書類番号
2

2019年3月6日
物 価 統 計 室

宿泊料のウェブスクレイピングによる価格取集及び指数作成方法について

１．経緯・目的

政府における経済統計の改善の取組として、より正確な景気判断に資する基礎統計改善、国民経済計算の加工・推計手法の改善等のため、「公的統計の整備に関する基本的な計画」（平成30年３月６日閣議決定）等において、消費者物価指数の2020年基準改定におけるインターネット販売価格の採用の可否を検討し、2018年度までに結論を得ることとされている。

このため、近年の消費者のインターネットを利用した購入割合の状況を踏まえ、2018年10月の物価指数研究会において検討を進めることとした「旅行サービス」の３品目（航空運賃、宿泊料、外国パック旅行費）のうち、宿泊料について、ウェブスクレイピング技術を用いた価格取集及び指数作成方法の検討を行った。

２．ウェブスクレイピングによる価格取集の状況

（１）価格の選定条件

下記①～③の選定条件の下、2018年６月以降、ウェブスクレイピングによりインターネット販売価格の取集を行った。

なお、試験的な価格取集を行うに当たり、事前に実施したアンケート結果[1]（図表１、２）を踏まえ、価格取集の時期及び取集対象を選定した。

①　価格取集時期及び取集サイト

宿泊日の１～３週間前及び１か月以上前に予約する者が多い傾向が見られたため、取集日（毎月上旬）の翌月及び翌々月の全日の価格(※)を取集した。

(※) 取集日が2018年６月の場合、同年７月１日～8月31日の全日の価格

また、インターネットによる予約の場合、旅行予約サイトを用いる者が最も多かったことから、取集サイトとして主要旅行会社３社の旅行予約サイトを選定した。

なお、ウェブスクレイピングでは、サイト構造による個別の設定が必要なことから、同一サイトから複数の価格を取集できる総合予約サイトから取集することは実務面からも効率的と言える。

[1] 2017年３月に実施したインターネットによるアンケート調査。調査対象は調査会社のパネルより直近１年間に宿泊を伴う国内旅行をした者（パックツアー利用者を除く）を抽出。年齢階級別人口比でサンプル数を配分。回答数2448人。

図表　１：予約時期と予約方法のクロス表（アンケート結果）

N=2448		予約時期				
		1 週間以内	1～3 週間前	1 か月以上前	不明	総計
予約方法	宿に直接電話	3%	4%	5%	1%	13%
	宿のホームページ	2%	7%	12%	1%	21%
	旅行予約サイト	7%	21%	29%	2%	59%
	店頭カウンター	0%	1%	2%	0%	3%
	その他	0%	0%	1%	0%	1%
	不明	0%	0%	1%	2%	3%
	総計	12%	33%	50%	6%	100%

②　宿泊施設

2018 年６～７月は特定地域の宿泊施設、８～９月は現行の宿泊料調査の対象宿泊施設に限定して価格取集を行っていたが、10 月以降は「宿泊旅行統計調査」（観光庁）の旅行目的地（都道府県）別宿泊者数などを基に、毎月 50 施設程度を追加して価格取集を行った。都道府県内の宿泊施設の選定に当たっては、収容人数などの施設規模を考慮した。

③　宿泊プラン

現行調査においては、旅館は『和室・１泊２食付き』、ホテルは『洋室・１泊朝食付き』のプランを指定している。アンケート結果においても現行調査で指定しているプランの利用が多いことが確認できたため、これらのプランの価格を取集した。（図表　２）

なお、出張向けのビジネスプラン、高価格帯のスイートルームなど、消費者物価の対象とすべきではないプランや品質差があるプランは、本来であれば集計から除外すべき価格であるが、一律な除外が困難なことから、これらを含めた全ての価格を取集した。ただし、後述する外れ値の除外処理により、一般的な宿泊料金に対して極端に高い（またはセールにより極端に安い）料金設定のプランは除外している。

図表　２：部屋の種類と食事の種類のクロス表（アンケート結果）

N=2448	洋室	和室	和洋室	その他	総計
食事なし	24%	4%	1%	1%	29%
朝食付き	24%	3%	1%	0%	29%
朝夕食付き	11%	22%	7%	0%	40%
朝･昼･夕食付き	1%	1%	0%	0%	2%
その他	0%	0%	0%	0%	0%
総計	60%	30%	9%	1%	100%

