\frac{\tau_2}{Q_i+\tau_2}\Big) + \sum_{i=1}^{m}\Big\{\psi'(\frac{n_i+\tau_1}{2}) - \psi'(\frac{\tau_1}{2})\Big\}, \\ 2L_{\tau_2} =& m\frac{\tau_1}{\tau_2} - \sum_{i=1}^{m} \frac{n_i+\tau_1}{Q_i+\tau_2},\end{aligned} \tag{3.5}$$

where $\partial Q_i/\partial\lambda = -n_i^2\gamma_i^2(\overline{y}_i - \overline{\boldsymbol{x}}_i^T\boldsymbol{\beta})^2$ for $\partial\gamma_i/\partial\lambda = -n_i\gamma_i^2$, and

$$\frac{\partial Q_i}{\partial \boldsymbol{\beta}} = -2\sum_{j=1}^{n_i}\{(y_{ij}-\overline{y}_i) - (\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i)^T\boldsymbol{\beta}\}(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i) - 2n_i\gamma_i(\overline{y}_i - \overline{\boldsymbol{x}}_i^T\boldsymbol{\beta})\overline{\boldsymbol{x}}_i. \tag{3.6}$$

The MLEs of $\boldsymbol{\beta}$, λ, τ_1 and τ_2 are the solution of the simultaneous equations $L_{\boldsymbol{\beta}} = 0$, $L_\lambda = 0$, $L_{\tau_1} = 0$ and $L_{\tau_2} = 0$, and the MLEs are denoted by $\widehat{\boldsymbol{\beta}}$, $\hat{\lambda}$, $\hat{\tau}_1$ and $\hat{\tau}_2$. The empirical Bayes estimator of $\xi_i = \overline{\boldsymbol{x}}_i^T\boldsymbol{\beta} + v_i$ is provided by

$$\hat{\xi}_i^{EB} = \hat{\xi}_i^B(\widehat{\boldsymbol{\beta}}, \hat{\lambda}) = \overline{\boldsymbol{x}}_i^T\widehat{\boldsymbol{\beta}} + (1-\hat{\gamma}_i)(\overline{y}_i - \overline{\boldsymbol{x}}_i^T\widehat{\boldsymbol{\beta}}), \tag{3.7}$$

where $\hat{\gamma}_i = \gamma_i(\hat{\lambda}) = 1/(n_i\hat{\lambda}+1)$.

3.2 Asymptotic properties of MLE

To evaluate the mean squared errors of the empirical Bayes estimator $\hat{\xi}_i^{EB}$ asymptotically, we need to derive asymptotic variances and covariances of the MLE when m tends to infinity. To derive asymptotic properties of the MLE, we assume the following:

(A1) The sample sizes n_i's are bounded below and above as $\underline{n} \le n_i \le \overline{n}$ for constants $\underline{n}$ and $\overline{n}$. The elements of $\boldsymbol{X}$ are uniformly bounded, $\boldsymbol{X}^T\boldsymbol{X}$ is positive definite and $\boldsymbol{X}^T\boldsymbol{X}/m$ converges to a positive definite matrix.

Since their asymptotic variances and covariances are expressed by the Fisher information matrix, we begin by deriving the Fisher information. Let $\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ be the Fisher information matrix of $\boldsymbol{\beta}$. The Fisher information matrix of $\boldsymbol{\theta} = (\lambda, \tau_1, \tau_2)^T$ and the inverse are denoted by

$$\boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}} = \begin{pmatrix} I_{\lambda\lambda} & I_{\lambda\tau_1} & I_{\lambda\tau_2} \\ I_{\lambda\tau_1} & I_{\tau_1\tau_1} & I_{\tau_1\tau_2} \\ I_{\lambda\tau_2} & I_{\tau_1\tau_2} & I_{\tau_2\tau_2} \end{pmatrix} \quad \text{and} \quad \boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} = \begin{pmatrix} I^{\lambda\lambda} & I^{\lambda\tau_1} & I^{\lambda\tau_2} \\ I^{\lambda\tau_1} & I^{\tau_1\tau_1} & I^{\tau_1\tau_2} \\ I^{\lambda\tau_2} & I^{\tau_1\tau_2} & I^{\tau_2\tau_2} \end{pmatrix}.$$

Then, exact expressions of the Fisher information matrices can be derived in the following theorem. The proof is given in the Appendix.

Theorem 3.1 *The Fisher information of $\boldsymbol{\beta}$ is given by*

$$\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \frac{\tau_1}{\tau_2}\sum_{i=1}^{m}\frac{n_i+\tau_1}{n_i+\tau_1+2}\Big\{\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i)^T + n_i\gamma_i\overline{\boldsymbol{x}}_i\overline{\boldsymbol{x}}_i^T\Big\}.$$

Also, $\boldsymbol{I}_{\boldsymbol{\beta}\lambda} = \mathbf{0}$, $\boldsymbol{I}_{\boldsymbol{\beta}\tau_1} = \mathbf{0}$ and $\boldsymbol{I}_{\boldsymbol{\beta}\tau_2} = \mathbf{0}$. The elements of $2\boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ are given by

$$\begin{aligned} 2I_{\lambda\lambda} =& \sum_{i=1}^{m}\frac{(n_i+\tau_1-1)n_i^2\gamma_i^2}{n_i+\tau_1+2}, \quad & 2I_{\lambda\tau_1} =& -\sum_{i=1}^{m}\frac{n_i\gamma_i}{n_i+\tau_1}, \\ 2I_{\lambda\tau_2} =& \frac{\tau_1}{\tau_2}\sum_{i=1}^{m}\frac{n_i\gamma_i}{n_i+\tau_1+2}, \quad & 2I_{\tau_1\tau_1} =& \frac{1}{2}\sum_{i=1}^{m}\{\psi''(\frac{\tau_1}{2}) - \psi''(\frac{n_i+\tau_1}{2})\}, \\ 2I_{\tau_1\tau_2} =& -\frac{1}{\tau_2}\sum_{i=1}^{m}\frac{n_i}{n_i+\tau_1}, \quad & 2I_{\tau_2\tau_2} =& \frac{\tau_1}{\tau_2^2}\sum_{i=1}^{m}\frac{n_i}{n_i+\tau_1+2}. \end{aligned}$$

It follows from Theorem 3.1 and assumption (A1) that $m^{-1}\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}} = O(1)$ and $m^{-1}\boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}} = O(1)$, and the limiting values of these quantities are away from zero. The following theorem is essential for approximating the MSE of $\hat{\xi}_i^{EB}$ asymptotically. The proof is given in the Appendix.

Theorem 3.2 *Assume condition* (A1). *Then, for $\widehat{\boldsymbol{\theta}} = (\hat{\lambda}, \hat{\tau}_1, \hat{\tau}_2)^T$, it holds that*

$$\begin{aligned} E[(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T|\boldsymbol{y}_i] =& (\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}})^{-1} + O_p(m^{-3/2}), \\ E[(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T|\boldsymbol{y}_i] =& (\boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}})^{-1} + O_p(m^{-3/2}), \\ E[(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T|\boldsymbol{y}_i] =& O_p(m^{-3/2})0 \end{aligned} \tag{3.8}$$

This implies that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\boldsymbol{y}_i = O_p(m^{-1/2})$ and $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|\boldsymbol{y}_i = O_p(m^{-1/2})$. Also, the conditional biases $E[\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\boldsymbol{y}_i] = O(m^{-1})$ and $E[\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|\boldsymbol{y}_i] = O(m^{-1})$.

4 Measures of Uncertainty of the Empirical Bayes Estimator

4.1 Second-order approximation of the conditional and unconditional MSEs

We shall derive a second-order approximation of the MSE of the empirical Bayes (EB) estimator and its second-order unbiased estimator. Recall that we want to predict $\xi_i = \overline{\boldsymbol{x}}_i^T\boldsymbol{\beta} + v_i$ with EB $\hat{\xi}_i^{EB} = \hat{\xi}_i^B(\widehat{\boldsymbol{\beta}}, \hat{\lambda}) = \overline{\boldsymbol{x}}_i^T\widehat{\boldsymbol{\beta}} + \hat{v}_i(\widehat{\boldsymbol{\beta}}, \hat{\lambda})$. For measuring uncertainty of EB, we use the conditional and unconditional mean squared errors (MSE) defined by

$$\begin{aligned} cMSE(\omega; \hat{\xi}_i^{EB}|\boldsymbol{y}_i) =& E[(\hat{\xi}_i^{EB} - \xi_i)^2|\boldsymbol{y}_i], \\ MSE(\omega; \hat{\xi}_i^{EB}) =& E[(\hat{\xi}_i^{EB} - \xi_i)^2]. \end{aligned}$$

The conditional and unconditional MSEs can be decomposed as

$$\begin{aligned} cMSE(\omega;\hat{\xi}_i^{EB}|\boldsymbol{y}_i) =& E[\{\xi_i-\hat{\xi}_i^B(\boldsymbol{\beta},\lambda)\}^2|\boldsymbol{y}_i]+E[\{\hat{\xi}_i^B(\widehat{\boldsymbol{\beta}},\hat{\lambda})-\hat{\xi}_i^B(\boldsymbol{\beta},\lambda)\}^2|\boldsymbol{y}_i] \\ =& g_1^c(\omega|\boldsymbol{y}_i)+g_2^c(\omega|\boldsymbol{y}_i), \quad \text{(say)} \end{aligned} \tag{4.1}$$

and

$$\begin{aligned} MSE(\omega;\hat{\xi}_i^{EB}) =& E[\{\xi_i-\hat{\xi}_i^B(\boldsymbol{\beta},\lambda)\}^2]+E[\{\hat{\xi}_i^B(\widehat{\boldsymbol{\beta}},\hat{\lambda})-\hat{\xi}_i^B(\boldsymbol{\beta},\lambda)\}^2] \\ =& g_1(\omega)+g_2(\omega). \quad \text{(say)} \end{aligned} \tag{4.2}$$

The first term $g_1^c(\omega|\boldsymbol{y}_i)$ is the posterior variance of ξ_i given $\boldsymbol{y}$, namely,

$$g_1^c(\omega|\boldsymbol{y}_i)=Var(\xi_i|\boldsymbol{y}_i)=\frac{\lambda}{n_i\lambda+1}E[\eta_i^{-1}|\boldsymbol{y}_i]=\frac{\lambda}{n_i\lambda+1}\frac{Q_i+\tau_2}{n_i+\tau_1-2}, \tag{4.3}$$

where Q_i is given in (3.3). Similarly, $g_1(\omega)$ is given by

$$g_1(\omega)=E[Var(\xi_i|\boldsymbol{y}_i)]=\frac{\lambda}{n_i\lambda+1}E[\eta_i^{-1}]=\frac{\lambda}{n_i\lambda+1}\frac{\tau_2}{\tau_1-2}. \tag{4.4}$$

Noting that $g_1^c(\omega|\boldsymbol{y}_i)=O_p(1)$, $g_2^c(\omega|\boldsymbol{y}_i)=O_p(m^{-1})$, $g_1(\omega)=O(1)$ and $g_2(\omega)=O(m^{-1})$, we can see that the difference between the cMSE and MSE appears in the leading or the first-order terms. This is an interesting fact, because the difference is small and appears in the second-order terms in the classical normal theory mixed models as demonstrated by Booth and Hobert (1998). They also showed that the difference is significant and appears in the first-order terms for distributions far from normality. Noting that the random dispersion model (2.2) is not a normal distribution, but close to a t-distribution, we observe that the above fact coincides with their assertion.

In the case of the HNER model, $Var(\xi_i|\boldsymbol{y}_i)$ is identical to $E[Var(\xi_i|\boldsymbol{y}_i)]$ since y_i has a normal distribution, and is given by $\sigma_i^2\lambda/(n_i\lambda+1)$. Thus, it should be noted that we cannot estimate the first-order term $\sigma_i^2\lambda/(n_i\lambda+1)$ consistently in the HNER model, since n_i is bounded. However, we can estimate $g_1^c(\omega|\boldsymbol{y}_i)$ and $g_1(\omega)$ consistently in the RHNER model (2.2) since λ, τ_1 and τ_2 are estimated consistently.

Theorem 4.1 *Under assumption* (A1), *the conditional MSE of* $\hat{\xi}_i^{EB}$ *is approximated as*

$$\begin{aligned} cMSE(\lambda,\boldsymbol{\tau};\hat{\xi}_i^{EB}|\boldsymbol{y}_i) =& \frac{1-\gamma_i}{n_i}\frac{Q_i+\tau_2}{n_i+\tau_1-2}+\gamma_i^2\overline{\boldsymbol{x}}_i^T\left(\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}\right)^{-1}\overline{\boldsymbol{x}}_i \\ &+n_i^2\gamma_i^4(\overline{y}_i-\overline{\boldsymbol{x}}_i^T\boldsymbol{\beta})^2I^{\lambda\lambda}+O_p(m^{-3/2}), \end{aligned} \tag{4.5}$$

for $\gamma_i=1/(n_i\lambda+1)$, *and the unconditional MSE is approximated as*

$$\begin{aligned} MSE(\lambda,\boldsymbol{\tau};\hat{\xi}_i^{EB}) =& \frac{1-\gamma_i}{n_i}\frac{\tau_2}{\tau_1-2}+\gamma_i^2\overline{\boldsymbol{x}}_i^T\left(\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}\right)^{-1}\overline{\boldsymbol{x}}_i \\ &+n_i\gamma_i^3\frac{\tau_2}{\tau_1-2}I^{\lambda\lambda}+O(m^{-3/2}). \end{aligned} \tag{4.6}$$

4.2 Second-order unbiased estimators of the conditional and unconditional MSEs

We now derive second-order unbiased estimators of the unconditional and conditional MSEs. Since it is hard to derive second-order biases of the MLEs of $\boldsymbol{\beta}$, λ, τ_1 and τ_2, we could not provide analytical second-order unbiased estimators of the MSEs. Instead, we use the parametric bootstrap methods, which provide second-order unbiased MSE estimators.

We begin by treating the unconditional case. The parametric bootstrap sample in this case is denoted as

$$y_{ij}^* = \boldsymbol{x}_{ij}^T \widehat{\boldsymbol{\beta}} + v_i^* + \varepsilon_{ij}^*, \quad i = 1, \dots, m; \; j = 1, \dots, n_i, \tag{4.7}$$

where v_i^*'s and ε_{ij}^*'s are conditionally mutually independent given η_i^*'s and

$$\begin{aligned} v_i^* | \eta_i^* \sim & \mathcal{N}(0, \hat{\lambda}/\eta_i^*), \\ \varepsilon_{ij}^* | \eta_i^* \sim & \mathcal{N}(0, 1/\eta_i^*), \\ \eta_i^* \sim & \sim \mathcal{G}a(\hat{\tau}_1/2, 2/\hat{\tau}_2). \end{aligned} \tag{4.8}$$

The estimator of the unconditional MSE, $MSE(\lambda, \boldsymbol{\tau}; \hat{\xi}_i^{EB})$, is given by

$$mse^*(\hat{\xi}_i^{EB}) = \hat{g}_1^* + \hat{g}_2^*,$$

where

$$\begin{aligned} \hat{g}_1^* =& 2g_1(\hat{\lambda}, \widehat{\boldsymbol{\tau}}) - E_*[g_1(\hat{\lambda}^*, \widehat{\boldsymbol{\tau}}^*)], \\ \hat{g}_2^* =& \hat{\gamma}_i^2 E^*[\{\overline{\boldsymbol{x}}_i^T(\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}})\}^2] + n_i \hat{\gamma}_i^3 \frac{\hat{\tau}_2}{\hat{\tau}_1 - 2} E^*[(\hat{\lambda}^* - \hat{\lambda})^2]. \end{aligned}$$

Proposition 4.1 *Assume the condition* (A1). *Then,*

$$E[mse^*(\hat{\xi}_i^{EB})] = MSE(\lambda, \boldsymbol{\tau}; \hat{\xi}_i^{EB}) + O(m^{-3/2}).$$

