【資料3-1】

人流データを活用した国内宿泊者数の試算

総務省統計品質管理推進室

令和 6 年 2 月 21 日

目次

取組について

総務省では、ビッグデータ（BD）を活用した試行的な取組の一環として、携帯端末の位置情報から作成された「メッシュ型人流データ」を用いて、都道府県単位の宿泊者数の推計に関する研究を実施

第 IV 期公的統計基本計画[1]（抄）

第 2　6　統計各分野の取組

(3) 観光に関する統計の精度向上

また、人流データを活用した宿泊動向の足下予測等、ビッグデータの利活用についても研究を進める。

研究概要

- 人流データを用いて宿泊者数の近似値を算出。「宿泊旅行統計調査」の宿泊者数をベンチマークとして、統計の公表に先行する形で日本人宿泊者数を試算

[1]令和５年３月 28 日閣議決定/ Link

ビッグデータの利活用として

「ビッグデータの更なる活用の方向性～政策の質の向上を目指して～」[1]

BD の活用事例を既存の公的統計との関係性から次の 3 つに整理

I 既存の公的統計の中での活用

II 調査実施者が既存の公的統計の結果公表時に併せて行う分析での活用

III 既存の公的統計では捉えることのできなかった新たな指標の作成

3 つを併進 → 「BD の活用の裾野の拡大」、「ノウハウや事例の蓄積」、「各種データの検証」に有効

- BD の速報性を活かし、公的統計の公表に “先行した指標” を得る点で、上記 III に位置

[1] 令和 4 年 6 月 2 日　ビッグデータ等の利活用に関する産官学協議のための連携会議決定/ Link

宿泊者数への着目

数ある指標の中で、なぜ宿泊者数を題材としたか？

直面した課題

- 膨大かつ雑多な BD の情報から、対象となる人たちをどのように特定するか？
- 個々人の子細な動きを捉えた BD のマイクロデータは入手と扱いが極めて困難

導き出したアイデア

- 深夜、早朝は人の動きも少なく、ホテル周辺人口の多くが宿泊者なのでは？
- 一地点にとどまる宿泊者なら、集計された BD からでも研究の実現可能性あり

メッシュ統計の活用

BD と統計の接合にはメッシュ統計の枠組みを活用

- メッシュ単位で集計された人流 BD を、多くのベンダーが提供
- 宿泊施設の立地情報はオープンデータから入手可能
- 統計の表章（e.g. 都道府県）にあわせて該当メッシュの人数を集計、統計と照合

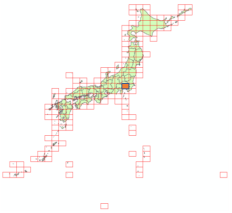


図 1. メッシュ統計のイメージ

BD × メッシュ → 宿泊施設の全数調査に相当する情報を抽出できる可能性

宿泊者数推計の概要　Step 1

宿泊施設周辺人口の特定

1. 宿泊施設が位置する 500m メッシュを特定
2. 携帯位置情報データ（人流データ）から当該メッシュの午前４時台の滞在人口を求める
3. 滞在人口から常住人口を控除

この値を県単位で月次集計したものを本研究では「指標値」と呼ぶ

図 2. 宿泊施設が位置するメッシュ（イメージ）

宿泊者数推計の概要　Step 2

回帰補正

- 公的統計の結果が公表されている年月（T 期）までで、Step 1 で求めた指標値の系列を統計の宿泊者数の系列に回帰
- 複数の推計モデルを想定し、データへの当てはまりが良い式を選択