（２）価格取集の結果

上記（１）の選定条件で取集した個別価格数は図表　３のとおりである。個別価格数を施設数、サイト数、日数で除した単位あたりの価格数は、一時的な取集の減少があった 2018 年 7 月を除くと 23 価格～32 価格で推移しており、十分な価格数が得られていると言える。

なお、2018 年 7 月の価格数の減少はプラン数に上限を設けて取集を制限した

結果であり、ウェブスクレイピングは安定して実行できている。

図表　３：ウェブスクレイピングによる価格取集の結果

取集年月	予約月	個別価格数	施設数	単位あたり価格数	平均価格（円）
2018/06	7 月	519181	230	24.27	16,598
	8 月	489515	230	22.89	17,710
2018/07	8 月	377495	229	17.73	17,177
	9 月	378404	229	18.36	15,735
2018/08	9 月	890815	313	31.62	18,973
	10 月	900645	313	30.94	18,831
2018/09	10 月	825352	314	28.26	18,728
	11 月	758055	314	26.82	18,334
2018/10	11 月	864618	355	27.06	17,827
	12 月	923240	355	27.96	17,580
2018/11	12 月	964409	400	25.92	17,381
	1 月	1003384	400	26.97	16,665
2018/12	1 月	1061248	444	25.70	16,350
	2 月	1024270	444	27.46	17,249
2019/01	2 月	1107386	491	24.25	17,040
	3 月	1290077	491	31.28	16,988

注１）６～７月は、特定地域の宿泊施設の価格を取集。集計に当たっては、１月取集対象施設に限定している。
注２）８～９月は、現行の宿泊料調査の対象宿泊施設に限定して価格を取集。
注３）10 月以降は、毎月対象施設を 50 程度ずつ追加して価格を取集。

３．価格指数の試算

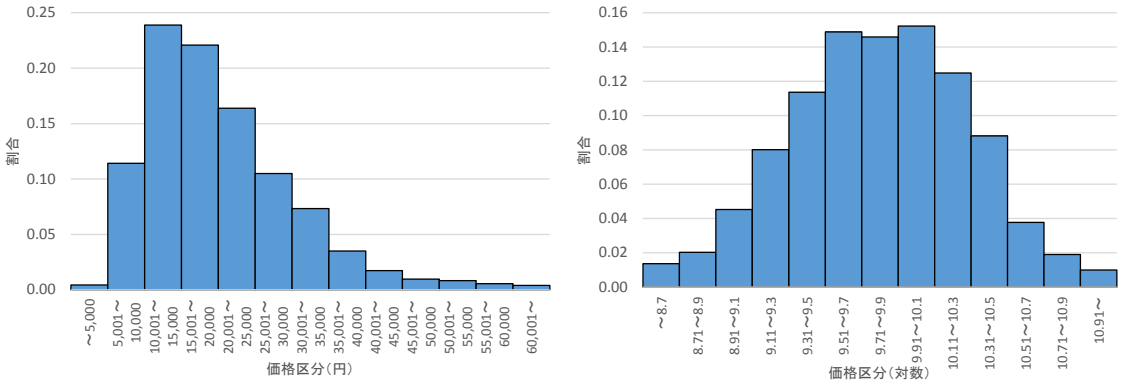
（１）外れ値の除外

上述のように、検索条件に合致する全てのプランを取集しているため、極端に価格が高い又は低い値が取集されることがある。このような価格帯のプランは、他の価格との間に大きな品質差があることや、タイムセールのような一時的に安い価格が設定されている可能性があり、価格指数を作成する上では外れ値として除外することが適切であると考えられる。ここでは、外れ値の除外方法を検討する。

取集した個別価格の分布（図表　４）を見ると、右側の高価格帯に裾広がりの分布をしているため、対数変換を行い正規分布に近づけた上で外れ値の除外処理を行う。詳細な手順は以下のとおり。