We next consider the conditional case. Keeping $\boldsymbol{y}_i = (y_{i1}, \dots, y_{in_i})^T$ fixed, a bootstrap sample $\boldsymbol{y}_k = (y_{k1}, \dots, y_{kn_k})^T$ is generated from (4.7) for $k \neq i$. Noting that $\boldsymbol{y}_i$ is fixed, we construct the estimators $\widehat{\boldsymbol{\beta}}^*_{(i)}$, $\hat{\lambda}^*_{(i)}$, $\hat{\tau}^*_{1(i)}$ and $\hat{\tau}^*_{2(i)}$ from $\boldsymbol{y}_i$ and the bootstrap sample

$$\boldsymbol{y}_1^*, \dots, \boldsymbol{y}_{i-1}^*, \boldsymbol{y}_i, \boldsymbol{y}_{i+1}^*, \dots, \boldsymbol{y}_m^* \tag{4.9}$$

with the same technique as used to obtain the estimator $\widehat{\boldsymbol{\beta}}$, $\hat{\lambda}$, $\hat{\tau}_1$ and $\hat{\tau}_2$. Let $E_*[\cdot|\boldsymbol{y}_i]$ be the expectation with regard to the bootstrap sample (4.9). The conditional MSE is given by $cMSE(\omega; \hat{\xi}_i^{EB}|\boldsymbol{y}_i) = g_1^c(\omega|\boldsymbol{y}_i) + g_2^c(\omega|\boldsymbol{y}_i)$, where $g_1^c(\omega|\boldsymbol{y}_i) = E[\{\xi_i - \hat{\xi}_i^B(\boldsymbol{\beta}, \lambda)\}^2|\boldsymbol{y}_i]$ and $g_2^c(\omega|\boldsymbol{y}_i) = E[\{\hat{\xi}_i^B(\widehat{\boldsymbol{\beta}}, \hat{\lambda}) - \hat{\xi}_i^B(\boldsymbol{\beta}, \lambda)\}^2|\boldsymbol{y}_i]$ from (4.1). Since $g_1^c(\omega|\boldsymbol{y}_i) = n_i^{-1}(1 - \gamma_i(\lambda))(Q_i(\boldsymbol{y}_i, \boldsymbol{\beta}, \lambda) + \tau_2)/(n_i + \tau_1 - 2)$ from (4.3), a second-order unbiased estimator of $g_1^c(\omega|\boldsymbol{y}_i)$ is given by

$$\hat{g}_1^{c*} = 2g_1(\boldsymbol{y}_i, \widehat{\boldsymbol{\beta}}, \hat{\lambda}, \widehat{\boldsymbol{\tau}}) - E_*\left[g_1(\boldsymbol{y}_i, \widehat{\boldsymbol{\beta}}^*_{(i)}, \hat{\lambda}^*_{(i)}, \widehat{\boldsymbol{\tau}}^*_{(i)})|\boldsymbol{y}_i\right].$$

Then, it can be verified that $E[\hat{g}_1^{c*}|\boldsymbol{y}_i] = g_1^c(\omega|\boldsymbol{y}_i) + o_p(m^{-1})$. $g_2^c(\omega|\boldsymbol{y}_i)$, is estimated via parametric bootstrap as

$$\hat{g}_2^{c*} = E^*\left[\{\hat{\xi}_i^{B*}(\widehat{\boldsymbol{\beta}}^*_{(i)}, \hat{\lambda}^*_{(i)}) - \hat{\xi}_i^{B*}(\widehat{\boldsymbol{\beta}}, \hat{\lambda})\}^2|\boldsymbol{y}_i\right].$$

Thus,

$$cmse^*(\hat{\xi}_i^{EB}|\boldsymbol{y}_i) = \hat{g}_1^{c*} + \hat{g}_2^{c*}. \tag{4.10}$$

Theorem 4.2 *Under the condition* (A1), *the estimator* (4.10) *is a second-order unbiased estimator of cMSE, namely*

$$E[cmse^*(\hat{\xi}_i^{EB}|\boldsymbol{y}_i)|\boldsymbol{y}_i] = cMSE(\omega; \hat{\xi}_i^{EB}|\boldsymbol{y}_i) + o_p(m^{-1}).$$

5 Application to PLP data in Japan

We now investigate empirical performances of the suggested model, the empirical Bayes estimator and the second-order unbiased estimators of the conditional and unconditional MSEs through analysis of real data. The data used here originates from the posted land price data along the Keikyu train line in 2001. This train line connects the suburbs in the Kanagawa prefecture to the Tokyo metropolitan area. Those who live in the suburbs in the Kanagawa prefecture take this line to work or study in Tokyo everyday. Thus, it is expected that the land price depends on the distance from Tokyo. The posted land price data are available for 52 stations on the Keikyu train line, and we consider each station as a small area, namely, $m = 52$. For the i-th station, data of n_i land spots are available, where n_i varies around 4 and some areas have only one observation.

To investigate variability in each area, the boxplots are drawn for all areas. For nine selected areas among areas with more than 4 observations, we draw the boxplots in Figure 1, which clearly indicates that the posted land price has the large within-area variation and the conventional NER model (which assumes homogeneity of variance) does not seem to be appropriate.

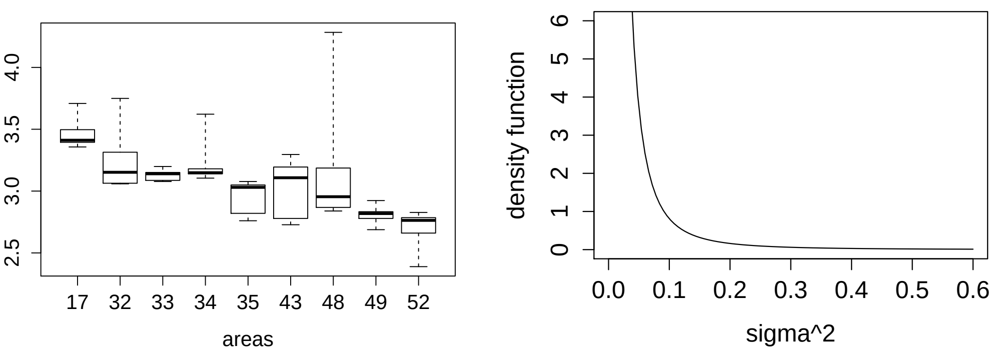


Figure 1: Boxplots of the Posted Land Price Data for Selected Areas (left) and the Estimated Density Function of $\sigma_i^2 = 1/\eta_i$ (right).

For $j = 1, \ldots, n_i$, y_{ij} denotes the value which is transformed by logarithm from the posted land price (Yen/10,000) for the unit meter squares of the j-th spot, T_i is the time to take from the nearby station i to the Tokyo station around 8:30 in the morning, D_{ij} is the value of geographical distance from the spot j to the station i and FAR_{ij} denotes the floor-area ratio, or ratio of building volume to lot area of the spot j. These values of T_i, D_{ij} and FAR_{ij} are transformed by logarithm. Since these data have within-area variability as indicated in Figure 1 (left), we use the RHNER model

$$y_{ij} = \beta_0 + FAR_{ij}\beta_1 + T_i\beta_2 + D_{ij}\beta_3 + v_i + \varepsilon_{ij}, \tag{5.1}$$

where v_i and ε_{ij} are mutually independent and distributed as $\mathcal{N}(0, \lambda\sigma_i^2)$ and $\mathcal{N}(0, \sigma_i^2)$, and $\eta_i (= 1/\sigma_i^2)$ is independently distributed as $\Gamma(\tau_1/2, 2/\tau_2)$.

The estimates of the parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \lambda, \tau_1, \tau_2)$ are

$$\widehat{\beta}_0 = 5.69,\ \widehat{\beta}_1 = 0.11,\ \widehat{\beta}_2 = -0.63,\ \widehat{\beta}_3 = -0.08,\ \hat{\lambda} = 0.22,\ \hat{\tau}_1 = 2.93,\ \hat{\tau}_2 = 0.04.$$

It is interesting to point out that the estimated regression function is a decreasing function of T_i and D_{ij}, which means that the land price y_{ij} tends to decrease as the time from Tokyo or distance from nearest station increases. Since $\hat{\tau}_1 = 2.93$ and $\hat{\tau}_2 = 0.04$, the distribution of η_i has a large mean about 73 and a heavy tail. Since the estimated value of τ_1 is smaller than 4, the variance of η_i or σ_i^2 does not exist, which agrees the observation that the posted land price data has great variability as indicated by the boxplots in Figure 1. Figure 1 (right) draws the estimated density function of $\sigma_i^2 = 1/\eta_i$ where η_i has $\Gamma(\hat{\tau}_1/2, 2/\hat{\tau}_2)$, so that the distribution of σ_i^2 has a small mean, but a heavy tail.

The predicted values of $\overline{\boldsymbol{x}}_i'\boldsymbol{\beta} + v_i$ and their conditional and unconditional MSE estimates, which can be obtained based on $1,000$ bootstrap samples, are given in Table 4. It is revealed from Table 4 that the estimates of the unconditional MSE get smaller as n_i gets larger. On the other hands, the estimates of the conditional MSE do not have a similar property, because the conditional MSE is affected by not only n_i but also the observed values as indicated in Table 3. It is interesting to point out that, in area 48, the estimated conditional MSE is relatively large while the estimated unconditional MSE is not large. Noting that this area has great variability as shown in Figure 1, it seems that the conditional MSE can capture the variability of areas.

Table 1: Values of EBLUP and Estimates of Unconditional and Conditional MSES for Selected fifteen Areas. (Estimates of MSE and cMSE are multiplied by 100)

Area	n_i	EBLUP	$\widehat{\text{MSE}}$	$\widehat{\text{cMSE}}$
1	1	4.02	5.19	0.58
4	2	3.91	4.34	0.49
5	5	3.96	3.13	2.31
8	3	3.86	3.83	0.33
17	7	3.50	2.66	1.04
25	7	3.39	2.65	1.37
26	4	3.45	3.42	1.88
32	6	3.22	2.86	2.68
33	8	3.12	2.48	1.90
34	11	3.16	2.09	1.10
35	7	2.99	2.65	3.58
43	6	3.02	2.86	3.73
48	6	3.07	2.86	5.11
49	10	2.82	2.21	2.69
52	6	2.76	2.87	6.55

References

[1] Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.*, **83**, 28-36.

[2] Booth, J. S. and Hobert, P. (1998). Standard errors of prediction in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **93**, 262 - 272.

[3] Butar, F.B. and Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small-area estimators. *J. Statist. Plan. Inf.*, **112**, 63-76.

[4] Datta, G.S. (2009). Model-based approach to small area estimation. In *Handbook of Statistics*, **29B**. Editors: D. Pfeffermann and C.R. Rao. North Holland, New York, pp 251-288.

[5] Datta, G.S., Rao, J.N.K. and Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, **92**, 183-196.

[6] Fay, R.E. and Herriot, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.*, **74**, 269-277.

[7] Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statist. Science*, **9**, 55-93.

[8] Jiang, J. and Nguyen, T. (2012). Small area estimation via heteroscedastic nested-error regression. *Canad. J. Statist.*, **40**, 588-603.

[9] Mardia, K.V. and Marshall, R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**, 135-146.

[10] Pfeffermann, D. (2002). Small area estimation: new developments and directions. *International Statistical Institute Review*, **70**, 125-143.

[11] Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *J. Amer. Statist. Assoc.*, **85**, 163-171.

[12] Rao, J.N.K. (2003). *Small Area Estimation.* Wiley.

[13] Sweeting, T.J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *Ann. Statist.*, **8**, 1375-1381.

乗法モデルとベンチマーク問題

川久保 友超 [1]
with Malay Ghosh[2], 久保川 達也 [3]

[1] 東京大学・経済・D2

[2]University of Florida

[3] 東京大学

2015 年 1 月 30 日
科研コンファレンス「経済統計・政府統計の数理的基礎と応用」
東京大学

はじめに

小地域推定

- ある予算を市町村（小地域）ごとに配分する際に，政策の根拠として各小地域の所得の平均を知りたいとする.
- しかし予算制約などの都合で小地域レベルでは十分なサンプルを得られない場合，標本平均では推定誤差が大きく，不適切な予算配分が行われる危険性がある.

→小地域ごとの推定精度を上げたい

はじめに

Fay–Herriot model（Fay and Herriot, 1979）

- 小地域 i の興味ある特性の direct estimator（標本平均など）を z_i とし，z_i を以下のような 2 段階のモデリングを行う．

$$
\begin{aligned}
&\textit{Level } 1: \ (\textit{sampling model}): \ z_i = \phi_i + \varepsilon_i; \\
&\textit{Level } 2: \ (\textit{linking model}): \ \phi_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + v_i,
\end{aligned}
$$

  ε_i は sampling error，v_i は地域効果（random effect）で，
  $\varepsilon_i \sim \mathcal{N}(0, d_i)$，$v_i \sim \mathcal{N}(0, \tau^2)$，ただし d_i は既知．

- モデルにもとづいて ϕ_i を推定する．
  → model based estimator:

$$
\hat{\phi}_i = \frac{\hat{\tau}^2}{d_i + \hat{\tau}^2} z_i + \frac{d_i}{d_i + \hat{\tau}^2} \boldsymbol{x}_i^t \widehat{\boldsymbol{\beta}}(\hat{\tau}^2)
$$

- direct estimator z_i を $\boldsymbol{x}_i^t \boldsymbol{\beta}$ の方向に縮小する作用がはたらき，小地域の推定が安定する．

はじめに

ベンチマーク法

- model based estimator の総和を，direct estimator の総和（ベンチマーク）に一致させる． → benchmarked estimator
  - model misspecification に対して頑健.
  - 官庁統計における要請（統計の整合性の観点から）.
- 制約付きベイズ推定量（Ghosh, 1992）が小地域の benchmarked estimator としてよく用いられる.

はじめに

本研究の概要

- 所得など正の値のみをとり歪んでいるデータに対し，対数変換したものが Fay–Herriot モデルとなるような乗法モデルとして，transformed Fay–Herriot model（Slud and Maiti, 2006）を考える.
- weighted Kullback–Leibler loss のもとで，benchmarked estimator を導出する．benchmarked estimator は non-benchmarked model based estimator の定数倍という自然なかたちとなる.
- 推定量のリスクの漸近不偏推定量を構成する.

モデルとベンチマーク制約

transformed Fay–Herriot model

- 各地域 $i = 1, \ldots, m$ に対して，正の direct estimarot y_i が乗法モデル $y_i = \theta_i \eta_i$ にしたがっている．推定の対象は θ_i.
- $z_i = \log y_i$，$\phi_i = \log \theta_i$，$\varepsilon_i = \log \eta_i$ とすると，z_i が以下の Fay–Herriot model にしたがう．

$$\begin{aligned} z_i =& \phi_i + \varepsilon_i, \quad \phi_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + v_i, \\ v_i \sim& \mathcal{N}(0, \tau^2), \quad \varepsilon_i \sim \mathcal{N}(0, d_i) \end{aligned}$$

  d_i は既知の定数，τ^2 は未知パラメータ，$v_1, \ldots, v_m, \varepsilon_1, \ldots, \varepsilon_m$ は互いに独立．
- $\boldsymbol{\beta}$ は事前分布 $\boldsymbol{\beta} \sim uniform(\mathbb{R}^p)$ をもつとする（階層ベイズモデル）．
- θ_i をベイズ推定し，階層ベイズ（HB）推定量 $\hat{\theta}_i^{\mathrm{HB}}$ を得る．
  $\rightarrow$（制約無し）model based estimator
- ベンチマーク制約のもとで事後リスク最小化を行い，制約付き階層ベイズ推定量 $\hat{\theta}_i^{\mathrm{CHB}}$ を得る．
  $\rightarrow$ benchmarked estimator

モデルとベンチマーク制約

ベンチマーク制約

$$\sum_{i=1}^{m} w_i \hat{\theta}_i = \sum_{i=1}^{m} w_i y_i$$

- 上式の制約下での事後リスク最小化問題の解が，制約付き階層ベイズ（CHB）推定量.
- $L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ を損失関数とすると，ラグランジュ関数は以下で定義される.

$$LM(\widehat{\boldsymbol{\theta}}, \lambda) = E[L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})|\mathbf{y}] + \lambda \left\{ \sum_{i=1}^{m} w_i \hat{\theta}_i - \sum_{i=1}^{m} w_i y_i \right\},$$

- 推定量は，損失関数の取り方に依存する.
- θ_i の推定に対して適切で，かつ CHB 推定量が陽に書け自然なかたちになる損失関数を提案したい.

モデルとベンチマーク制約

損失関数の選択

- 損失関数の候補をいくつか考える.

$$
\begin{aligned}
L_{\mathrm{Q}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= \sum_{i=1}^{m} \xi_i (\hat{\theta}_i - \theta_i)^2, \\
L_{\mathrm{TQ}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= \sum_{i=1}^{m} \xi_i (\log \hat{\theta}_i - \log \theta_i)^2, \\
L_{\mathrm{KL}}^{\xi}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= \sum_{i=1}^{m} \xi_i \left\{ \hat{\theta}_i/\theta_i - \log(\hat{\theta}_i/\theta_i) - 1 \right\},
\end{aligned}
$$

- L_{Q}：制約無しベイズ推定量は，事後平均 $E(\theta_i|\boldsymbol{y})$．しかし，CB 推定量が負値をとりうる．また尺度母数 θ_i の推定に二乗損失は適当でない．
- L_{TQ}：CB 推定量が陽に書けない．
- L_{KL}^{ξ}：制約無しベイズ推定量は，調和平均 $1/E(\theta_i^{-1}|\boldsymbol{y})$．

モデルとベンチマーク制約

weighted Kullback–Leibler loss

$$L^w_{\mathrm{KL}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sum_{i=1}^{m} w_i \theta_i \left\{ \hat{\theta}_i/\theta_i - \log(\hat{\theta}_i/\theta_i) - 1 \right\}$$

- weighted KL loss のもとでの θ_i の（制約無し）階層ベイズ推定量は，$\hat{\theta}_i^{\mathrm{HB}}(\tau^2) = E(\theta_i|\boldsymbol{y})$（事後平均）となる．
- 以下の $\hat{\theta}_i^{\mathrm{CHB}}$ は，weighted KL loss とベンチマーク制約のもとでの，制約付き階層ベイズ推定量．

$$\hat{\theta}_i^{\mathrm{CHB}}(\tau^2) = \hat{\theta}_i^{\mathrm{HB}}(\tau^2) \frac{\sum_{j=1}^m w_j y_j}{\sum_{j=1}^m w_j \hat{\theta}_j^{\mathrm{HB}}(\tau^2)}.$$

- τ^2 を推定量 $\hat{\tau}^2$ でおきかえて，
  - $\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2) \rightarrow$（制約無し）model based estimator
  - $\hat{\theta}_i^{\mathrm{CHB}}(\hat{\tau}^2) \rightarrow$ benchmarked estimator

リスクの推定

制約無し model based estimator の推定リスク

- weighted KL loss にもとづいた，$\hat{\theta}_i^{\mathrm{HB}}$ の推定リスクの 2 次漸近不偏推定量（$O(m^{-1})$ の項まで）を構成する．
- $\hat{\theta}_i^{\mathrm{HB}}(\tau^2) = E(\theta_i|\boldsymbol{y}, \tau^2)$ に気を付けると，リスクは以下に分解される．

$$
\begin{aligned}
& R_\omega\{\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2)\} \\
=& E[\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2) - \theta_i - \theta_i \log\{\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2)/\theta_i\}] \\
=& E[-\theta_i \log\{\hat{\theta}_i^B/\theta_i\}] + E[\hat{\theta}_i^{\mathrm{HB}}(\tau^2) - \theta_i - \theta_i \log\{\hat{\theta}_i^{\mathrm{HB}}(\tau^2)/\hat{\theta}_i^B\}] \\
& + E[\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2) - \hat{\theta}_i^{\mathrm{HB}}(\tau^2) - \theta_i \log\{\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2)/\hat{\theta}_i^{\mathrm{HB}}(\tau^2)\}] \\
=& I_1 + I_2 + I_3 \ (\mathrm{say}).
\end{aligned}
$$

ただし，$\hat{\theta}_i^{\mathrm{B}} = \hat{\theta}_i^{\mathrm{B}}(\boldsymbol{\omega}) = E(\theta_i|\boldsymbol{y}, \boldsymbol{\omega})$, $\boldsymbol{\omega} = (\boldsymbol{\beta}^t, \tau^2)^t$.