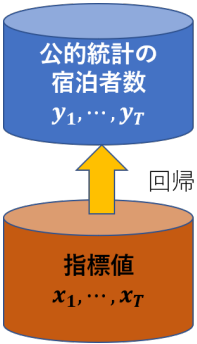


図 3. 回帰補正

宿泊者数推計の概要　Step 3

直近の宿泊者数を推計

1. 最新月の人流データから Step 1 の指標値を計算
2. Step 2 で得た推計式に指標値を当てはめ、直近（$T+1$、$T+2$ 期）の人数を外挿

毎月、最新のデータセットに更新して、一連の流れを逐次実行

公的統計の公表より時期的に 1 ヶ月半早く、宿泊者数の動向を推測

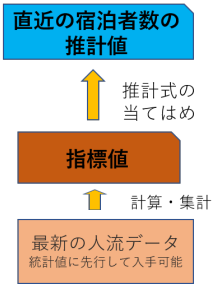


図 4. 直近の宿泊者数の推定

前回報告からの発展

令和４年２月会議報告[1]からの研究の改善、発展

- 推計の対象を神奈川県と京都府の２府県から、全国の都道府県に拡大
- 推計に用いるモデル群を拡大
  - 前年同期比を変数にしたモデルの追加
  - 直近期間のトレンドを考慮したモデルの採用 [1]
- 推計作業のシステム化、自動化 → 迅速な結果の取得

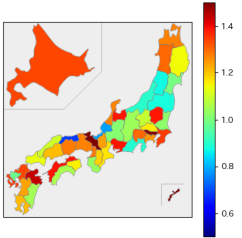


図 5. 2023 年 12 月の状況　2019 年同月比

[1]文献 [2] 参照

人流データ（携帯位置情報データ）

宿泊施設が立地するメッシュの滞在人口[1] [2] を捕捉　(日本人のみ)

使用データ　株式会社 Agoop「流動人口データ」

位置情報　スマホアプリから GPS などを通して取得

時間の粒度　時間単位で日次集計（平日、休日の別あり）

空間の粒度　全国を 500m メッシュでカバー

推計で使用する指標値は、宿泊施設が立地するメッシュ[3] ごとに下記を求めて、県別に月次集計

午前４時台の滞在人口 –「国勢調査」の常住人口　（「国勢調査」の値は時間的に最も近い調査年のものを用いる）

[1] 各メッシュの値は全体が日本人の総人口に合致するように拡大推計して提供されている

[2] 該当メッシュへの滞在が１時間に満たないケースは、その長さ（分）に応じた調整

[3] 宿泊施設の位置情報は OpenStreetMap / Link から取得。その後、緯度経度の情報をメッシュコードに変換

「宿泊旅行統計調査」

日本人延べ宿泊者数のベンチマークとして用いる

実施主体　観光庁

調査時期　毎月[1]

調査対象　ホテル、旅館などの事業所

- 従業員数 10 人以上 → 全数調査
- 従業員数９人以下 → 無作為抽出によるサンプル調査

使用変数　各都道府県の日本人延べ宿泊者数[2][3]

[1]推計では調査の翌々月末に公表される第 2 次速報値を使用

[2]例えば、１人の人が２泊すると２人となる

[3]延べ宿泊者の総数から外国人宿泊者数の値を引いた値を日本人宿泊者とみなしている

指標値と統計値

用語の再確認

指標値　人流データから「国勢調査」常住地人口を引いて作成した値

統計値　「宿泊旅行統計調査」の延べ宿泊者数

両者の関係はどうなっているのか？

指標値と統計値の関係 —全国の概況—

統計値と指標値の月次推移　全国合計

棒 → 実人数　青：統計値、橙：指標値　1e8→1 億

折れ線 → 両者の比率　統計値／指標値

- 指標値 ≫ 統計値の関係
- 差の要因として、例えば考えられることは
  - 宿泊施設周辺における早朝の活動人口（労働者や交通など）
  - 実際の観測人数に対して重みがつけられていることの影響