図表　４：個別価格の分布（９月取集・１１月予約）

左：実数　右：対数



① 予約サイト別(s)、予約日別(a)、宿泊施設別(b)、プラン別(c)の個別価格を$P_{s,a,b,c}$とし、これを対数変換する。

$$Y_{s,a,b,c} = \log(P_{s,a,b,c}) \quad \cdots (1)$$

② 予約サイト別、予約日別、宿泊施設別の平均価格と標準偏差を計算する。（$N_{s,a,b}$は予約サイト別、予約日別、宿泊施設別のプラン数）

$$Y_{s,a,b} = \frac{1}{N_{s,a,b}} \sum_{c}^{N_{s,a,b}} Y_{s,a,b,c} \quad \cdots (2)$$

$$\sigma_{s,a,b} = \sqrt{\frac{1}{N_{s,a,b}-1} \sum_{c}^{N_{s,a,b}} \left(Y_{s,a,b,c} - Y_{s,a,b}\right)^2} \quad \cdots (3)$$

③ 予約サイト別、予約日別、宿泊施設別に、平均価格との差が標準偏差の絶対値の 3 倍を超える個別価格を外れ値とする。

$$\left|Y_{s,a,b,c} - Y_{s,a,b}\right| > 3\sigma_{s,a,b} \quad \cdots (4)$$

データクリーニングにより除外された個別価格数は、全期間を通しておよそ0.5%程度となった。また、データクリーニング後の平均価格は 60 円程度低くなった。（図表 5）

図表 5：データクリーニングの結果

取集年月	予約日	データクリーニング前		データクリーニング後		除外された個別価格数	価格差（円）
		個別価格数	平均価格（円）	個別価格数	平均価格（円）		
2018/06	7 月	519181	16,598	516380	16,529	2801	−69
	8 月	489515	17,710	486870	17,643	2645	−66
2018/07	8 月	377495	17,177	375875	17,125	1620	−53
	9 月	378404	15,735	376941	15,696	1463	−40
2018/08	9 月	890815	18,973	886568	18,922	4247	−51
	10 月	900645	18,831	896201	18,769	4444	−62
2018/09	10 月	825352	18,728	821183	18,665	4169	−63
	11 月	758055	18,334	754258	18,270	3797	−63
2018/10	11 月	864618	17,827	860242	17,755	4376	−72
	12 月	923240	17,580	918937	17,512	4303	−68
2018/11	12 月	964409	17,381	959663	17,313	4746	−68
	1 月	1003384	16,665	998674	16,598	4710	−67
2018/12	1 月	1061248	16,350	1056019	16,281	5229	−69
	2 月	1024270	17,249	1019409	17,177	4861	−72
2019/01	2 月	1107386	17,040	1102322	16,978	5064	−62
	3 月	1290077	16,988	1284253	16,923	5824	−64

さらに、外れ値を除外した個別価格について、予約サイト別、予約日別、宿泊施設別の平均価格を計算し、これらを属性とするデータテーブルを作成する。（$N'_{s,a,b}$は外れ値を除いた個別価格数）

$$Y'_{s,a,b} = \frac{1}{N'_{s,a,b}} \sum_{c}^{N'_{s,a,b}} Y_{s,a,b,c} \quad \cdots (5)$$

（５）式の平均価格の計算では、同一の予約サイト、予約日、宿泊施設のプラン別個別価格において、プランの違いにより多少の品質差はあるものの、平均することで品質が均一化されるとみなしている。また、幾何平均による指数化により、均一化された品質が相殺されていると考えることもできる。

また、これを指数変換することで、幾何平均価格となる。

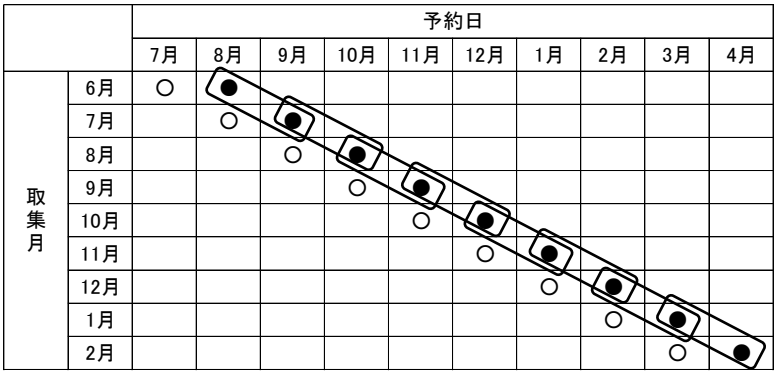
$$P_{s,a,b} = \exp\left[Y'_{s,a,b}\right] \quad \cdots (6)$$