- I_1, I_2 は解析的な評価が比較的容易だが，I_3 は超パラメータ τ^2 の推定リスクを測っており評価が難しい．

→解析的な手法（テイラー展開）と数値的な手法（parametric bootstrap）

リスクの推定

制約無し model based estimator の推定リスク（解析的手法）

- $l_1 = g_{1i}(\omega)$, $l_2 = g_{2i}(\omega) + O(m^{-2})$, $l_3 = g_{3i}(\omega) + O(m^{-3/2})$ となり,

$$R_\omega\{\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2)\} = g_{1i}(\omega) + g_{2i}(\omega) + g_{3i}(\omega) + O(m^{-3/2}),$$

  とリスクが近似できる．ただし,
  $g_{1i}(\omega) = O(1)$, $g_{2i}(\omega) = O(m^{-1})$, $g_{3i}(\omega) = O(m^{-1})$.

- $g_{1i}(\widehat{\omega})$ は 2 次バイアス $b_{1i}(\omega) = O(m^{-1})$ が生じる．

$$E\{g_{1i}(\widehat{\omega})\} = g_{1i}(\omega) + b_{1i}(\omega) + O(m^{-3/2})$$

- これらの表現がすべて解析的に得られる場合,

$$\widehat{R}_\omega\{\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2)\} = g_{1i}(\widehat{\omega}) - b_{1i}(\widehat{\omega}) + g_{2i}(\widehat{\omega}) + g_{3i}(\widehat{\omega})$$

  としてリスクの 2 次漸近不偏推定量を構成できる．

リスクの推定

制約無し model based estimator の推定リスク（ブートストラップ）

- 推定量で構成した以下のモデルからブートストラップ標本を得る.
- $i = 1, \ldots, m$ に対して，正の観測 y_i^* が乗法モデル $y_i^* = \theta_i^* \eta_i^*$ にしたがっている.
- $z_i^* = \log y_i^*, \phi_i^* = \log \theta_i^*, \varepsilon_i^* = \log \eta_i^*$ とすると，z_i^* が以下の対数線形モデルにしたがう.

$$\begin{aligned} z_i^* &= \phi_i^* + \varepsilon_i^*, \quad \phi_i^* = \boldsymbol{x}_i^t \widehat{\boldsymbol{\beta}}(\hat{\tau}^2) + u_i^*, \\ u_i^* &\sim \mathcal{N}(0, \hat{\tau}^2), \quad \varepsilon_i^* \sim \mathcal{N}(0, d_i) \end{aligned}$$

ただし，$\widehat{\boldsymbol{\beta}}(\hat{\tau}^2)$ と $\hat{\tau}^2 = \hat{\tau}^2(\boldsymbol{y})$ は，データ $\boldsymbol{y}$ にもとづいた推定量.

- $l_1^* = \{g_{1i}(\widehat{\boldsymbol{\omega}})\}^2 / E_*\{g_{1i}(\widehat{\boldsymbol{\omega}}^*)\}$，$l_2^* = g_{2i}(\widehat{\boldsymbol{\omega}})$，さらに

$$l_3^* = E_*[\hat{\theta}_i^{\mathrm{HB}*}(\hat{\tau}^{2*}) - \hat{\theta}_i^{\mathrm{HB}*}(\hat{\tau}^2) - \hat{\theta}_i^{\mathrm{B}*}(\widehat{\boldsymbol{\omega}}) \log\{\hat{\theta}_i^{\mathrm{HB}*}(\hat{\tau}^{2*}) / \hat{\theta}_i^{\mathrm{HB}*}(\hat{\tau}^2)\}]$$

としたとき，$\widehat{R}^*\{\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2)\} = l_1^* + l_2^* + l_3^*$ がリスクの 2 次漸近不偏推定量.

リスクの推定

benchmarked estimator の推定リスク

- benchmarked estimator $\hat{\theta}_i^{\mathrm{CHB}}(\hat{\tau}^2)$ のリスクを推定する.

$$\hat{\theta}_i^{\mathrm{CHB}}(\hat{\tau}^2) = \hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2)\frac{\sum_{j=1}^m w_j y_j}{\sum_{j=1}^m w_j \hat{\theta}_j^{\mathrm{HB}}(\hat{\tau}^2)}.$$

- $\hat{\theta}_i^{\mathrm{CHB}}(\hat{\tau}^2)$ のリスクは $\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2)$ のリスクを用いて，以下のように表現できる.

$$\begin{aligned}
R_\omega\{\hat{\theta}_i^{\mathrm{CHB}}(\hat{\tau}^2)\} =& R_\omega\{\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2)\} - E\left[\hat{\theta}_i^{\mathrm{B}}(\omega)\log\left(\frac{\sum_{j=1}^m w_j y_j}{\sum_{j=1}^m w_j \hat{\theta}_j^{\mathrm{HB}}(\hat{\tau}^2)}\right)\right] \\
& + E\left[\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2)\left\{\frac{\sum_{j=1}^m w_j y_j}{\sum_{j=1}^m w_j \hat{\theta}_j^{\mathrm{HB}}(\hat{\tau}^2)} - 1\right\}\right] \\
=& R_\omega\{\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2)\} - K_1 + K_2.
\end{aligned}$$

- $\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2)$ と $\hat{\theta}_i^{\mathrm{CHB}}(\hat{\tau}^2)$ のリスクの差 $-K_1 + K_2$ を推定すればよい.

リスクの推定

benchmarked estimator の推定リスク（つづき）

- $-K_1 + K_2 = E[\widehat{K}] + K_3$ と書き換えられる．ただし，

$$\widehat{K} = \hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2)\left\{\frac{\sum_{j=1}^m w_j y_j}{\sum_{j=1}^m w_j \hat{\theta}_j^{\mathrm{HB}}(\hat{\tau}^2)} - \log\left(\frac{\sum_{i=1}^m w_i y_i}{\sum_{j=1}^m w_j \hat{\theta}_j^{\mathrm{HB}}(\hat{\tau}^2)}\right) - 1\right\},$$

$$K_3 = E\left[\left\{\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2) - \hat{\theta}_i^{\mathrm{B}}(\omega)\right\}\log\left\{\frac{\sum_{i=1}^m w_i y_i}{\sum_{j=1}^m w_j \hat{\theta}_j^{\mathrm{HB}}(\hat{\tau}^2)}\right\}\right].$$

- $\widehat{K}$ は $E[\widehat{K}]$ の不偏推定量だから，$K_3 = O(m^{-1})$ をパラメトリック・ブートストラップ法で推定する→ K_3^* を得る．
- $\widehat{R}^*\{\hat{\theta}_i^{\mathrm{CHB}}(\hat{\tau}^2)\} = \widehat{R}^*\{\hat{\theta}_i^{\mathrm{HB}}(\hat{\tau}^2)\} + \widehat{K} + K_3^*$ が $\hat{\theta}_i^{\mathrm{CHB}}(\hat{\tau}^2)$ のリスクの 2 次漸近不偏推定量．

実データにもとづく例

小地域の教育支出の平均推定

- 2011 年 11 月の各県庁所在地の教育支出の平均を推定する.
- 家計調査「教育支出」を direct estimator とする.
- 家計調査は毎月行われるが，小地域ではサンプル数が小さく推定が安定しない.
- そこで transformed Fay–Herriot model をあてはめ，（制約無しの）model based estimate, benchmarked estimate をそれぞれ求める．さらにリスクの推定も行う.
- 補助情報（モデルの説明変数）として，5 年に 1 度行われる全国消費実態調査の 2009 年の「教育支出」のデータを用いる．全国消費実態調査は大規模な調査のため，推定の精度が高い.

実データにもとづく例

- 各県庁所在地をあらわす添え字 $i = 1, \ldots, m, (m = 47)$ に対して，i 番目の都市の 2011 年 11 月の家計調査「教育支出」データ（単位は千円）を y_i とする（direct estimator），
- 説明変数として，i 番目の都市の 2009 年全国消費実態調査「教育支出」データを x_{i1} とする．
- $z_i = \log y_i$ に対し，以下のモデルをあてはめる．

$$\begin{aligned} z_i =& \phi_i + \varepsilon_i, \quad \phi_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + u_i, \\ u_i \sim& \mathcal{N}(0, \tau^2), \quad \varepsilon_i \sim \mathcal{N}(0, d_i) \end{aligned}$$

  ただし $\boldsymbol{x}_i^t = (1, x_{i1})$ とする．
- モデル上は誤差項の分散 d_i は既知だが，同じ都市の過去 10 年の 11 月の「教育支出」の標本分散で推定する．
- i 番目の都市の標本数を n_i とし，ウェイトは標本数で重みをつけた $w_i = n_i / \sum_{j=1}^{47} n_j$ を用いる．

実データにもとづく例

推定結果

- $\sum_{i=1}^m w_i y_i / \sum_{j=1}^m w_j \hat{\theta}_j^{\mathrm{HB}}(\hat{\tau}^2)$ の値は 1.0876 となった．よって，HB 推定値を 1.0876 倍すれば，CHB 推定値が得られる．
- direct estimator y_i のリスクも推定した．
- CHB 推定量は制約付きの事後リスク最小解なので，制約無しの事後リスク最小解である HB よりも，リスクは大きくなる．しかしほとんどの地域で y_i のリスクよりは小さい．

県	n_i	d_i	y_i	HB	CHB	$\widehat{R}_{\mathrm{HB}}$	$\widehat{R}^*_{\mathrm{HB}}$	$\widehat{R}^*_{\mathrm{CHB}}$	$\widehat{R}_{\boldsymbol{y}}$
茨城	95	9.56	8.10	8.89	9.65	19.5	19.1	22.4	45.6
栃木	95	26.96	10.03	9.48	10.29	23.6	23.4	27.3	134.7
群馬	94	8.41	5.21	7.71	8.37	19.7	19.4	22.3	42.0
埼玉	95	3.93	12.33	12.73	13.83	19.9	19.7	25.0	26.0
千葉	94	37.83	30.71	13.10	14.22	32.1	31.5	37.0	240.3
東京	386	2.30	15.45	14.45	15.68	12.4	12.7	18.3	14.2
神奈川	142	15.99	23.25	14.06	15.27	30.0	29.3	35.0	98.4

Table : HB 推定値，CHB 推定値とそれらのリスクの推定値

- Fay, R.E. and Herriot, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.*, **74**, 269-277.
- Ghosh, M. (1992). Constrained Bayes estimation with applications. *J. American Statist. Assoc.*, **87**, 533-540.
- Slud, E.V. and Maiti, T. (2006). Mean-squared error estimation in transformed Fay-Herriot models. *J. Royal Statist. Soc.*, **B 68**, 239-257.

データ変換と小地域推定

菅澤 翔之助

東京大学大学院経済学研究科修士 2 年

2015 年 1 月 30 日

経済統計・政府統計の基礎と応用

小地域推定 (設定・目的・方法)

設定

m 個のクラスター (地域) があり，各地域内で n_i 個 $(i = 1, \ldots, m)$ のサンプル $(y_{ij}, \mathbf{x}'_{ij})$, $j = 1, \ldots, n_i$ が得られている.

クラスター数 m は大きいが n_i はあまり大きくない.

目的

各地域の平均を精度良く推定したい.

($\bar{y}_i$ で推定した場合は n_i が小さい故に不安定な推定量になってしまう．)

手法

他の地域の情報や共変量の情報を「うまく」用いる.

具体的には Nested Error Regression Model(NERM) と呼ばれる以下のモデルを用いる.

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m$$

$v_i \sim N(0, \tau^2)$：変量効果, $\varepsilon_{ij} \sim N(0, \sigma^2)$：誤差項, v_i, ε_{ij} は互いに独立.

個々の地域の特徴を v_i で表現し全体としての特徴を $\boldsymbol{\beta}$ で表現している.

小地域推定 (設定・目的・方法)

前スライドのモデルをもとに

$$\mu_i = \bar{\mathbf{x}}_i' \boldsymbol{\beta} + v_i$$

の予測量を構成すると

$$\hat{\mu}_i = \bar{\mathbf{x}}_i' \hat{\boldsymbol{\beta}} + \frac{n_i \hat{\tau}^2}{\hat{\sigma}^2 + n_i \hat{\tau}^2} (\bar{y}_i - \bar{\mathbf{x}}_i' \hat{\boldsymbol{\beta}})$$

となる.

これを各地域の平均の推定量として用いることで安定した推定を行うことができる.

実は NERM は線形混合モデル (LMM) と呼ばれるモデルのクラスに属する．

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \varepsilon,$$
$$\mathbf{b} \sim N_r(\mathbf{0}, \mathbf{G}), \quad \varepsilon \sim N_N(\mathbf{0}, \mathbf{R})$$

このモデルにおいて $\mu = \mathbf{c}'\boldsymbol{\beta} + \mathbf{d}'\mathbf{b}$ の予測量は

$$\hat{\mu} = \mathbf{c}'\hat{\boldsymbol{\beta}} + \mathbf{d}'\hat{\mathbf{G}}\mathbf{Z}'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

で与えられる．(EBLUP: 経験最良不偏予測量)

問題提起

小地域推定に限らず地域毎 (クラスター毎) に観測されたデータにおいて y_{ij} が正の値となるデータは頻出.

LMM は観測 $\mathbf{y}$ に正規分布を仮定しているため，正のデータに対してそのまま適用するのは妥当でない.

そのような状況では対数変換を用いて $\log y_{ij}$ に対して LMM を当てはめるが一般的.

本研究の目的

Box-Cox(BC) 変換を用いた LMM を導入し，そのモデルに対する推定および予測を考え，小地域推定へ応用する.

Box-Cox 変換

$$h(x,\lambda) = (x^\lambda - 1)/\lambda, \quad \lambda \neq 0, \qquad h(x,0) = \log x$$

$\Rightarrow$ 対数変換 ($\lambda = 0$) と恒等変換 ($\lambda = 1$) を含む.

変換パラメータもデータから推定することによって柔軟なモデリングが可能になる.

本研究の目的

具体的な研究の目的は以下の通り.

- BC 変換を取り入れた LMM を提案する.

  特に小地域推定で NERM を用いた応用を視野にいれている.

- すべてのパラメータの一致推定量を構成する.

  実は BC 変換で最尤法により λ を推定すると一致性を持たないのでその点を解消する.

- 予測量を構成し，その不確実性を測るために予測区間を構成する.

  予測量のリスクを測ることは応用上非常に重要.

BC-LMM

次のような Box-Cox transformed LMM (BC-LMM) を考える.

$$h(\mathbf{y}_i, \lambda) = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, m$$

$\mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}^q$ および $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N \in \mathbb{R}^{n_i}$ は互いに独立に
$\mathbf{b}_i \sim N_q(0, \mathbf{R}(\boldsymbol{\psi}))$, $\boldsymbol{\varepsilon}_i \sim N_{n_i}(0, \mathbf{G}_i(\boldsymbol{\psi}))$.

$\boldsymbol{\psi}$ は分散パラメータ.

$q = 1$ として $\mathbf{Z}_i = \mathbf{1}_{n_i}$ とすると以下のような Box-Cox 変換を取り入れた NERM(BC-NERM) が得られる.

$$h(y_{ij}, \lambda) = \mathbf{x}_{ij}' \boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad j = 1, \ldots, n_i, \;\; i = 1, \ldots, m$$

パラメータ推定

λ が所与のもとでは通常の LMM の理論が使える.