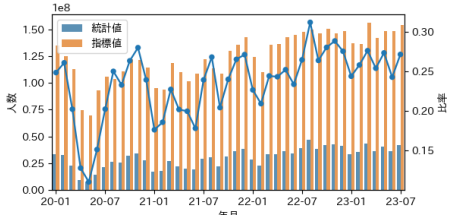


図 6. 指標値と統計値の月次推移

指標値と統計値の関係 —全国の概況—　続き

統計値と指標値の月次推移　全国合計

棒 → 実人数　青：統計値、橙：指標値　1e8→1 億

折れ線 → 両者の比率　統計値／指標値

- 比率は 0.2 から 0.3 の間でおおむね安定
  - 2020 年 4 月前後は人の動きが大きく制限され時期であり、宿泊数自体が大幅に落ち込む
  - しかし、その後のコロナ禍の中、社会環境が大きく変動する時期において、両者の動きは密接にリンクしている

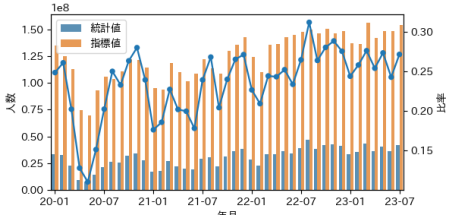


図 7. 指標値と統計値の月次推移

指標値と統計値の関係 —県別の比率—

比率（統計値／指標値）の平均

図 6と同じ 2020 年 1 月ー 2023 年 7 月の期間

比率が大きいほど、色が濃い

- 多くの県では 0.2 から 0.3 近辺の範囲におさまっている
- 県の特性による、ばらつきがみられる
  - 大都市がある県でおおむね値が小
  - 甲信、北関東での比率が大

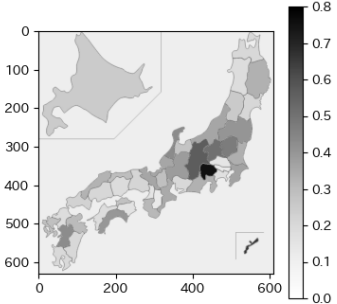


図 8. 指標値と統計値の比率

指標値と統計値の相関 —県別の相関—

両系列の相関係数を県別に提示

図 6と同じ 2020 年 1 月ー 2023 年 7 月の期間

相関が高いほど、色が濃い

- 全体として高い相関
  - 特に、宿泊者数の多い県では相関が高

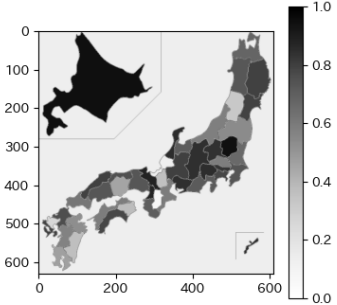


図 9. 指標値と統計値の相関

統計値と指標値の関係からいえること

- 指標値は宿泊者数そのものを代替しえない
- しかし、相関の高さから指標値の変化は宿泊者数の変化を表現している可能性があり、統計値を用いて補正ができる可能性
- 都道府県の状況の違いは考慮する必要がある

回帰による推計 —方針—

実際の宿泊者数を近似するように指標値を補正

- 指標値を統計値に回帰して、その係数を縮小係数に
- 統計の公表値が得られる期間で回帰に用いるデータを作成
- 県別に推計し、県固有の効果は制御

知りたいことは、直近の宿泊者数

- 最も当てはまりの良い推計式を用いて、最新の指標値から宿泊者数の推計値を外挿
- 翌月中旬には宿泊状況が分かる　(公的統計の公表より 1 ヶ月半早い)

回帰に用いたモデル群

回帰に用いる変数の種類と期間の違いにより、14 個の単回帰モデルを候補に設定[1]

- 変数：“実人数”、“前年同月比”、“対数前年同月比” の３種
- 回帰させるデータの期間
  - 分析の始期を固定したもの
  - 自動線形区分回帰（文献 [1]）を適用し、そのアルゴリズムに基づき期間を分割したもの
  - 終期はいずれも、試算時点で得られる最新の統計データの月まで