（２）欠測値の補完

データクリーニング後の平均価格において、予約日、宿泊施設を設定してサイト検索した結果、サイト上に個別価格が表示されなかった場合は、この検索条件での平均値が計算できないため、データテーブルに欠測値が生じる。指数計算において欠測値を無視した平均価格の計算（完全ケース分析）では、曜日別の欠測に差があることで欠測がランダムでなくなり、平均価格に偏りが生じる場合がある。また、平均価格の計算段階での代入（平均値補完）では、平均の計算順序により指数計算結果が変わってしまうため、注意が必要である。ここでは、実測値データセットの回帰分析から欠測値を推定し補完する手法（回帰補完）を検討する。

なお、後述のように、指数算式は前月比連鎖方式を想定しているため、連続する２か月間のプールデータを用いることで、２か月間の取集対象施設の差異の調整も回帰補完に含ませることを考えている。今回の試算では、取集データのうち２か月先のデータテーブルを前月と当月の２か月分使用する（図表 ６）。

図表 ６：データセット



回帰補完[2]

まず、価格が欠測となったレコードを除いたデータセットを作成し、価格を被説明変数$\mathrm{y_{obs}}$、検索条件

・宿泊日　　：$\boldsymbol{x_1} = (x_{1,1}, \cdots, x_{1,A-1})$　　A: 前月と当月の合計日数
・予約サイト　：$\boldsymbol{x_2} = (x_{2,1}, \cdots, x_{2,S-1})$　　S: 予約サイト数$(3\ \text{サイト})$
・宿泊施設　　：$\boldsymbol{x_3} = (x_{3,1}, \cdots, x_{3,B-1})$　　B: 宿泊施設数

を説明変数（ダミー変数）とする回帰分析を行う。

$$y_{\mathrm{obs}} = \alpha + \boldsymbol{\beta_1} \cdot \boldsymbol{x_1} + \boldsymbol{\beta_2} \cdot \boldsymbol{x_2} + \boldsymbol{\beta_3} \cdot \boldsymbol{x_3} + \varepsilon \qquad \cdots (7)$$

次に、推定された回帰モデルに基づいて、価格が欠測となったレコードの属性情報（宿泊日:$\boldsymbol{x'}_1$、予約サイト:$\boldsymbol{x'}_2$、宿泊施設:$\boldsymbol{x'}_3$）を用いて、価格の推計値$\widehat{\mathrm{y_{mis}}}$を計算し、補完値として代入する。

$$\widehat{y_{\mathrm{mis}}} = \hat{\alpha} + \widehat{\boldsymbol{\beta_1}} \cdot \boldsymbol{x'}_1 + \widehat{\boldsymbol{\beta_2}} \cdot \boldsymbol{x'}_2 + \widehat{\boldsymbol{\beta_3}} \cdot \boldsymbol{x'}_3 \qquad \cdots (8)$$

補完後のデータセットにおいて、月ごとに平均価格を計算すると、それらは宿泊施設と予約サイトが２か月間で共通した、全ての宿泊日に価格が存在するデータセットの平均価格となる。

また、２か月間の平均価格の差から、幾何平均による価格比を計算することができる。

$$\exp\left[\frac{1}{N_t}\sum y_t - \frac{1}{N_{t-1}}\sum y_{t-1}\right] = \exp\left[\frac{1}{N_t}\sum \log(p_t) - \frac{1}{N_{t-1}}\sum \log(p_{t-1})\right]$$