$\Rightarrow \boldsymbol{\beta}, \boldsymbol{\psi}$ ともに λ の関数として $\hat{\boldsymbol{\beta}}(\lambda), \hat{\boldsymbol{\psi}}(\lambda)$ として表せる.

変換パラメータ λ の推定が本質的.

以下の方程式の解として推定.

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\{ h(y_{ij}, \lambda) - \mathbf{x}'_{ij} \hat{\boldsymbol{\beta}}(\lambda) \right\}^3 = 0,$$

「なせ変換するか？—分布を正規分布に近付けるため」

ということを考えると自然な推定手法.

パラメータ推定

λ が推定できれば $\boldsymbol{\beta}, \boldsymbol{\psi}$ は plug-in して $\hat{\boldsymbol{\beta}}(\hat{\lambda}), \hat{\boldsymbol{\psi}}(\hat{\lambda})$ と推定できる.

以上のように得られた推定量に対して以下が示せる.

$\theta = (\boldsymbol{\beta}', \boldsymbol{\psi}', \lambda)'$ とすると

$$\hat{\theta} - \theta = O_p(m^{-1/2}), \quad E[\hat{\theta} - \theta] = O(m^{-1})$$

λ の推定に対して ML を用いるとこの結果は得られない.

予測

以下の量の予測を考える．

$$h^{-1}(\mu, \lambda) = (\lambda\mu + 1)^{1/\lambda}, \qquad \mu = \mathbf{c}_1'\boldsymbol{\beta} + \mathbf{c}_2'\mathbf{b},$$

ただし $\mathbf{b} = (\mathbf{b}_1', \ldots, \mathbf{b}_m')'$ で $\mathbf{c}_1, \mathbf{c}_2$ は既知の定数ベクトル．

(固定効果 β と変量効果 b の線形結合 μ を逆変換した量)

μ の EBLUP は

$$\hat{\mu}^{\mathrm{EBLUP}} = \mathbf{c}_1'\hat{\boldsymbol{\beta}} + \mathbf{c}_2'\hat{\mathbf{b}}^{\mathrm{BP}}(\hat{\psi}, \hat{\lambda}),$$

ただし

$$\hat{\mathbf{b}}_i^{\mathrm{BP}}(\psi, \lambda) = \mathbf{R}(\psi)\mathbf{Z}_i'\boldsymbol{\Sigma}_i(\psi)^{-1}\left\{h(\mathbf{y}_i, \lambda) - \mathbf{X}_i\boldsymbol{\beta}\right\}.$$

$\Rightarrow$ $h^{-1}(\mu, \lambda)$ を $h^{-1}(\hat{\mu}^{\mathrm{EBLUP}}, \hat{\lambda})$ で予測する．

予測区間

前スライドの予測量 $h^{-1}(\hat{\mu}^{\mathrm{EBLUP}}, \hat{\lambda})$ に対してそのリスクを見積もるための予測区間を構成する.

$\Rightarrow$ plug-in による naive な方法 —— $O(m^{-1})$ の精度.

$\Rightarrow$ parametric bootstrap による方法 —— $O(m^{-3/2})$ の精度.

以下の分布を近似するのが本質的.

$$T = \hat{\sigma}^{-1}\left\{\frac{(\lambda\mu+1)^{\hat{\lambda}/\lambda}-1}{\hat{\lambda}} - \hat{\mu}^{\mathrm{EBLUP}}\right\}$$

$$\sigma^2 = \mathbf{c}_2'\mathrm{diag}\,(\mathbf{V}_1, \ldots, \mathbf{V}_m)\mathbf{c}_2, \quad \mathbf{V}_i = \mathbf{R}(\psi) - \mathbf{R}(\psi)\mathbf{Z}_i'\boldsymbol{\Sigma}_i^{-1}(\psi)\mathbf{Z}_i\mathbf{R}(\psi).$$

この統計量は $m \to \infty$ で $N(0,1)$ に収束する (ように構成されている).

予測区間

前スライドの T の分布の分位点 q_L, q_U がわかれば

$$I_m = \left[\left\{ \hat{\lambda}(\hat{\mu}^{\mathrm{EBLUP}} + q_1 \hat{\sigma}) + 1 \right\}^{1/\hat{\lambda}}, \left\{ \hat{\lambda}(\hat{\mu}^{\mathrm{EBLUP}} + q_2 \hat{\sigma}) + 1 \right\}^{1/\hat{\lambda}} \right]$$

のように予測区間が求まる.

T の分布を N(0,1) で近似 $\Rightarrow O(m^{-1})$ の精度

T の分布を parametric bootstrap で近似 $\Rightarrow O(m^{-3/2})$ の精度

Simulation

提案した推定量の MSE を計算し,ML で λ を推定したケースとの比較.

データ生成過程：

$$h(y_{ij}, \lambda) = \beta_0 + \beta_1 x_{ij} + v_i + \varepsilon_{ij}, \quad i = 1, \ldots, 15, \quad j = 1, \ldots, 5$$

$v_i \sim N(0, \boldsymbol{\sigma}_1^2)$, $\varepsilon_{ij} \sim N(0, \boldsymbol{\sigma}_2^2)$, $\boldsymbol{\sigma}_1 = 1, \ \ \boldsymbol{\sigma}_2 = 1.5$

$x_{ij} \sim U(4, 8)$. (最初に発生させて各 run では固定)

$\beta_0 = 1, \ \beta_1 = 2.$

Simulation

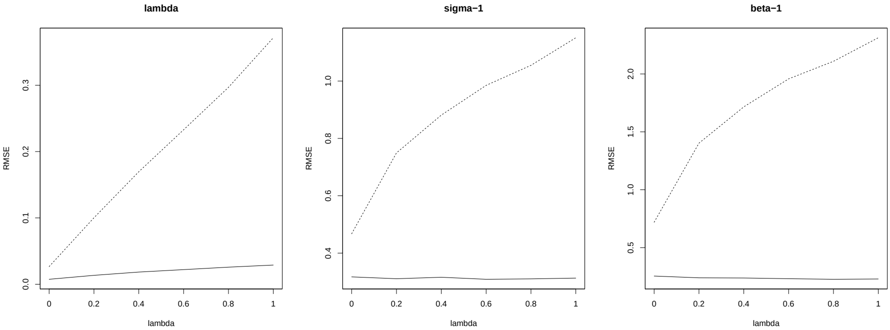


Figure: 5000 回の simulation から計算した推定量の MSE. 実線が提案推定量であり破線が λ を ML で推定したケース.

λ の ML による推定は精度悪く，それが別のパラメータの推定にも波及してしまう.

実データへの応用

2001 年の京急線の公示地価 (PLP) データに BC-NERM を適用する.

y_{ij} : 公示地価 (1000 円単位)
FAR_{ij} : 容積率
DST_{ij} : 最寄り駅からの距離
TRN_{i} : 駅 i から東京駅までの所用時間 (分)

$$h(y_{ij}, \lambda) = \beta_0 + \beta_1 \log(\mathrm{FAR}_{ij}) + \beta_2 \log(\mathrm{TRN}_i) + \beta_3 \log(\mathrm{DST}_{ij}) + v_i + \varepsilon_{ij},$$

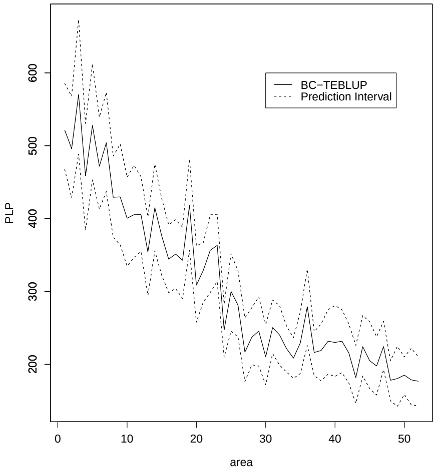
$v_i \sim N(0, \tau^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$

パラメータ推定値は

$\hat{\lambda}$	$\hat{\tau}$	$\hat{\boldsymbol{\sigma}}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
0.43	0.67	2.21	49.0	2.05	−7.10	−1.18

実データへの応用

1,000 回のブートストラップにより予測区間を構成すると以下のようになる.



まとめ

本研究では Box-Cox 変換を取り入れた LMM(BC-LMM) を考え，その推定および予測問題について考えた．

$\Rightarrow$ 変換パラメータ λ についてシンプルな推定手法を提案し，その漸近的性質を示した．

$\Rightarrow$ BC-LMM における予測量を提案し，そのリスク評価のための予測区間を構成する手法を提案した．

$\Rightarrow$ BC-LMM の具体例として BC-NERM を考え，PLP データに応用した．

観測誤差と線形制約を伴う真のデータの推定に関する新たな近接法（未定稿）

千木良 弘朗　　山本 拓

東北大学大学院経済学研究科　日本大学経済学部

研究集会「経済統計・政府統計の数理的基礎と応用」

2015 年 1 月 30 日 東京大学

1 はじめに

“観測誤差と線形制約を伴う真のデータ” とは?

例. 国民経済計算における供給表と使用表

供給表:

商品＼産業	製造業	サービス業	総計
製造品	900	100	1000
サービス	300	500	800
総計	1200	600	

使用表:

商品＼産業	製造業	サービス業	消費	総計
製造品	150	50	750	950
サービス	250	100	250	600
賃金	400	250		
営業余剰	350	100		
総計	1150	500		

$\Rightarrow$ 　　　、　　　、　　　、　　　 は等しくなければならない

$\Rightarrow$ 真のデータでは等しいが、観測誤差のために等しくなくなっている

真のデータの推定

供給表:

商品 ＼ 産業	製造業	サービス業	総計
製造品	880	90	970
サービス	260	460	720
総計	1140	550	

使用表:

商品 ＼ 産業	製造業	サービス業	消費	総計
製造品	160	40	770	970
サービス	270	150	300	720
賃金	410	280		
営業余剰	300	80		
総計	1140	550		

⇒ 、 、 、 を等しくするよう観測誤差を修正する

$\Rightarrow$ **修正法は国によって違う。アメリカとオランダは統計学の理論に基づいたストーン法（Stone，Champernowne and Meade (1942)）を採用**

$\Rightarrow$ **本稿では、ストーン法も部分的に使うが、全く新しいアプローチを導入する**

発表の構成

2. ストーン法と先行研究の概観
3. 新しいアプローチ
4. モンテカルロ実験
5. まとめ

3 ストーン法と先行研究の概観

モデル

$$\underset{(N\times 1)}{\boldsymbol{X}} = \boldsymbol{X}^* + \boldsymbol{u},\ \boldsymbol{u} \sim (\boldsymbol{0}, \boldsymbol{\Omega}_{\boldsymbol{u}})$$

$$\underset{(R\times N)}{\boldsymbol{A}'} \boldsymbol{X}^* = \boldsymbol{h}$$

$$\begin{cases} \boldsymbol{X}\text{: 観測されたデータ} \\ \boldsymbol{X}^*\text{: 真のデータ} \\ \boldsymbol{u}\text{: 観測誤差} \\ \boldsymbol{A}\text{、}\boldsymbol{h}\text{: } \boldsymbol{X}^*\text{が満たすべき線形制約} \end{cases}$$

例．国民経済計算における供給表と使用表

供給表:

商品 ＼ 産業	製造業	サービス業	総計
製造品	x_1	x_2	
サービス	x_3	x_4	
総計			

使用表:

商品 ＼ 産業	製造業	サービス業	消費	総計
製造品	x_5	x_6	x_7	
サービス	x_8	x_9	x_{10}	
賃金	x_{11}	x_{12}		
営業余剰	x_{13}	x_{14}		
総計				

$\rightarrow A'X^* = h$ は、

$$\begin{bmatrix} 1 & 1 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1^* \\ \vdots \\ x_{14}^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

ストーン法

$$\min(X - X^*)'\Omega_u^{-1}(X - X^*),\ \text{s.t.} A'X^* = h$$

$$\to \boxed{\widehat{X}^* = X - \Omega_u A(A'\Omega_u A)^{-1}(A'X - h)}$$

別の解釈:
簡単化のため $h = 0$ とすると制約条件は $A'X^* = 0$ で

$$\begin{aligned}
\widehat{X}^* &= X - \Omega_u A(A'\Omega_u A)^{-1}A'X \\
&= (I - \tilde{A}(A'\tilde{A})^{-1}A')X \qquad (\tilde{A} = \Omega_u A) \\
&= A_\perp(\tilde{A}'_\perp A_\perp)^{-1}\tilde{A}'_\perp X \qquad (\text{“}\perp\text{” は直行補空間})
\end{aligned}$$

→ **X^* が $A_\perp$ に属している（フルランク N ではなくランク落ちして $N - R$）ため、X も $A_\perp$ に属するよう変換しているのがストーン法**

Ω_uの設定について

ストーン法はΩ_uが解らないと実行不能

→ **統計理論に基づいた解決法はそう多くなく、ある程度恣意的に決められることが多い**

		業種						
		農林水産業	鉱業	製造業	建　設　業	電気・ガス・水道業	卸売・小売業	金融・保険業
商品	農林水産業	1,981.2	1.5	8,412.4	250.1	1.7	1,605.1	0.0
	鉱業	0.3	8.4	8,506.8	1,037.0	2,336.4	3.5	0.0
	製造業	3,510.3	281.2	138,525.2	28,727.9	1,669.4	5,702.9	1,517.6
	建　設　業	85.5	9.7	1,410.0	208.1	1,131.2	554.8	160.1
	電気・ガス・水道業	106.8	43.3	6,522.2	500.4	1,335.1	1,128.4	220.5
	卸売・小売業	4.0	0.0	0.0	0.0	0.0	672.0	0.0
	金融・保険業	137.3	42.1	1,419.1	480.4	224.5	1,723.5	1,150.0

野木森 稔, 「加重最小二乗法を利用したバランシング・モデル-SUT バランスシステム開発に向けた一考察」, 季刊国民経済計算, 147, 69–89, 2012 年
より抜粋

○ 統計理論に基づいた解決法

Weale (1992) 等のアイデア:

$$X_t = X_t^* + u_t,\ u_t \sim (0, \Omega_u),\ t = 1, \cdots, T$$

$$\begin{cases} X_t、X_t^*、u_t \text{は定常} \\ Cov(X_t^*, u_t) = 0 \end{cases}$$

とすると、簡単化のため $E(X_t^*) = 0$ として

$$\begin{aligned} \boxed{\widehat{A'\Omega_u}} &= \frac{1}{T}\sum_{t=1}^{T} A' X_t X_t' \\ &= \frac{1}{T}\sum_{t=1}^{T} A' u_t (X_t^* + u_t)' \xrightarrow{p} \boxed{A'\Omega_u} \end{aligned}$$

$\Rightarrow$ Ω_u ではなく $A'\Omega_u$ の一致推定でストーン法を実行可能にしている

先行研究の問題点と本稿での新しいアプローチ

問題点:

$$\begin{cases} \widehat{A'\Omega_u}\text{は}N << T\text{でないとパフォーマンスが悪い} \\ X_t^*\text{に定常性を仮定するのは非現実的} \\ \text{ストーン法では}u_t\text{を一致推定できない} \\ \Omega_u\text{は一致推定できない} \end{cases}$$

$\Rightarrow$ **新しいアプローチ:**

$$\begin{cases} \text{大きな}N\text{に対して良好なパフォーマンス} \\ X_t^*\text{に定常性を仮定しない} \\ u_t\text{（の一部）を一致推定する} \\ \Omega_u\text{を一致推定する} \end{cases}$$

3 新しいアプローチ

モデル

観測誤差にファクターモデルを適用する

$$X_t = X_t^* + u_t = X_t^* + \underset{(N\times q)}{C}\underset{(q\times 1)}{F_t} + \varepsilon_t,\ t = 1, \cdots, T$$

$$\begin{cases} C\,\text{はファクター負荷} \\ F_t\text{はファクター} \\ q\,\text{はファクター数} \end{cases}$$

$\Rightarrow$ **CF_tを一致推定して観測誤差を修正する。ε_tについてはストーン法で修正**

（簡単化のため、以下では$h = 0$とする）

モデルの推定の直感的な概観

$$X_t = X_t^* + CF_t + \varepsilon_t,\ \varepsilon_t \sim (0, \Omega),\ \Omega = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_N^2 \end{bmatrix}$$