[1]詳細は Appendix、および文献 [3] を参照

モデルの選択

- 観測データ（のみ）の当てはまりを考慮
- モデル選択の指標：直近 6 か月の平均二乗パーセント誤差の平方根（RMSPE）
- 推計のつど（各月、各県）、用いるデータに照らし選択する

全国の概況 —2023 年 12 月の結果—

前年同月比と 2019 年同月比　　分母の値には「宿泊旅行統計調査」の第２次速報値を使用

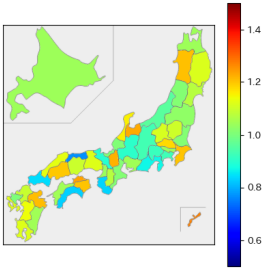


図 10. 前年同月比

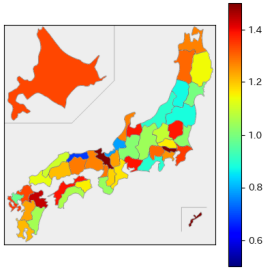


図 11. 2019 年同月比

推計の精度評価 —設定—

- 推計後に公表された統計値をテストデータにして、推計値と統計値の乖離率を評価
- ここでの推計値は 1 か月先の値
- 期間：2022 年 1 月–2023 年 7 月 (19 か月)

乖離率

各月の推計に用いた回帰（訓練）データの終期を T として、得られた推計式に指標値を適用した $\hat{y}_{T+1}$ と、事後的に得られた統計値 y_{T+1} から、次式について、

$$\frac{|\hat{y}_{T+1} - y_{T+1}|}{y_{T+1}},$$

上記の 19 か月 $\times$47 都道府県分求める

乖離率の分布 —日本全体—

乖離率（絶対値）の分布を提示

47 県 $\times$ 19 ヶ月 $=$ 873 のデータポイント

- 乖離率が大きくなるにしたがって単調減少
  - 0.2（=20%）以内に全体の8 割強
  - 0.1(=10%) 以内に全体の 5 割強
- 平均値は 0.11、中央値は 0.09

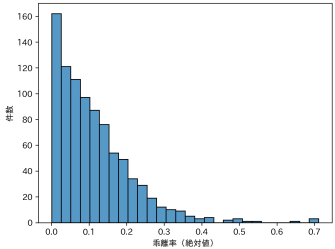


図 12. 乖離率の分布

乖離率の分布 —県別—

県別に乖離率（絶対値）の平均を提示

- 県によってばらつきがあり
- 宿泊者数が多い県で乖離率が小さい傾向

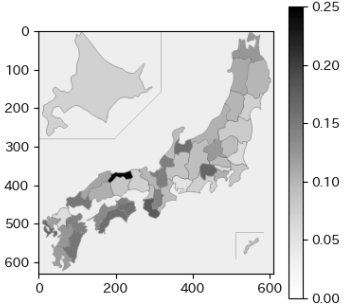


図 13. 県別の乖離率

推計方式の課題

ただし，平均して乖離率が小さい県においても、乖離率の大きい月がある

データに起因する問題以外にも、モデル選択にあたって、

- 観測データのみの当てはまりで評価している
- 選択の指標が 6 か月の情報しか用いていない

そのため、過剰適合（over fitting）が発生。推計方式に改良の余地あり

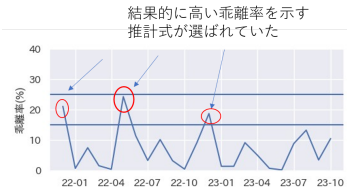


図 14. 例：東京の乖離率の推移

まとめ 1

目的と方法

- 携帯端末の位置情報から得られた人数を用いて、日本人延べ宿泊者数を都道府県別に試算
  - 宿泊施設の位置する 500m メッシュでの、午前 4 時台の滞在人数に着目
  - 公的統計の情報を用いて、その値を回帰補正
  - 回帰係数と最新の人流データを用いて、直近の宿泊者数を外挿
  - 翌月の中旬、公的統計の公表に 1.5 ヶ月先行して試算が可能