$$= \frac{(\prod p_t)^{1/N_t}}{(\prod p_{t-1})^{1/N_{t-1}}} \qquad \cdots (9)$$

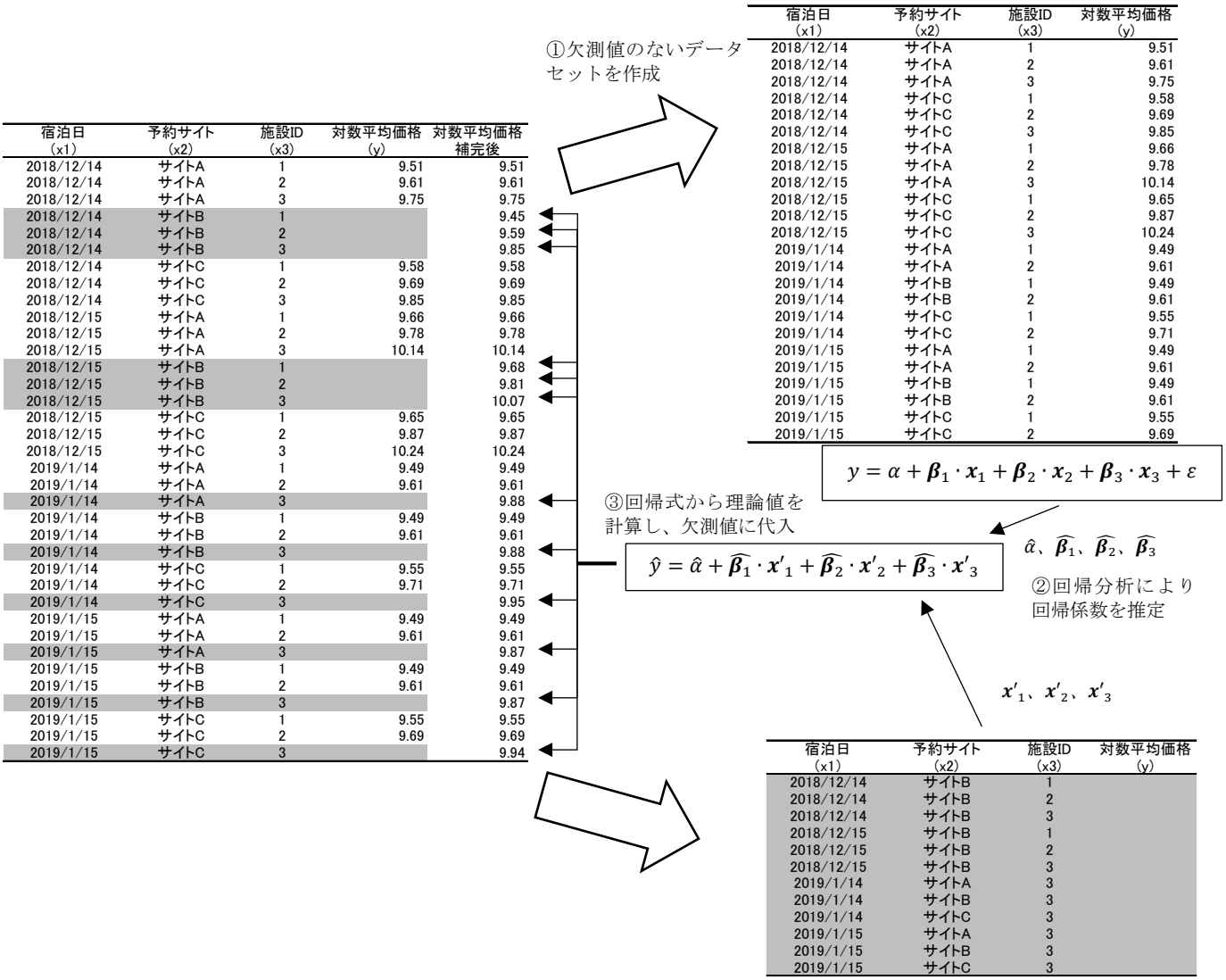
これを前月の指数に乗ずることで、当月の価格指数を求める。

$$I_t = I_{t-1} \times \frac{(\prod p_t)^{1/N_t}}{(\prod p_{t-1})^{1/N_{t-1}}} \qquad \cdots (10)$$

連続する２か月間のデータセットを用いて回帰分析を行うことにより、当月に新たに取集できた価格や、当月から予約を受け入れなくなったような月単位での宿泊施設の出入りによる平均価格の変動も同じ回帰係数によりまとめて調整することができる。