$\Downarrow$

$$A_N' X_t = A_N' C F_t + A_N' \varepsilon_t,\ A_N = \underset{(R\times R)}{A\ \Upsilon_N^{-1}},\ \Upsilon_N = \begin{bmatrix} n_1 & & 0 \\ & \ddots & \\ 0 & & n_R \end{bmatrix},$$

n_rはAの第r列のnon-zero要素の数

$\Downarrow$

この式にBai (2003)の方法を適用して$A_N'\widehat{CH'}^{-1}$と$\widehat{H'F}_t$を得る

$$\Downarrow$$

$Cov(X_t^*, \varepsilon_t) = 0$、$Cov(F_t, \varepsilon_t) = 0$を仮定すると

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} \Upsilon_N(A'_N X_t - A'_N \widehat{CH'}^{-1}\widehat{H'}F_t)X'_t &\approx \frac{1}{T}\sum_{t=1}^{T}\widehat{A'}\varepsilon_t X'_t \\ &\approx \frac{1}{T}\sum_{t=1}^{T}\widehat{A'}\varepsilon_t\varepsilon'_t \\ &\approx A'\Omega\end{aligned}$$

$$\Downarrow$$

この式を積率条件としてGMMで$\hat{\sigma}_1^2, \cdots, \hat{\sigma}_N^2$を得る

$$\Downarrow$$

$$\frac{1}{T}\sum_{t=1}^{T} A'_N X_t X'_t \approx A'_N C \Omega_F C' + A'_N \Omega \qquad (Cov(X^*_t, F_t) = 0\text{を追加})$$

$$= A'_N C H'^{-1} H' \Omega_F H H^{-1} C' + A'_N \Omega$$

において、

$$\underbrace{\frac{1}{T}\sum_{t=1}^{T} A'_N X_t X'_t}_{\text{計算可能}} = \underbrace{A'_N C H'^{-1} H' \Omega_F H}_{\widehat{A'_N C H'^{-1}}\text{ と }\widehat{H' F_t}\text{ で推定可能}} \underbrace{H^{-1} C'}_{\text{未知パラメーター}} + \underbrace{A'_N \Omega}_{\hat{\sigma}_1^2, \cdots, \hat{\sigma}_N^2\text{ で推定可能}}$$

$$\Downarrow$$

この式から $H^{-1}C'$ を逆算して $\widehat{H^{-1}C'}$ を得る

仮定:

1. $E(F_t)=0$、$F_t, \varepsilon_t \sim I(0)$

2. $\boxed{A\text{の第}r\text{列と第}s\text{列で共通する項目の数}c_{rs}\text{が十分小さい}}$

（全てのrについて$\sum_{s=1}^{R} c_{rs}$が$R\to\infty$としても発散しない）

3. $\boxed{q<R}$

4. Aの全行には少なくとも1つnon-zero要素がある

5. $\boxed{X_t^* \sim I(0)}$

6. $N, R, T \to \infty$ $\boxed{\text{with } \frac{N^K}{\sqrt{T}R^{1-W}} \to 0}$

（$\sum_{r=1}^{R} n_R = O(N^K)$, $1 \le K \le 2$、
$n_1, \cdots, n_R$の内で$O(1)$なものの個数を$O(R^W)$, $0 \le W \le 1$）

$$\widehat{CH'^{-1}}\widehat{H'F_t} \xrightarrow{p} CF_t \quad \text{（観測誤差（の一部）の一致推定）}$$

$$\hat{\sigma}_1^2, \cdots, \hat{\sigma}_N^2 \xrightarrow{p} \sigma_1^2, \cdots, \sigma_N^2 \quad \text{（ストーン法の改善）}$$

実際のprocedure

ファクター-ストーン法:

$\widetilde{X_t^*} = X_t - \widehat{CH'^{-1}}\widehat{H'F_t}$に、$\hat{\Omega} = \begin{bmatrix} \hat{\sigma}_1^2 & & 0 \\ & \ddots & \\ 0 & & \hat{\sigma}_N^2 \end{bmatrix}$を使ってストーン法を適用

$\rightarrow \widehat{X_t^*} = \widetilde{X_t^*} - \hat{\Omega}A(A'\hat{\Omega}A)^{-1}A'\widetilde{X_t^*}$

4 モンテカルロ実験

DGPと実験の設定

$$X_t = X_t^* + u_t = X_t^* + CF_t + \varepsilon_t$$

$$\begin{cases} X_t^* \sim \text{トレンド定常VAR(1)} \\ F_t \sim \text{期待値0の定常VAR(1)} \qquad (C\Omega_F C' / Var(X_t^*) \approx 5\%) \\ \varepsilon_t \sim \text{期待値0の定常VAR(1)} \qquad (\Omega / Var(X_t^*) \approx 5\%) \\ X_t^*、F_t、\varepsilon_t \text{は全て正規乱数で互いに独立} \end{cases}$$

$\Rightarrow Var(X_t)$の内、$Var(X_t^*)$が約90%でΩ_uが約10%

$Var(X_t^*)$、$C\Omega_F C'$、Ω全てにおいて最小の分散と最大の分散には約500〜1000倍の差をつけ、不均一性を表現

- X_t^* のVAR(1)パラメーターと定数・トレンド項は乱数で生成して固定
- F_t と ε_t のVAR(1)パラメーターは乱数で生成して固定
- C は乱数で生成して固定
- A は $-1, 0, 1, 2$ しかとらないsparse行列として乱数で生成して固定
- $q = 2$ で既知
- $N = 25, 100, 400$、$R = N/4, \sqrt{N}$、$T = 25, 50, 800$
- 繰り返し回数1000回

各修正法の具体的な計算

○ 原始的なストーン法（Stone1）

$$\Omega_u = I_N \text{として} \widehat{X_t^*} = X_t - A(A'A)^{-1}A'X_t$$

○ Weale (1992) の方法（Weale）

demean、detrend した X_t^d より $\widehat{A'\Omega_u} = \frac{1}{T}\sum_{t=1}^{T} A'X_tX_t^{d\prime}$ として $\widehat{X}_t^* = X_t - \widehat{\Omega_u A}(\widehat{A'\Omega_u}A)^{-1}A'X_t$

○ ファクター-ストーン法（F-Stone）

先述の通りだが、

$$\begin{cases} \frac{1}{T}\sum_{t=1}^{T} \Upsilon_N(A_N'X_t - A_N'\widehat{CH'}^{-1}\widehat{H'F}_t)X_t^{d\prime} \approx A'\Omega \\ \frac{1}{T}\sum_{t=1}^{T} A_N'X_tX_t^{d\prime} \approx A_N'CH'^{-1}H'\Omega_F HH^{-1}C' + A_N'\Omega \end{cases}$$

のように X_t^d で計算する。また、$\hat{\sigma}_i^2 \leq 0$ だった時は $\hat{\sigma}_i^2 = 0.01 \times \frac{1}{T}\sum_{t=1}^{T}(x_{it}^d)^2$ とする

推定精度の尺度

MSE$= E\{(X_t^* - \widehat{X}_t^*)'(X_t^* - \widehat{X}_t^*)\}$だと、

- 変動の大きい$x_{it}^* - \hat{x}_{it}^*$だけにdominateされてしまい不公平
- NやRを変えるとDGPのパラメーターの値の変化を反映してしまい、純粋なsample sizeの効果を捉えられない

$\Rightarrow$ 基準化MSE（sMSE）：

$$\text{sMSE}= \frac{1}{N}E\left\{(X_t^* - \widehat{X}_t^*)'\begin{bmatrix} Var(x_1^*) & & 0 \\ & \ddots & \\ 0 & & Var(x_N^*)\end{bmatrix}^{-1}(X_t^* - \widehat{X}_t^*)\right\}$$

$$= \frac{1}{N}E\left\{\sum_{i=1}^{N}\frac{(x_{it}^* - \hat{x}_{it}^*)^2}{Var(x_i^*)}\right\}$$

$$\Rightarrow \begin{cases} x^*_{it} = \mu_i + \delta_i t + z_{it},\ E(z_{it}) = 0,\ z_{it} \sim I(0) \\ \hat{x}^*_{it} = \mu_i + \delta_i t \quad \text{（確定項だけ捉えている“最低限”の推定量）} \end{cases}$$

とすると、$E\left\{\frac{(x^*_{it}-\hat{x}^*_{it})^2}{Var(x^*_i)}\right\} = E\left\{\frac{z^2_{it}}{Var(x^*_i)}\right\} = 1$

$$\Rightarrow \begin{cases} \text{sMSE} < 1 \ \text{•••最低限の規準はクリア} \\ \text{sMSE} \geq 1 \ \text{•••推定精度に問題あり} \end{cases}$$

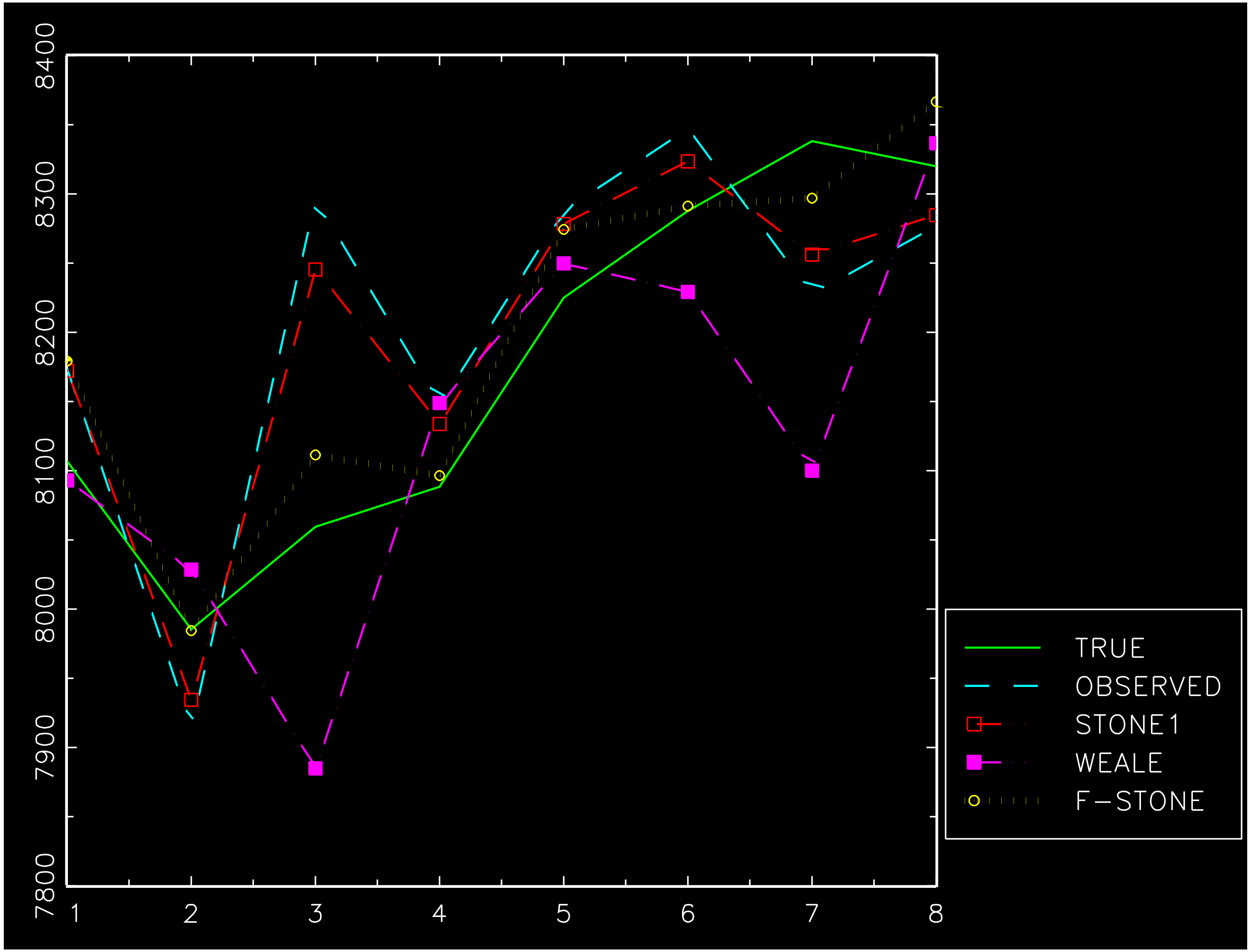
（実際はsMSEが$t = 1, \cdots, T$について得られるので$\frac{1}{T}\sum_{t=1}^{T} \text{sMSE}_t$を計算）

sMSE

		Stone1			Weale			F-Stone		
N	R	T			T			T		
		25	50	800	25	50	800	25	50	800
25	5	0.137	0.136	0.136	0.294	0.178	0.062	0.144	0.109	0.064
	7	0.152	0.152	0.151	0.398	0.222	0.060	0.143	0.106	0.057
100	10	0.138	0.139	0.139	0.599	0.307	0.068	0.144	0.110	0.059
	25	0.192	0.192	0.193	.	0.632	0.081	0.147	0.106	0.052
400	20	0.116	0.115	0.115	2.005	0.524	0.081	0.135	0.105	0.053
	100	0.137	0.138	0.138	.	.	0.195	0.139	0.104	0.045

実験データの時系列プロット

$N = 100$、$R = 25$、$T = 50$



5 まとめ

本稿の貢献

観測誤差を伴う真のデータの推定について、

$\left\{ \begin{array}{l} \boxed{\text{観測誤差（の一部）の一致推定による直接的な推定}} \\ \boxed{\text{（それ以外の観測誤差に関する）修正精度の改善}} \end{array} \right.$

を提案した

推定精度

sMSE という尺度では、概ね $\boxed{\text{F-Stone} > \text{Stone1} > \text{Weale}}$

トレンド・季節性とマクロ時系列：(非定常)変数誤差問題
(Trend, seasonality and macro time series : Nonstationary errors-in-variables)

国友直人・佐藤整尚

2015年1月30日

1. A General Problem

Let y_{ij} be the $i-$th observation of the $j-$th time series at t_i^n for $i = 1, \cdots, n; j = 1, \cdots, p; 0 = t_0^n \leq t_1^n \leq \cdots \leq t_n^n = T$. We set $n = T$, $t_i^n - t_{i-1}^n = 1$, $\mathbf{y}_i = (y_{1i}, \cdots, y_{pi})^{'}$ be a $p \times 1$ vector and $\mathbf{Y}_n = (\mathbf{y}_i^{'})$ $(= (y_{ij}))$ be an $n \times p$ matrix of observations. ($\mathbf{y}_0$ is the initial observation vector.) We consider the situation when the underlying non-stationary trends $\mathbf{x}_i$ $(= (x_{ji}))$ at t_i^n $(i = 1, \cdots, n)$ are not necessarily the same as the observed time series and let $\mathbf{c}_i^{'} = (c_{1i}, \cdots, c_{pi})$, $\mathbf{s}_i^{'} = (s_{1i}, \cdots, s_{pi})$ and $\mathbf{v}_i^{'} = (v_{1i}, \cdots, v_{pi})$ be the vectors of the seasonal components, the cycle components and the noise components at t_i^n, respectively, which are independent of $\mathbf{x}_i$. Then we use the additive decomposition (see Kitagawa (2010))

$$\mathbf{y}_i = \mathbf{x}_i + \mathbf{s}_i + \mathbf{c}_i + \mathbf{v}_i \ , \tag{1.1}$$

where $\mathbf{x}_i$ are a sequence of non-stationary trend components satisfying $\Delta \mathbf{x}_i = (1 - \mathcal{L})\mathbf{x}_i = \mathbf{w}_i^{(x)}$ (with $\mathcal{L}\mathbf{x}_i = \mathbf{x}_{i-1}$, $\mathcal{E}(\mathbf{w}_i^{(x)}) = \mathbf{0}$, $\mathcal{E}(\mathbf{w}_i^{(x)}\mathbf{w}_i^{(x)'}) = \mathbf{\Sigma}_x$), $\mathbf{s}_i$ are a sequence of seasonal components satisfying $(1 + \mathcal{L} + \cdots + \mathcal{L}^{s-1})\mathbf{s}_i = \mathbf{w}_i^{(s)}$ (with $\mathcal{E}(\mathbf{w}_i^{(s)}) = \mathbf{0}$, $\mathcal{E}(\mathbf{w}_i^{(s)}\mathbf{w}_i^{(s)'}) = \mathbf{\Sigma}_s$), $\mathbf{c}_i$ are a sequence of stationary cycle components (with $\mathcal{E}(\mathbf{c}_i) = \mathbf{0}$, $\mathcal{E}(\mathbf{c}_i\mathbf{c}_i^{'}) = \mathbf{\Sigma}_c$) and $\mathbf{v}_i$ are a sequence of independent noise components (with $\mathcal{E}(\mathbf{v}_i) = \mathbf{0}$, $\mathcal{E}(\mathbf{v}_i\mathbf{v}_i^{'}) = \mathbf{\Sigma}_v$).
We assume that $\mathbf{w}_i^{(x)}, \mathbf{w}_i^{(s)}$ and $\mathbf{v}_i$ are the sequence of i.i.d. random variables with $\mathbf{\Sigma}_v$ being positive definite and finite, and the random variables $\mathbf{w}_i^{(x)}, \mathbf{w}_i^{(s)}$, $\mathbf{c}_i$ and $\mathbf{v}_i$ are mutually independent. We also set $\mathbf{c}_i = \mathbf{0}$ in this occasion, but we can generalize the following arguments into more general directions with $\mathbf{c}_i \neq \mathbf{0}$ $(i = 1, \cdots, n)$.

The main purpose of this study is to estimate the structural relationships among the hidden random variables; the trend components and seaonal components in particular when we have stationary and non-stationary errors-in-variables models. Let $\boldsymbol{\beta}$ be a $p \times 1$ vector and we want to estimate

$$\boldsymbol{\beta}^{'}\mathbf{x}_i = o_p(1) \tag{1.2}$$

when we have the observations of $p \times 1$ vectors $\mathbf{y}_i$ $(i = 1, \cdots, n)$. More generally, let $\mathbf{B}$ be a $q \times p$ $(q \leq p)$ matrix and we want to estimate

$$\mathbf{B}\mathbf{x}_i = o_p(1); \tag{1.3}$$

when we have the observations of $p \times 1$ vectors $\mathbf{y}_i$ $(i = 1, \cdots, n)$. Similarly, some structural relations among seasonal components can be written as

$$\mathbf{B}_s\mathbf{s}_i = o_p(1) \ , \tag{1.4}$$

and they imply that the observed multivariate time series have common seasonality.