まとめ２

結果

- 全体的に見れば、一定の推計精度を実現
- すべての県、時点での精度保証には難あり
  - 地域ブロックなど、地理的範囲や人口規模を確保した上で傾向把握に
- 指標値と統計値の乖離要因を解明することは、引き続きの検討課題

今後の研究の可能性

- 性別や年齢等、宿泊者の属性による細分化
- 市区町村や複数の行政区画にまたがる観光エリアでの推計
- 外国人宿泊者数の捕捉
- 遊園地の入場者数など、特定の場所に一定時間とどまる人数の人流 DB での把握　　など

まとめ３

BD 利活用、そのパイロット研究としての意義

- BD 由来の情報と公的統計値との関連性を実証
- 研究過程における BD の特性把握（資料 3-2 参照）
- データ間の共通座標として、メッシュ統計の有用性を提示
- 使用データはすべて一般に入手可能であり、研究の再現性と汎用性

メッシュ統計の概要

[label=mesh] メッシュ統計とは、緯度・経度に基づき地域を隙間なく網の目（メッシュ）の区域に分けて、それぞれの区域に関する統計データを編成したもの

- ほぼ同一の大きさ・形であるため、地域メッシュ相互間の事象の計量的比較が容易
- 位置や区画が固定されているため、行政区域や地形の変化の影響を受けず、時系列比較が容易
- メッシュデータの合算により、任意の地域のデータの入手・分析が容易

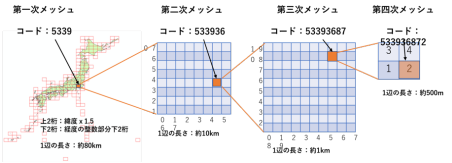


図 15. 地域メッシュ統計の概要

用いた計算基盤について

扱うメッシュの数は多く、用いるデータのサイズは大

- データの集計、分析 → 相応の計算能力が必要
- 結果の公表 → 継続性、迅速性を実現するために処理の自動化が必要

双方を可能にする基盤として MESHSTATS を利用

- MESHSTATS とは大量のメッシュ統計を世界規模で取り扱うことができる分析基盤
- 横浜市立大学大学院データサイエンス研究科　佐藤彰洋研究室で MESHSTATS の機能開発が行われ、（一社）世界メッシュ研究所 / Link がその運用を担当している

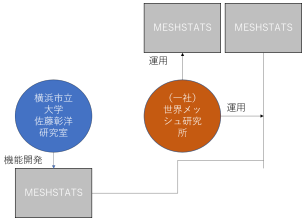


図 16. MESHSTATS の概念図

用いた計算基盤について　続き

MESHSTATS 上での処理は大きく 3 つの
モジュール群から構成

1. データの取り込み、指標値作成
2. 推計の実行
3. 結果の取りまとめ
   （フロントエンド）

   1,2 には API 機能の実装あり

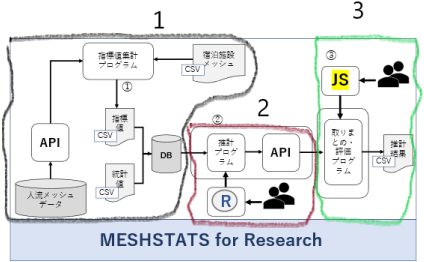


図 17. MESHSTATS での処理の流れ

指標値の計算

指標値 x_t は次のように計算される.

$$x_t = \sum_{i \in \mathcal{M}} \mathbf{1}_A(i) \left\{ pop_{it} - \widetilde{pop}_i(t) \right\}.$$