[2] 内閣府経済社会総合研究所景気統計部(2017)「欠測値補完に関する調査研究報告書【詳細版】を参考にした。

図表 7：回帰補完



宿泊日 (x1)	予約サイト (x2)	施設ID (x3)	対数平均価格 (y)	対数平均価格 補完後
2018/12/14	サイトA	1	9.51	9.51
2018/12/14	サイトA	2	9.61	9.61
2018/12/14	サイトA	3	9.75	9.75
2018/12/14	サイトB	1		9.45
2018/12/14	サイトB	2		9.59
2018/12/14	サイトB	3		9.85
2018/12/14	サイトC	1	9.58	9.58
2018/12/14	サイトC	2	9.69	9.69
2018/12/14	サイトC	3	9.85	9.85
2018/12/15	サイトA	1	9.66	9.66
2018/12/15	サイトA	2	9.78	9.78
2018/12/15	サイトA	3	10.14	10.14
2018/12/15	サイトB	1		9.68
2018/12/15	サイトB	2		9.81
2018/12/15	サイトB	3		10.07
2018/12/15	サイトC	1	9.65	9.65
2018/12/15	サイトC	2	9.87	9.87
2018/12/15	サイトC	3	10.24	10.24
2019/1/14	サイトA	1	9.49	9.49
2019/1/14	サイトA	2	9.61	9.61
2019/1/14	サイトA	3		9.88
2019/1/14	サイトB	1	9.49	9.49
2019/1/14	サイトB	2	9.61	9.61
2019/1/14	サイトB	3		9.88
2019/1/14	サイトC	1	9.55	9.55
2019/1/14	サイトC	2	9.71	9.71
2019/1/14	サイトC	3		9.95
2019/1/15	サイトA	1	9.49	9.49
2019/1/15	サイトA	2	9.61	9.61
2019/1/15	サイトA	3		9.87
2019/1/15	サイトB	1	9.49	9.49
2019/1/15	サイトB	2	9.61	9.61
2019/1/15	サイトB	3		9.87
2019/1/15	サイトC	1	9.55	9.55
2019/1/15	サイトC	2	9.69	9.69
2019/1/15	サイトC	3		9.94

宿泊日 (x1)	予約サイト (x2)	施設ID (x3)	対数平均価格 (y)
2018/12/14	サイトA	1	9.51
2018/12/14	サイトA	2	9.61
2018/12/14	サイトA	3	9.75
2018/12/14	サイトC	1	9.58
2018/12/14	サイトC	2	9.69
2018/12/14	サイトC	3	9.85
2018/12/15	サイトA	1	9.66
2018/12/15	サイトA	2	9.78
2018/12/15	サイトA	3	10.14
2018/12/15	サイトC	1	9.65
2018/12/15	サイトC	2	9.87
2018/12/15	サイトC	3	10.24
2019/1/14	サイトA	1	9.49
2019/1/14	サイトA	2	9.61
2019/1/14	サイトB	1	9.49
2019/1/14	サイトB	2	9.61
2019/1/14	サイトC	1	9.55
2019/1/14	サイトC	2	9.71
2019/1/15	サイトA	1	9.49
2019/1/15	サイトA	2	9.61
2019/1/15	サイトB	1	9.49
2019/1/15	サイトB	2	9.61
2019/1/15	サイトC	1	9.55
2019/1/15	サイトC	2	9.69

宿泊日 (x1)	予約サイト (x2)	施設ID (x3)	対数平均価格 (y)
2018/12/14	サイトB	1	
2018/12/14	サイトB	2	
2018/12/14	サイトB	3	
2018/12/15	サイトB	1	
2018/12/15	サイトB	2	
2018/12/15	サイトB	3	
2019/1/14	サイトA	3	
2019/1/14	サイトB	3	
2019/1/14	サイトC	3	
2019/1/15	サイトA	3	
2019/1/15	サイトB	3	
2019/1/15	サイトC	3	

図表 8：回帰補完結果

予約月	補完前 平均価格数	補完した 平均価格数 の合計	うち欠測による補完	うち施設数増加を調整する補完	補完後 平均価格数
2018/09	32054	10036	9856	180	42090
2018/10	40158	26820	9471	17349	66978
2018/11	48059	9952	9220	732	58011
2018/12	52473	12492	8526	3966	64965
2019/01	61140	13632	8703	4929	74772
2019/02	65511	13254	8547	4707	78765
2019/03	74857	13289	7664	5625	88146