2. Simple Cases

Consider some examples when $p = 2$ and $\mathbf{c}_i = \mathbf{0}$ (i.e. no-cycle components).
Example 1 : Assume that the random variables $x_{1i} = \nu_i = \beta_2\mu_i$ and $x_{2i} = \mu_i$ satisfy $\mu_i = \mu_{i-1} + w_i^{(x)}$ $(i = 1, \cdots, n)$ and $w_i^{(x)}$ are i.i.d. random variables with $\mathcal{E}(w_i^{(x)}) = 0$ and $\mathcal{E}(w_i^{(x)2}) = \sigma_x^2$. Then we can write

$$\mathbf{y}_i = \begin{pmatrix} \beta_2 \\ 1 \end{pmatrix} \mu_i + \mathbf{v}_i \ . \tag{2.1}$$

Since μ_i follows the random walk model,

$$\frac{1}{n^2}\sum_{i=1}^{n} \mu_i^2 \xrightarrow{p} \sigma_x^2 \int_0^1 B_s^2 ds \ , \tag{2.2}$$

where B_s is the standard Brownian Motion on $[0, 1]$.
If we multiply the vector $\boldsymbol{\beta}^{'} = (1, -\beta_2)$ to (2.1) from the left, we have the relation

$$\boldsymbol{\beta}^{'}\mathbf{y}_i = u_i \ (= \boldsymbol{\beta}^{'}\mathbf{v}_i) \ , \tag{2.3}$$

which is a structural equation.

Example 2 : We take the case when $\mathbf{x}_i = \boldsymbol{\mu}_i$, and $\boldsymbol{\mu}_i = \boldsymbol{\mu}_{i-1} + \mathbf{w}_i^{(x)}$, which has been often called **spurious regression**. It can be written as

$$\mathbf{y}_i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \boldsymbol{\mu}_i + \mathbf{v}_i \tag{2.4}$$

and the dimension of random walk is 2 and $\boldsymbol{\beta}^{'}\mathbf{y}_i = \boldsymbol{\beta}^{'}\boldsymbol{\mu}_i + u_i$, $u_i = \boldsymbol{\beta}_x^{'}\mathbf{v}_i$ for any $\boldsymbol{\beta} \neq \mathbf{0}$ (the term of $\boldsymbol{\beta}^{'}\boldsymbol{\mu}_i$ cannot be disappeared).

Example 3 : Assume that the random variables $s_{1i} = \nu_i^{(s)} = \beta_2^{(s)}\mu_i^{(s)}$ and $s_{2i} = \mu_i^{(s)}$

satisfy $\mu_i^{(s)} = \mu_{i-s}^{(s)} + w_i^{(s)}$ $(s \geq 1\ ;\ i = 1, \cdots, n)$ and $w_i^{(s)}$ are i.i.d. random variables with $\mathcal{E}(w_i^{(s)}) = 0$ and $\mathcal{E}(w_i^{(s)2}) = \sigma_s^2$. Then we can write

$$\mathbf{y}_i = \begin{pmatrix} \beta_2^{(s)} \\ 1 \end{pmatrix} \mu_i + \mathbf{v}_i \ . \tag{2.5}$$

If we multiply the vector $\boldsymbol{\beta}_s^{'} = (1, -\beta_2^{(s)})$ to (2.5) from the left, we have the relation

$$\boldsymbol{\beta}_s^{'}\mathbf{y}_i = u_i \ (= \boldsymbol{\beta}_s^{'}\mathbf{v}_i) \tag{2.6}$$

and $\mathbf{y}_i$ has the common seasonal components.

3. The Case without Seasonality

Let $p \geq 2$ and $\mathbf{s}_i = \mathbf{c}_i = \mathbf{0}$, we have the representation

$$\mathbf{y}_i = \mathbf{x}_i + \mathbf{v}_i = \mathbf{\Pi}\boldsymbol{\mu}_i + \mathbf{v}_i \ , \tag{3.1}$$

where $\mathcal{E}(\mathbf{w}_i^{(x)}) = \mathbf{0}$, $\mathcal{E}(\mathbf{w}_i^{(x)}\mathbf{w}_i^{(x)'}) = \mathbf{\Sigma}_x$ and $\mathbf{w}_i^{(x)} = \Delta\mathbf{x}_i$. We assume that the rank of $p \times q$ matrix $\mathbf{\Pi}$ is q $(\leq p)$, $\boldsymbol{\mu}_i$ are $q \times 1$ vectors, and there exists a $q \times p$ matrix $\mathbf{B}$ such that $\mathbf{B}\mathbf{y}_i = u_i$ $(= \mathbf{B}\mathbf{v}_i)$, which are the set of q structural equations.

We consider the situation when $\Delta\mathbf{x}_i$ and $\mathbf{v}_i$ $(i = 1, \cdots, n)$ are independent and they are independently, identically and normally distributed as $N_p(\mathbf{0}, \mathbf{\Sigma}_x)$ and $N_p(\mathbf{0}, \mathbf{\Sigma}_v)$, respectively. We use an $n \times p$ matrix $\mathbf{Y}_n = (\mathbf{y}_i^{'})$ and consider the distribution of $np \times 1$ random vector $(\mathbf{y}_1^{'}, \cdots, \mathbf{y}_n^{'})^{'}$. Given the initial condition $\mathbf{y}_0$, we have

$$\mathbf{Y}_n \sim N_{n\times p}\left(\mathbf{1}_n \cdot \mathbf{y}_0^{'}, \mathbf{I}_n \otimes \mathbf{\Sigma}_v + \mathbf{C}_n\mathbf{C}_n^{'} \otimes \mathbf{\Sigma}_x\right) \ , \tag{3.2}$$

where $\mathbf{1}_n^{'} = (1, \cdots, 1)$ and

$$\mathbf{C}_n = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ 1 & \cdots & 1 & 1 & 0 \\ 1 & \cdots & 1 & 1 & 1 \end{pmatrix}_{n\times n} \ . \tag{3.3}$$

Then given the initial condition $\mathbf{y}_0$ the maximum likelihood (ML) estimator can be defined as the solution of maximizing the log-likelihood function except a constant as

$$\begin{aligned} L_n^* &= \log|\mathbf{I}_n \otimes \mathbf{\Sigma}_v + \mathbf{C}_n\mathbf{C}_n^{'} \otimes \mathbf{\Sigma}_x|^{-1/2} - \frac{1}{2}[vec(\mathbf{Y}_n - \bar{\mathbf{Y}}_0)^{'}]^{'}[\mathbf{I}_n \otimes \mathbf{\Sigma}_v + \mathbf{C}_n\mathbf{C}_n^{'} \otimes \mathbf{\Sigma}_x]^{-1} \\ &\qquad \times[vec(\mathbf{Y}_n - \bar{\mathbf{Y}}_0)^{'}] \end{aligned}$$

and

$$\bar{\mathbf{Y}}_0 = \mathbf{1}_n \cdot \mathbf{y}_0^{'} . \tag{3.4}$$

We transform $\mathbf{Y}_n$ to $\mathbf{Z}_n$ $(= (\mathbf{z}_k^{'}))$ by

$$\mathbf{Z}_n = \mathbf{P}_n \mathbf{C}_n^{-1} \left(\mathbf{Y}_n - \bar{\mathbf{Y}}_0 \right) \tag{3.5}$$

where

$$\mathbf{C}_n^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}_{n \times n} , \tag{3.6}$$

$$\mathbf{P}_n = (p_{jk}^{(n)}) \ , \ p_{jk}^{(n)} = \sqrt{\frac{2}{n+\frac{1}{2}}} \cos \left[\frac{2\pi}{2n+1} (k - \frac{1}{2})(j - \frac{1}{2}) \right] . \tag{3.7}$$

By using the spectral decomposition $\mathbf{C}_n^{-1} \mathbf{C}_n^{'-1} = \mathbf{P}_n \mathbf{D}_n \mathbf{P}_n^{'}$ and $\mathbf{D}_n$ is a diagonal matrix with the k-th element

$$d_k = 2[1 - \cos(\pi(\frac{2k-1}{2n+1}))] \ (k = 1, \cdots, n) .$$

Then the log-likelihood function is proportional to

$$L_n = \sum_{k=1}^{n} \log |a_{kn} \mathbf{\Sigma}_v + \mathbf{\Sigma}_x|^{-1/2} - \frac{1}{2} \sum_{k=1}^{n} \mathbf{z}_k^{'} [a_{kn} \mathbf{\Sigma}_v + \mathbf{\Sigma}_x]^{-1} \mathbf{z}_k , \tag{3.8}$$

where

$$a_{kn} \ (= d_k) \ = 4 \sin^2 \left[\frac{\pi}{2} \left(\frac{2k-1}{2n+1} \right) \right] \ (k = 1, \cdots, n) . \tag{3.9}$$

Since we are dealing with the errors-in-variables model, there is an issue if we can identify the structural equation of our interest. When $\mathbf{x}_i$ are i.i.d. random variables, for instance, the coefficient parameters are not identified without some further restrictions. In the classical case we need to impose some conditions on the covariance matrix (such as the homoscedasticity and zero covariance) when we have the functional relationship model in Example 1 except x_i $(= \mu_i \ ; i = 1, \cdots, n)$ with

$$\frac{1}{n} \sum_{i=1}^{n} \mu_i^2 \xrightarrow{p} \sigma_x^2 . \tag{3.10}$$

(See Fuller (1987) for instance.) Here we say that the parameter vector $\boldsymbol{\theta}$ $(= (\theta_j))$ is identified if $\boldsymbol{\theta} \neq \boldsymbol{\theta}^{'}$ implies that $L_n(\theta) \neq L_n(\theta^{'})$.

We illustrate our arguments on the likelihood function when $q = 1$. We take $\boldsymbol{\theta}$ $(= \boldsymbol{b})$ and apply the matrix formulae that for a positive definite $\mathbf{A}$ we have

$$|\mathbf{A} + \mathbf{b}\mathbf{b}^{'}| = |\mathbf{A}|[1 + \mathbf{b}^{'} \mathbf{A}^{-1} \mathbf{b}]$$

and

$$[\mathbf{A}+\mathbf{b}\mathbf{b}^{'}]^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{b}[1+\mathbf{b}^{'}\mathbf{A}^{-1}\mathbf{b}]^{-1}\mathbf{b}^{'}\mathbf{A}^{-1}$$

for $\mathbf{A}=a_{kn}\mathbf{\Sigma}_v$ $(k=1,\cdots,n)$, $\mathbf{b}=\sigma_\mu\mathbf{\Pi}$ and $\mathbf{b}_*=\mathbf{\Sigma}_v^{-1}\sigma_\mu\mathbf{\Pi}$.
Then L_n is proportional to (-1/2) times

$$(3.11)\quad L_{1n}=\sum_{k=1}^{n}\left[\log|\mathbf{\Sigma}_v|+\log(a_{kn}+\mathbf{b}^{'}\mathbf{\Sigma}_v^{-1}\mathbf{b})+a_{kn}^{-1}\mathbf{z}_k^{'}\mathbf{\Sigma}_v^{-1}\mathbf{z}_k-\frac{a_{kn}^{-1}(\mathbf{z}_k^{'}\mathbf{\Sigma}_v^{-1}\mathbf{b})^2}{a_{kn}+\mathbf{b}^{'}\mathbf{\Sigma}_v^{-1}\mathbf{b}}\right] .$$

We notice that L_{1n} is a concave function of $\mathbf{\Sigma}_v^{-1}$ and

$$\sum_{k=1}^{n}\frac{(\mathbf{z}_k^{'}\mathbf{\Sigma}_v^{-1}\mathbf{b})^2}{a_{kn}^2+a_{kn}\mathbf{b}^{'}\mathbf{\Sigma}_v^{-1}\mathbf{b}}=\mathbf{b}_*^{'}\sum_{k=1}^{n}[\frac{1}{a_{kn}(a_{kn}+c)}\mathbf{z}_k\mathbf{z}_k^{'}]\mathbf{b}_* ,$$

where we set $\mathbf{b}^{'}\mathbf{\Sigma}_v^{-1}\mathbf{b}=c$. Then given $\mathbf{b}_*^{'}\mathbf{\Sigma}_v\mathbf{b}_*=c$, it is a quadratic form and its minimum is the smallest characteristic root. Thus we have the next results.

Proposition 1 : Under the above assumptions, the structural parameter $\boldsymbol{\theta}$ (i.e. $\mathbf{B}$) is identified (up to a normalization).

Given the initial condition $\mathbf{y}_0$, the unconditional maximum likelihood (ML) estimator can be defined as the solution of maximizing the log-likelihood function.

Proposition 2 : Under the above assumptions, there exists a unique ML estimator for $\mathbf{B}$.

Next, given $\mathbf{b}^{'}\mathbf{\Sigma}_v^{-1}\mathbf{b}=\mathbf{b}_*^{'}\mathbf{\Sigma}_v\mathbf{b}_*=c$ (*a constant*), we approximate L_{1n} of the function of $\mathbf{\Sigma}_v$ and $\mathbf{b}_*$. If we use

$$(3.12)\qquad \mathbf{\Sigma}_v=\frac{1}{n}\sum_{k=n}^{n}a_{kn}^{-1}\mathbf{z}_k\mathbf{z}_k^{'} ,$$

and $1/[a_{kn}(a_{kn}+c)]=(1/c)[1/a_{kn}-1/(a_{kn}+c)]$, then given $\mathbf{b}_*^{'}\mathbf{\Sigma}_v\mathbf{b}_*=c$, the problem becomes to find a minimum of

$$(3.13)\qquad \mathbf{b}_*^{'}\sum_{k=1}^{n}[\frac{1}{a_{kn}+c}\mathbf{z}_k\mathbf{z}_k^{'}]\mathbf{b}_* .$$

Then it may be natural to use the minimization of

$$(3.14)\qquad \frac{1}{c}\mathbf{b}_*^{'}\mathbf{A}_m\mathbf{b}_*\ \ (\text{s.t. }\mathbf{b}_*^{'}\mathbf{\Sigma}_v\mathbf{b}_*=c,\mathbf{A}_m=\sum_{k=1}^{m_n}\mathbf{z}_k\mathbf{z}_k^{'}\ ),$$

where we use the fact that $c<a_{kn}+c<4+c$ and $a_{kn}\sim O(1/n)$ when $k_n=o(1)$. It is straightforward to extend the above likelihood analysis to the cases for more general q $(q\leq p)$.

4. Macro-SIML Estimation

The exact ML estimator of unknown parameters is a rather complicated function of all observations and it may depend crucially on the underlying assumptions including the Gaussianity. Then we need a simple robust procedure such that the assumptions of Gaussianity and the specifications of each components are not crucial for the estimation results.

Let denote $a_{k_n,n}$ and then we can evaluate that $a_{k_n,n} \to 0$ as $n \to \infty$ when $k_n = O(n^{\alpha})$ $(0 < \alpha < 1)$ since $\sin x \sim x$ as $x \to 0$. When k_n is small, we expect that $a_{k_n,n}$ is small. Then the separating information maximum likelihood (SIML) estimator of $\hat{\mathbf{\Sigma}}_x$ is defined by

$$\hat{\mathbf{\Sigma}}_{x,SIML} = \frac{1}{m_n}\sum_{k=1}^{m_n} \mathbf{z}_k \mathbf{z}_k^{'} \ . \tag{4.1}$$

(We need to use a consistent estimator for $\mathbf{\Sigma}_v$.) For $\hat{\mathbf{\Sigma}}_x$, the number of terms m_n should be dependent on n. Then we need the order requirement that $m_n = O(n^{\alpha})$ $(0 < \alpha < 1)$.

It may be natural to use the sample quantities $\hat{\mathbf{\Sigma}}_x$ $(= ((1/m_n)\sum_{k=1}^{m_n} z_{ik}z_{jk}))$ order to make statistical inference on $\mathbf{\Sigma}_x$. The estimation of the Pearson-correlation coefficent is a typical case, which is given by

$$\hat{\rho}_{ij} = \frac{\sum_{k=1}^{m_n} z_{ik}z_{jk}}{\sqrt{\sum_{k=1}^{m_n} z_{ik}^2}\sqrt{\sum_{k=1}^{m_n} z_{jk}^2}} \ . \tag{4.2}$$

Furthermore, we consider the estimation of the structural relationships in the non-stationary time series process satisfying (3.1). Hewe we notice that the present statistical problem could be regarded as the estimation of structural relationships with the covariance matrix $\mathbf{\Sigma}_x(\boldsymbol{\theta})$ with $\boldsymbol{\theta}$ being the vector of parameters. In the standard statistical multivariate analysis Anderson (1984) discussed the statistical models of estimating structural relationships among a set of variables and we have n independent observations on the underlying variables.