- i: 第４次メッシュ (約 500m) のコード，t: 時点（月次）
- pop_{it}: 人流データから得られたメッシュ i，t 期の午前４時台の滞在人数（月延べ数)[1]
- $\widetilde{pop}_i(t)$: pop_{it} に対応する「国勢調査」の常住地人口[2]　(t 期の月の日数で換算)
- A: 宿泊施設が存在するメッシュコードの集合
- $\mathbf{1}_A(i)$: i が A に属す場合は 1 を，それ以外は 0 を返す指示関数
- $\mathcal{M}$: 推計を行う都道府県に位置するメッシュコードの集合

[1]データは平日，休日別の 1 日あたり平均で提供されており，当該月の平日休日数に合わせて月の延べ人数を再集計
[2]t が 2020 年以降は 2020 年の「国勢調査」の値を，それ以前の時点では，前後，時間的に最も近い調査年の値を使用

回帰

指標値の集合を $\mathcal{X} \in \mathbb{R}$，宿泊者数の実績を示す統計値の集合を $\mathcal{Y} \in \mathbb{R}$ として，推計ルール $h : \mathcal{X} \to \mathcal{Y}$ を考える．
本研究では h を求めるアルゴリズムに原則、最小二乗法（OLS）による線形回帰を想定．

回帰に用いるデータ: 指標値（$x_t \in \mathcal{X}$）と統計値（$y_t \in \mathcal{Y}$）のペアデータ $S = \{(x_1, y_1), \ldots, (x_T, y_T)\}$ を構成．T は推計時における統計値の最新公表年月．

モデル その1

期間（始期）の異なる 14 個のデータセット $S_j\,(j = 1, \ldots, 14)$ を用意し，それぞれで OLS を実行.
適宜，変数変換を行ったデータを使用する

実人数を用いる 6 モデル $(j = 1, \ldots, 6)$

- y_t の推計モデルとして $h_S(x_t) = a + bx_t;\ a, b \in \mathbb{R}$ を想定
- 始期を固定した 3 つの期間，自動線形区分回帰による 3 つの設定，計 6 つの $S_j = \{(x_1, y_1), \ldots, (x_T, y_T)\}$ を用いて、それぞれで a, b を推定

モデル その2

前年同月比を用いる 4 モデル $(j = 7, \ldots, 10)$

- 指標値，統計値をそれぞれ $\phi_x : x_t \mapsto x_t/x_{t-12}$，$\phi_y : y_t \mapsto y_t/y_{t-12}$ で変数変換
- $\phi_y(y_t)$ の推計モデルとして $f(\phi_x(x_t)) = a + b \cdot \phi(x_t)$; $a, b \in \mathbb{R}$ を想定
- 始期を固定したものと、自動線形区分回帰による 3 つの設定区間，計 4 つの期間のデータセット S_j
- S_j に対応する変換後の系列 $S_j^{\phi} = \{(\phi_x(x_{13}), \phi_y(y_{13})), \ldots (\phi_x(x_T), \phi_y(y_T))\}$ を用いて，それぞれで a, b を OLS で推定
- y_t の推定モデルとしては $h_S = \phi_y^{-1} \circ f \circ \phi_x$
- 実人数に戻した時に負の値になる可能性がある

モデル その3

対数前年同月比を用いた 4 モデル ($j = 11, \ldots, 14$)

- 指標値，統計値をそれぞれ $\psi_x : x_t \mapsto \log_{10}(x_t/x_{t-12})$，$\psi_y : y_t \mapsto \log_{10}(y_t/y_{t-12})$ で変数変換
- $\psi_y(y_t)$ の推計モデルとして $g(\psi_x(x_t)) = a + b\psi(x_t)$; $a, b \in \mathbb{R}$ を想定
- データセット S_j は，始期を固定したものと、自動線形区分回帰による 3 つの設定区間，計 4 つ
- S_j に対応する変換後の系列 $S_j^{\psi} = \{(\psi_x(x_{13}), \psi_y(y_{13})), \ldots (\psi_x(x_T), \psi_y(y_T))\}$ を用い，a, b を OLS で推定
- 実測値 y_t の推定モデルとしては $h_S = \psi_y^{-1} \circ g \circ \psi_x$
- 実人数に戻したときに負にはならない
- 戻した際に誤差が大きくなる