We return to the case when $q = 1$. By using the arguments on the likelihood function, it may be natural to consider the characteristic equation

$$\left[\mathbf{\Sigma}_v^{-1}\hat{\mathbf{\Sigma}}_x\mathbf{\Sigma}_v^{-1} - \lambda\mathbf{\Sigma}_v^{-1}\right]\mathbf{b} = \mathbf{0} \ , \tag{4.3}$$

where λ is the (scalar) characteristic root. By using the relation $\mathbf{b}_* = \mathbf{\Sigma}_v^{-1}\mathbf{b}$ and we can represent

$$\left[\hat{\mathbf{\Sigma}}_x - \lambda\mathbf{\Sigma}_v\right]\mathbf{b}_* = \mathbf{0} \ . \tag{4.4}$$

If we take the smallest eigen-value and we take $\hat{\mathbf{\Sigma}}_{v,SIML}$, we have the $\mathbf{b}_{*,SIML}$.

Also under a set of regularity conditions we have that the smallest eigen-value

$$\lambda_1 \longrightarrow 0 \text{ (in probability)} \tag{4.5}$$

as $n \to \infty$. Then we may use the SILS (Separating Information Least Squares) method by solving

$$\hat{\mathbf{\Sigma}}_x \mathbf{b}_{*,SILS} = \mathbf{0} \ . \tag{4.6}$$

When $p = 2, q = 1$, $\boldsymbol{\beta} = (1, -\beta_2)'$ and $\mathbf{\Pi} = (\beta_2, 1)'$ ($\mathbf{\Sigma}_v = \sigma_v^2 \mathbf{I}_2$) in Example 1, then the SILS estimation becomes

$$\hat{\beta}_2 = \frac{\sum_{k=1}^{m_n} z_{1k} z_{2k}}{\sum_{k=1}^{m_n} z_{2k}^2} \ , \tag{4.7}$$

which is the regression coefficient of the first transformed variable on the second transformed variable in $\mathbf{z}_k$ $(= (z_{1k}, z_{2k})')$ $(k = 1, \cdots, m_n)$.

Some Remarks :

By using the SIML procedure, it is also straightforward to investigate the several structural relationships among trend variables at the same time. The SIML estimation can be defined by the smaller q $(\leq p)$ roots and the corresponding q $(\leq p)$ vectors of the characteristic equation. It may correspond to the standard situation in the statistical multivariate analysis except the fact that the classical multivariate analysis was based on the case when the observations are realizations of i.i.d. random variables without seasonality as well as non-stationarity in time series data sets.

We have done several simulations which are given in Appendix. The data length is 80, the number of simulations is 3000, $\alpha = 0.6$, and $m_n = n^{\alpha}$ in each case. In fugures *cor* $= 0.9$ means the true correlation coefficient $(= .9)$ among the trend components and *cor* is the SIML estimate. (vol1 is the correlation estimate based on the first differenced data and vol4 is the correlation estimate based on the seasonal differenced data with $s = 4$.)

When we have the basic model with the trend and noise components without the seasonal and cycle components, the optimal choice of $m_n = n^{\alpha}$ would be $\alpha = 0.8$, but it seems that the choice of $\alpha = 0.6$ would appropriate for the robustness of the results when we have seasonality as well as non-stationary trends.

5. Further Thoughts

Let $\mathbf{s}_i' = (s_{1i}, \cdots, s_{pi})$ be the vector of the seasonal components, which are independent of $\mathbf{v}_i$. Then we have

$$\mathbf{y}_i = \mathbf{s}_i + \mathbf{v}_i \tag{5.1}$$

where $\mathbf{s}_i$ are a sequence of seasonal components with $\mathcal{E}(\mathbf{w}_i^{(s)}) = \mathbf{0}$ and $\mathcal{E}(\mathbf{w}_i^{(s)} \mathbf{w}_i^{(s)'}) = \mathbf{\Sigma}_s$ with $\Delta_s \mathbf{s}_i = \mathbf{s}_i - \mathbf{s}_{i-s} = \mathbf{w}_i^{(s)}$ $(s \geq 2)$. We assume that $\mathbf{\Sigma}_s$ is positive definite and finite. When we transform the observed data by using the seasonal difference operator, we have

$$(1-\mathcal{L}^s)\mathbf{y}_i = \mathbf{w}_i^{(s)} + (1-\mathcal{L}^s)\mathbf{v}_i \ . \tag{5.2}$$

For the seasonal transformation let $n = ms$ $(s \geq 2)$ and transform $\mathbf{Y}_n$ to $\mathbf{Z}_n^{(s)}$ $(= (\mathbf{z}_k^{(s)'}))$ by

$$\mathbf{Z}_n^{(s)} = \mathbf{P}_n^{(s)}\mathbf{C}_n^{(s)-1}\left(\mathbf{Y}_n - \bar{\mathbf{Y}}_0\right) \tag{5.3}$$

where $\mathbf{C}_n^{(s)-1} = \mathbf{C}_m^{-1} \otimes \mathbf{I}_s$ and

$$\mathbf{C}_m^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}_{m\times m} , \tag{5.4}$$

$$\mathbf{P}_n^{(s)} = \mathbf{P}_m \otimes \mathbf{I}_s \ . \tag{5.5}$$

By using the spectral decomposition $\mathbf{C}_n^{(s)-1}\mathbf{C}_n^{(s)'-1} = (\mathbf{P}_m \otimes \mathbf{I}_s)(\mathbf{D}_m^{(s)} \otimes \mathbf{I}_s)(\mathbf{P}_m' \otimes \mathbf{I}_s)$ and $\mathbf{D}_m^{(s)}$ is a diagonal matrix with

$$a_{km} = 2[1-\cos(\pi(\frac{2k-1}{2m+1}))] = 4\sin^2(\pi/2)[(2k-1)/(2m+1)] \ (k=1,\cdots,m)$$

with the s multiplicities.
Then we can develop the likelihood analysis as the case when $s=1$.

We consider the case when we have

$$\mathbf{y}_i = \mathbf{x}_i + \mathbf{s}_i + \mathbf{v}_i \tag{5.6}$$

where $\mathbf{x}_i$ are a sequence of trend components and $\mathbf{s}_i$ are a sequence of seasonal components. When we transform the observed data by using the seasonal difference operator, we have

$$(1-\mathcal{L}^s)\mathbf{y}_i = (1+\mathcal{L}+\cdots+\mathcal{L}^{s-1})\Delta\mathbf{x}_i + (1-\mathcal{L}^s)\mathbf{s}_i + (1-\mathcal{L}^s)\mathbf{v}_i \ . \tag{5.7}$$

Then there can be alternative possibilities of transformations of $\mathbf{Y}_n$, but we may use $\mathbf{Z}_n^{(s)*}$ $(= (\mathbf{z}_k^{(s)*'}))$ by

$$\mathbf{Z}_n^{(s)*} = \mathbf{P}_n\mathbf{C}_n^{(s)-1}\left(\mathbf{Y}_n - \bar{\mathbf{Y}}_0\right) \ . \tag{5.8}$$

It seems that the likelihood analysis can be extended by using the orthogonality relations with components of $p_{jk}^{(n)}$. Let

$$\mathbf{\Lambda}_n = (\lambda_{jk}) = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ \vdots & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ 0 & \cdots & 1 & 1 & 0 \\ 0 & \cdots & 1 & \cdots & 1 \end{pmatrix}_{n\times n} \tag{5.9}$$

and

(5.10) $$\mathbf{P}_n^* = (p_{jk}^*) = \mathbf{P}_n \mathbf{\Lambda}_n \ .$$

Then

(5.11) $$\sum_{j=1}^{n} p_{kj}^* p_{k',j}^* \sim \delta(k,k') \frac{\sin^2 \frac{\pi}{2}\frac{2k-1}{2n+1}s}{\sin^2 \frac{\pi}{2}\frac{2k-1}{2n+1}} \sim s^2 \delta(k,k')$$

when $k/n \to 0$.

We consider the general case when we have

(5.12) $$\mathbf{y}_i = \mathbf{x}_i + \mathbf{s}_i + \mathbf{c}_i + \mathbf{v}_i$$

where $\mathbf{x}_i$ are a sequence of trend components, $\mathbf{s}_i$ are a sequence of seasonal components and $\mathbf{c}_i$ are a sequence of cycle components. When we transform the observed data by using the seasonal difference operator, we have

(5.13) $$(1-\mathcal{L}^s)\mathbf{y}_i = (1+\mathcal{L}+\cdots+\mathcal{L}^{s-1})\Delta\mathbf{x}_i + \mathbf{w}_i^{(s)} + (1-\mathcal{L}^s)\mathbf{c}_i + (1-\mathcal{L}^s)\mathbf{v}_i$$

with $(1-\mathcal{L}^s)\mathbf{s}_i = \mathbf{w}_i^{(s)}$.
Then again there are alternative possibilities of transformations of $\mathbf{Y}_n$, but we may use $\mathbf{Z}_n^{(s)*}$ $(= (\mathbf{z}_k^{(s)*'}))$ by

(5.14) $$\mathbf{Z}_n^{(s)*} = \mathbf{P}_n \mathbf{C}_n^{(s)-1} \left(\mathbf{Y}_n - \bar{\mathbf{Y}}_0 \right) \ .$$

It seems that the likelihood analysis can be extended in the second case by using the orthogonality relations with components of $p_{jk}^{(n)}$. Let

(5.15) $$\mathbf{P}_n^{**} = (p_{jk}^{**}) = \mathbf{P}_n(\mathbf{C}_m^{-1} \otimes \mathbf{I}_s) \ .$$

Then

(5.16) $$\sum_{j=1}^{n-s} p_{kj}^{**} p_{k',j}^{**} \sim 4\sin^2[\frac{\pi}{2}\frac{2k-1}{2n+1}s]$$

when $k=k'$. When $k \neq k'$,

(5.17) $$\sum_{j=1}^{n-s} p_{kj}^{**} p_{k',j}^{**} \sim O(\frac{1}{n}) \ .$$

Also for any finite l we have

(5.18) $$\sum_{j=1}^{n-s-l} p_{kj}^{**} p_{k',j+l}^{**} \sim O(\frac{1}{n})$$

we find that the covariances of the transformed cycle compotents $(\mathbf{c}_k^*) = \mathbf{P}_n^{**}(\mathbf{c}_j)$ are

(5.19) $$Cov(\mathbf{c}_j^*, \mathbf{c}_{j'}^*) = \rho^{|j-j'|} \times O(\frac{1}{n})$$

with $|\rho| < 1$.

References

[1] Anderson, T.W. (1984), "Estimating Linear Statistical Relationships," *Annals of Statistics*, 12, 1-45.

[2] Anderson, T.W. (2003), *An Introduction to Statistical Multivariate Analysis*, 3rd Edition, John-Wiley.

[3] Fuller, W. (1987), *Measurement Error Models*, John-Wiley.

[4] Kitagawa, G. (2010), *Introduction to Time Series Modeling*, CRA Press.

[5] Kunitomo, N. and S. Sato (2008, 2013), "Separating Information Maximum Likelihood Estimation of Realized Volatility and Covariance with Micro-Market Noise," Discussion Paper CIRJE-F-581, Graduate School of Economics, University of Tokyo, (http://www.e.u-tokyo.ac.jp/cirje/research/dp/2008), North American Journal of Economics and Finance, Elsevier, 26, 282-309.

[6] Kunitomo, N. and S. Sato (2011), "The SIML Estimation of Realized Volatility of Nikkei-225 Futures and Hedging Coefficient with Micro-Market Noise," *Mathematics and Computers in Simulation*, Vol.81, 1272-1289.

[7] 国友直人 (2011),「構造方程式モデルと計量経済学」, 朝倉書店。

[8] 国友直人 (2014), 「計測誤差と統計学」, 日本統計学会誌, Vol. 43-2, 157-183.

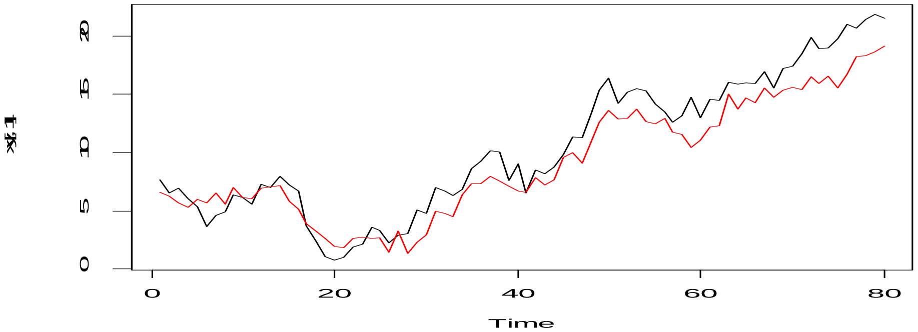
[9] 国友直人 (2015), 「(応用を目指す) 数理統計学」, 朝倉書店 (近刊)。

Appendix: Some Figures

Quarterly Data	alpha=0.6
n=80	nsim=3000

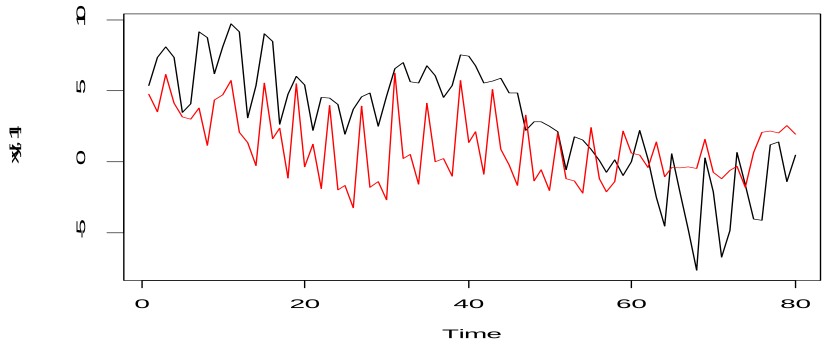
With Noise, no Seaonality

cor=0.9	cor	vol4	vol1
mean	0.852227	0.732646	0.491137
SD	0.087745	0.075951	0.094544
cor=0.0	cor	vol4	vol1
mean	0.007154	0.003388	0.001042
SD	0.277539	0.16811	0.119081



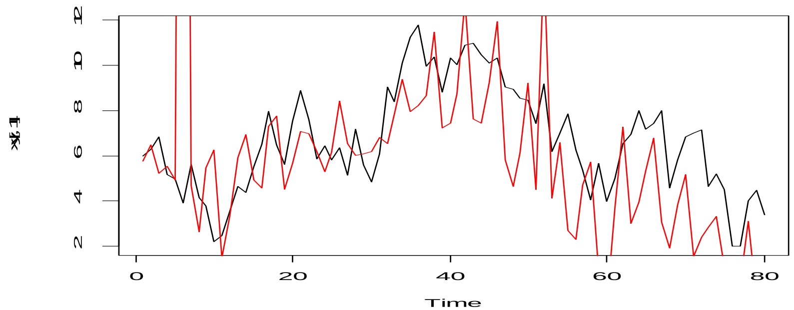
With noise and seasonality

cor=0.9	cor	vol4	vol1
mean	0.805092	0.662977	0.133043
SD	0.118186	0.088369	0.295366
cor=0.0	cor	vol4	vol1
mean	−0.00689	2.59E−03	0.004784
SD	0.278001	1.62E−01	0.287285



With noise, seasonality and outlier

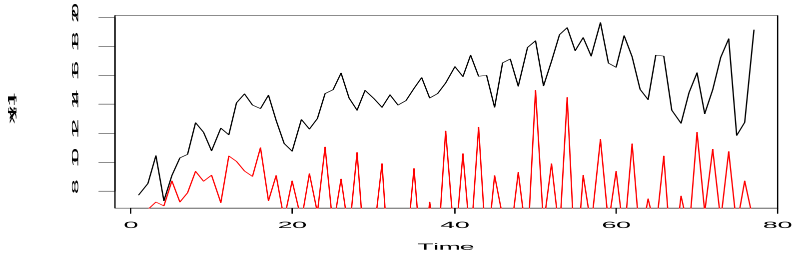
cor=0.9	cor	vol4	vol1
mean	0.584384	0.308292	0.063187
SD	0.209764	0.13995	0.217537
cor=0.0	cor	vol4	vol1
mean	0.005651	0.004289	−0.00455
SD	0.275338	0.128497	0.210767



With noise and changing seasonality

cor=0.9	cor	vol4	vol1
mean	0.672386	0.344135	0.033535
SD	0.196482	0.184532	0.190742

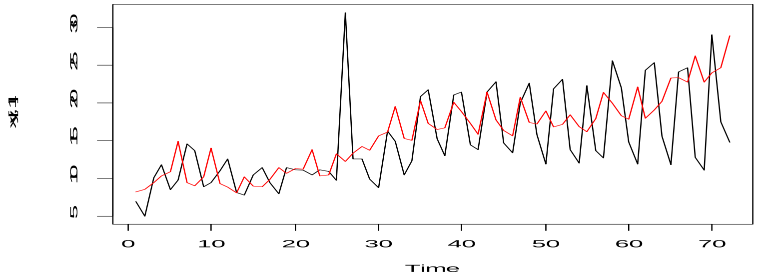
cor=0.0	cor	vol4	vol1
mean	0.002128	0.001937	0.002259
SD	0.283569	0.148525	0.183541



With noise, seasonality,outlier and changing seasonality

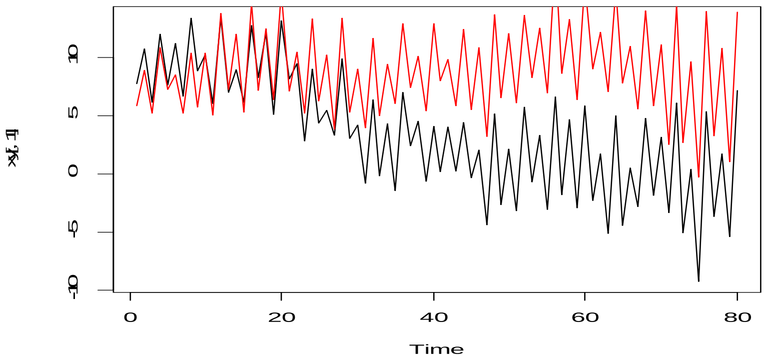
cor=0.9	cor	vol4	vol1
mean	0.491676	0.184133	0.019022
SD	0.244269	0.152994	0.160183

cor=0.0	cor	vol4	vol1
mean	−0.00467	0.000966	0.002659
SD	0.283498	0.133407	0.162206



With noise and seasonality

cor=0.0	cor	vol4	vol1
mean	0.044796	0.083241	0.673487
SD	0.282269	0.165334	0.237343



地域統計の季節調整問題

川崎能典（統計数理研究所）
国友直人（東京大学大学院経済学研究科）

科学研究費(基盤研究(A))研究集会
「経済統計・政府統計の基礎と応用」
2015年1月30日(金)
東京大学大学院経済学研究科・小島ホール

問題の提起

- 都道府県別時系列の季節調整
- 都道府県別時系列を集計して定義される地域別時系列(cf. 東北ブロック)の季節調整
- 両者は独立した統計処理
- 都道府県別季節調整済系列の合算と、地域別時系列に対する季調結果の不一致
- 両者の季調済系列間に何らかの整合性を持たせる必要があるのでは?

自然な制約～2変量を例に

- 第1県の時系列と季調 $y_{1t} = n_{1t} + s_{1t}$
- 第2県の時系列と季調 $y_{2t} = n_{2t} + s_{2t}$
  - n_{*t}はnon-seasonal partを、s_{*t}はseasonal partを表す。以下のN_t, S_tも同じ。
- 合算系列(より上位の地域ブロック統計に相当) $Y_t = y_{1t} + y_{2t}$ の季調 $Y_t = N_t + S_t$
- 自然な制約(期待) $N_t \approx n_{1t} + n_{2t}$, これを裏(除去したいもの)から見ると$S_t \approx s_{1t} + s_{2t}$

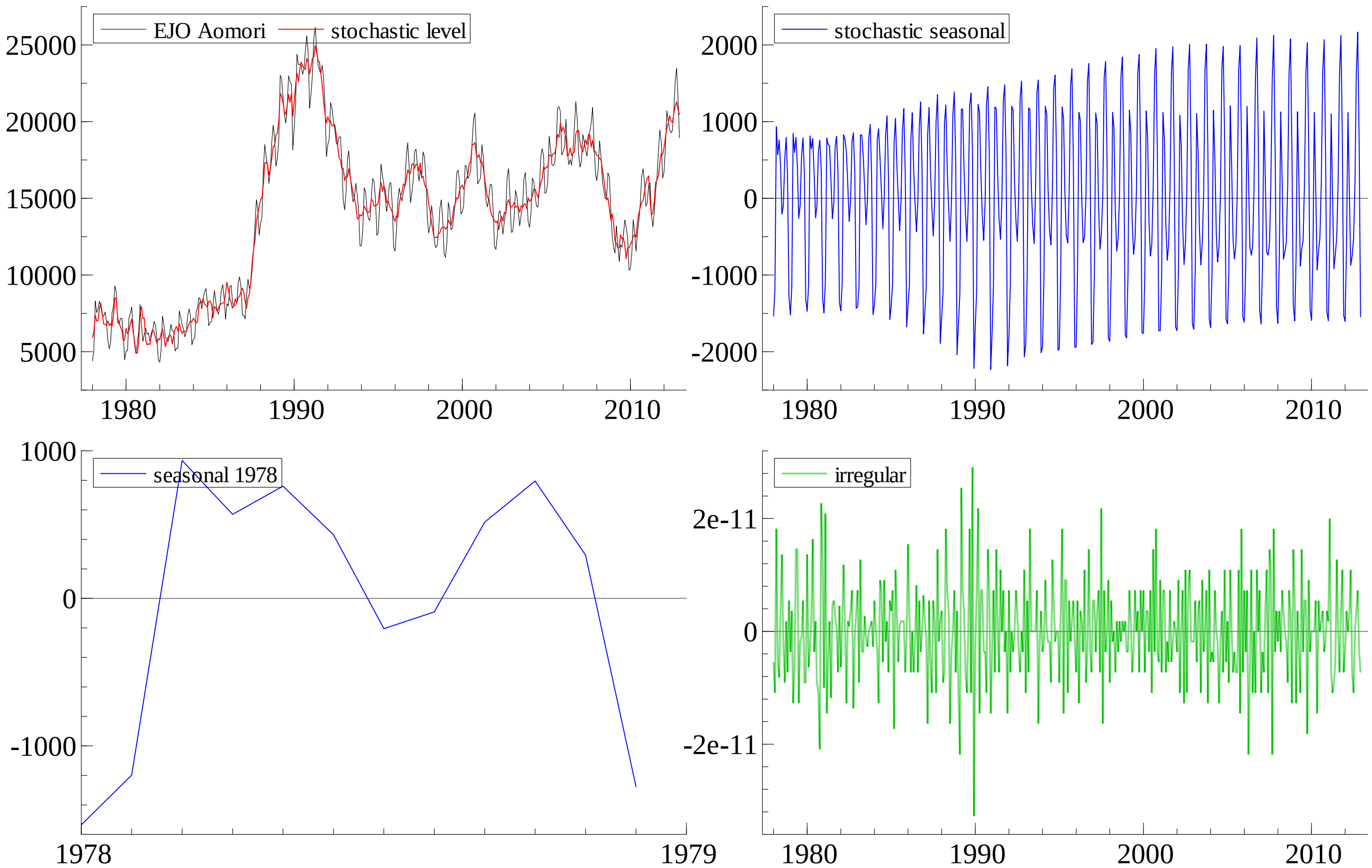
自然な制約(続)

- 例えば$N_t = TC_t + I_t$と分けて考えた時、$TC_t \approx TC_{1t} + TC_{2t}$を課すモデリングは可能だが、$I_t \approx I_{1t} + I_{2t}$は考えづらい
  - 前者は状態変数だが、後者は観測ノイズ
- $S_t \approx s_{1t} + s_{2t}$が成り立つ限りは$N_t \approx n_{1t} + n_{2t}$が期待できるので、集計系列(地域ブロックのデータ)の季節成分だけに、下位系列との整合性を課すモデリングをまず試す。

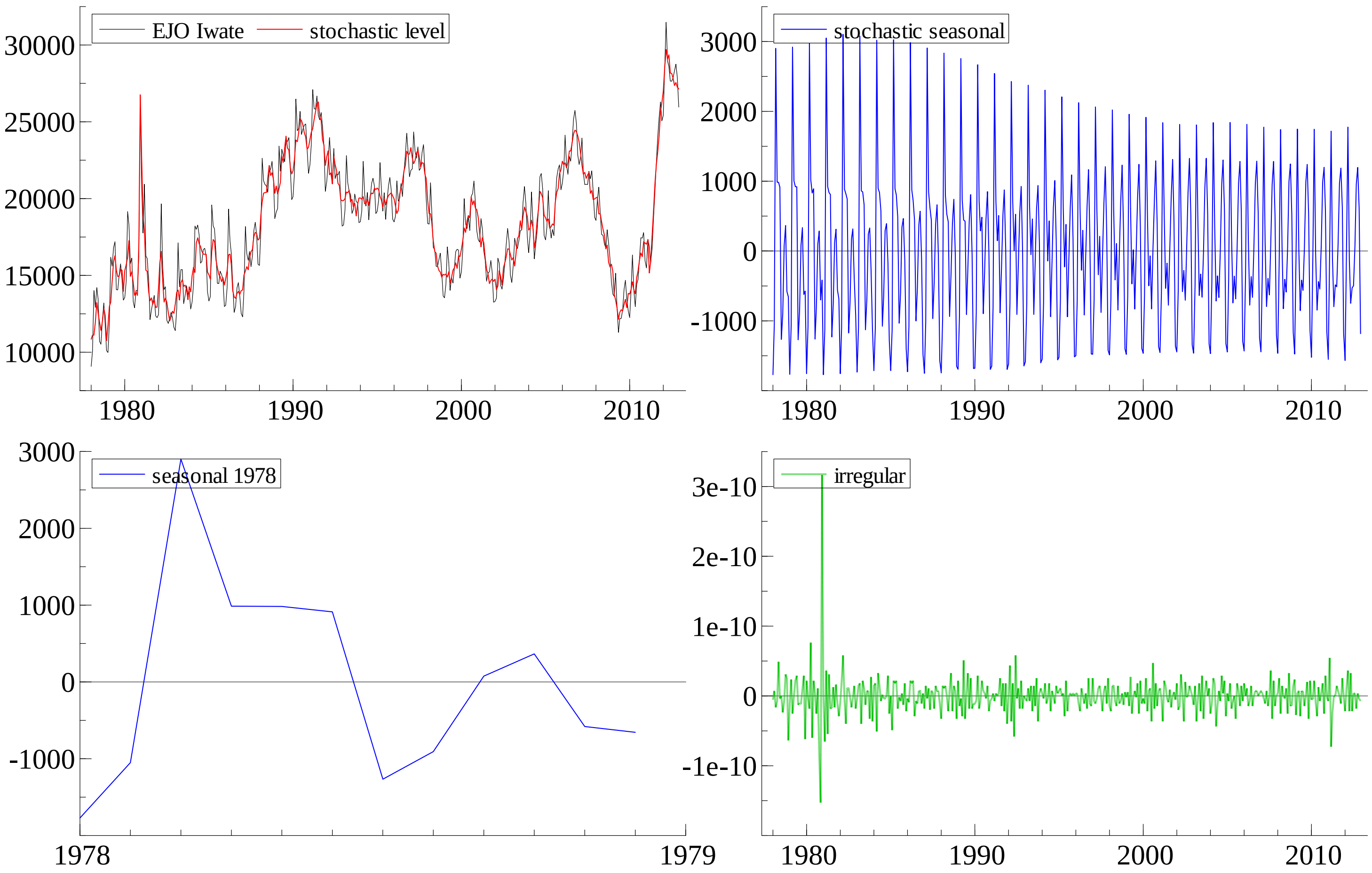
データ

- 職業安定業務統計 有効求人数(原系列)
  - 厚生労働省
- 青森県、岩手県、その合算を考える
- 1978年1月～2012年12月(時系列長420)

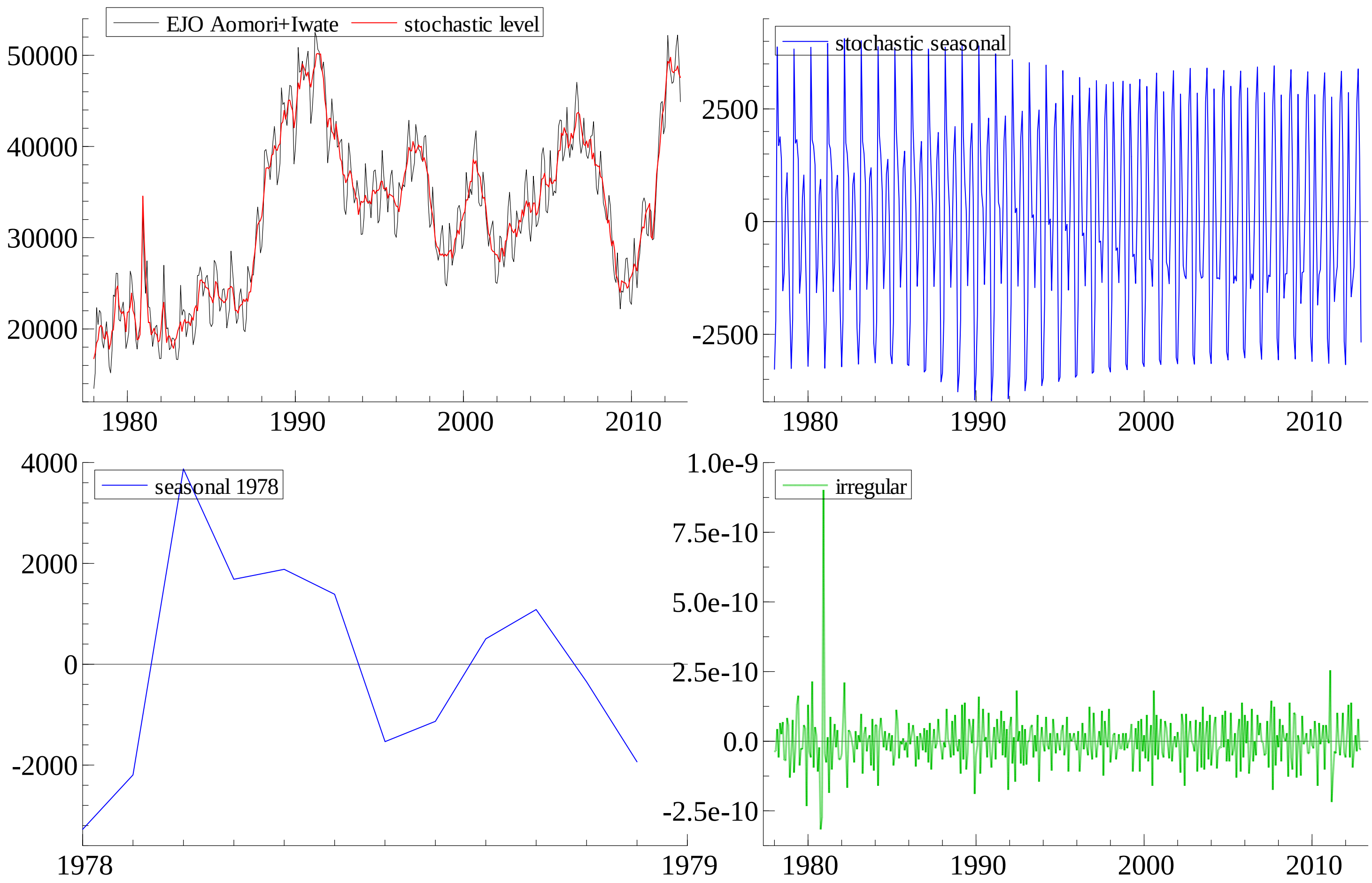
青森県の一変量解析


EJO Aomori
stochastic level
25000
20000
15000
10000
5000
1980
1990
2000
2010
stochastic seasonal
2000
1000
0
-1000
-2000
1980
1990
2000
2010
seasonal 1978
1000
0
-1000
1978
1979
irregular
2e-11
0
-2e-11
1980
1990
2000
2010

岩手県の一変量解析


EJO Iwate
stochastic level
30000
25000
20000
15000
10000
1980
1990
2000
2010
stochastic seasonal
3000
2000
1000
0
-1000
seasonal 1978
1978
1979
irregular
3e-10
2e-10
1e-10
-1e-10

青森+岩手合算系列の一変量解析



多変量状態空間モデル

観測方程式

$$Y_t = y_{1t} + y_{2,t} = \mu_t + s_{1t} + s_{2t} + \epsilon_t$$

$$y_{1t} = \mu_{1,t} + s_{1,t} + \epsilon_{1,t}$$

$$y_{2t} = \mu_{2,t} + s_{2,t} + \epsilon_{2,t}$$

状態方程式

システムノイズ、観測ノイズの分散共分散行列は対角を仮定

$$\mu_t = \mu_{t-1} + \eta_t, \mu_{i,t} = \mu_{i,t-1} + \eta_{i,t}\ (i = 1,2)$$

$$s_{i,t} = -\sum_{j=1}^{s-1} s_{i,t-j} + \omega_{i,t}\ (i = 1,2)$$

(暫定的な)解析結果

- 個別季節調整: 個別の対数尤度の総和
$$\hat{\ell} = -10256.17$$
- 多変量季節調整
$$\hat{\ell} = -10543.26$$
- まだ多変量状態空間モデルのパラメータ推定(最適化)がうまくいっておらず、暫定的な結果だが、多変量モデルは奏功していない…。
  - AIC流のバイアス補正で逆転できるような尤度の差ではない

考察

- 集計系列と、基礎となる個別系列間で整合性ある季節調整を考えたいが、観測ベクトル自体はおのずと完全多重共線状態。
- 個別モデルを束ねて形式的に多変量状態空間モデルとして推定してもパラメータが非現実的に大きくなりうまくいかない。完全多重共線かつ自由度が高すぎ。
- 制約(ファクター型の構造)が足りない?

考察(続):まとめに代えて

- システムノイズの次元落ちに関する丁寧な探索が必要だと、手間がかかりすぎて実用的とは言えない
- 不規則変動部分は(状態変数でないので)ファクター流にモデル化しづらい
- 観測ノイズは対角を仮定しているが推定してみると対角とは言いがたい。直交化は可能だが季節調整の解釈に問題は生じないか?
  - Nishio (2002) FTSM3