モデルの選択

各 S_j について，先述の方式を用いて具体的なパラメータの値を求めたモデルを $h^*_{S_j}$ とおき，その集合を $\mathcal{H}^* = \left\{h^*_{S_1}, \ldots, h^*_{S_{14}}\right\}$ とする．それらの中から下記の 6 か月平均 RMSPE

$$L(h^*_{S_j}) = \sqrt{\sum_{t=T-5}^{T} \left(\frac{h^*_{S_j}(x_t) - y_t}{y_t} \right)^2}$$

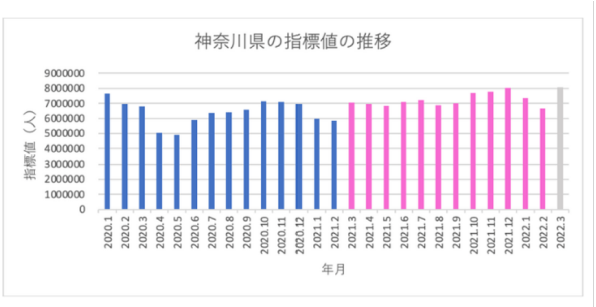
を最小にするモデル

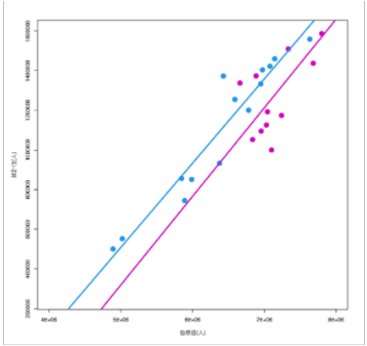
$$h^\star \in \underset{h^*_{S_j} \in \mathcal{H}^*}{\operatorname{argmin}} L(h^*_{S_j})$$

を直近の宿泊者数の試算に用いるモデルとする．

宿泊者数の推計値として $h^\star(x_{T+1})$，$h^\star(x_{T+2})$ を求める．

自動線形区分回帰のイメージ





トレンドの切り替わりを自動で検知して、区間を分割する方法

切り替わる区間を検知するために尤度比検定を行う

図 18. 自動区分線形回帰のイメージ

データの期間

終期はいずれも最新の統計値まで

始期を固定したものについては、変数のタイプに応じて下記５種類

変数	始期	備考
実人数	2017 年 01 月	人流データの基となるアプリユーザー数が 100 万人を超過
	2019 年 10 月	対象アプリの大規模な入れ替えあり
	2020 年 04 月	新型コロナウイルス感染症の流行拡大
前年同月比	2018 年 01 月	アプリユーザー数が前年同月に 100 万人を超過
対数前年同月比	2018 年 01 月	アプリユーザー数が前年同月に 100 万人を超過

自動線形区分回帰を用いたモデルについては、分割の最短期間を設定。いずれの変数とも 6 か月、12 か月、18 か月の 3 パターン

参考文献 I

[1] Aki-Hiro Sato.
A comprehensive analysis of time series segmentation on japanese stock prices.
Procedia Computer Science, 24:307–314, 2013.
17th Asia Pacific Symposium on Intelligent and Evolutionary Systems, IES2013.

[2] 総務省統計改革実行推進室.
「人流データを活用した宿泊旅行統計調査延べ宿泊者数の推計」.
https://www.soumu.go.jp/main_sosiki/singi/toukei/bigdata/02toukatsu01_04000416.html, 2022.
ビッグデータ等の利活用に関する産官学協議のための連携会議（第 16 回）資料 1.

[3] 松井伸司・佐藤彰洋.
人流メッシュ統計データを使用した宿泊者数推計シミュレーションについて.
ESTRELA, 351:16–22, 2023.