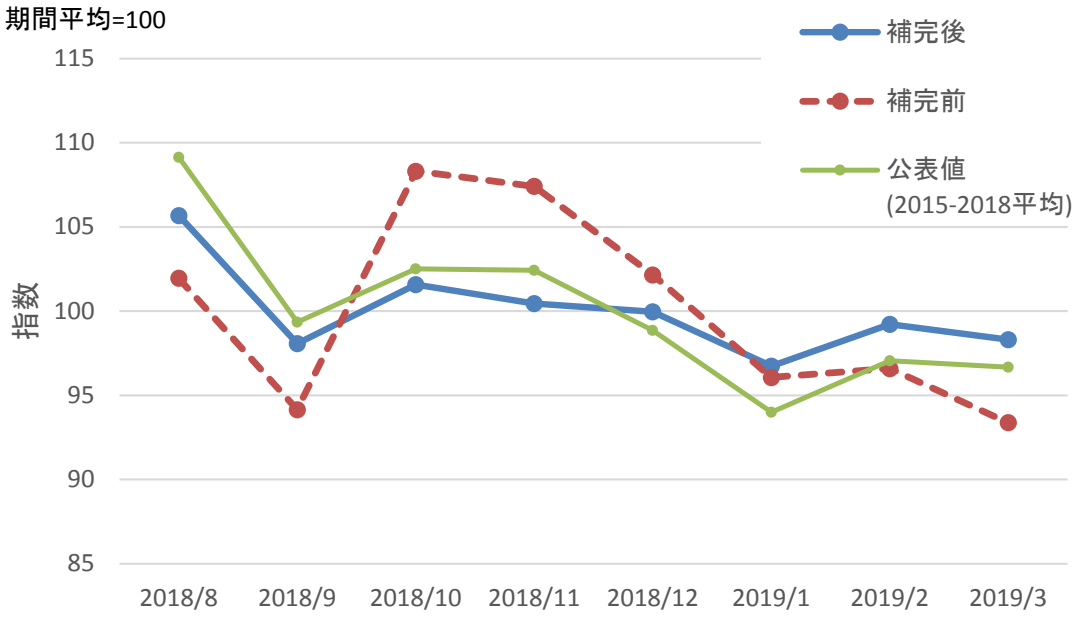
欠測値への補完結果は図表　８のとおりである。10 月以降は価格取集の対象施設を毎月追加しており、２か月間の施設数増加を調整する補完を除けば、欠測値補完数は補完後平均価格数の１～２割程度となっており、その割合は徐々に減少している。なお、10 月の施設数増加を調整する補完が急増しているのは、８月取集以降の予約サイトの検索条件を変更したため、前月と同一の施設数が大幅に減少したためである。

（３）価格指数の試算結果

価格指数の試算結果は図表　９のとおりである。2018 年 10 月以降は追加施設の価格水準差により、補完前の指数が上下に大きく動いていることがわかる。一方で、補完後の指数は月ごとの施設の違いによる平均価格の差を調整する効果により、時系列で安定した推移となっていることが分かる。

季節性を見るため、公表値の 2015 年～2018 年の４年間の平均値と比較したところ、補完後の指数はおおむね季節的な動きが捉えられていることが分かった。また、８月の指数は公表値に対して下方となったが、これは日並びの影響により 2018 年の公表値が大きく上昇したことによるものであり、ウェブスクレイピングにより毎日の価格を反映することで、調査日と日並びの関係による一時的な影響を取り除くことができる。逆に、12 月及び１月の指数は公表値に対して上方となったが、これは、公表値には年末年始の繁忙期の価格が反映されていない一方、試算値には反映されていることによる乖離であり、試算結果が実態を反映できているものと考えられる。

図表　９：指数試算結果



４．インターネット販売価格の採用の可否

これまでの検討結果より、以下の①～③の状況が確認できたことから、宿泊料については、消費者物価指数の 2020 年基準において、ウェブスクレイピングを活用してインターネット販売価格の取集を行い、それらの価格を用いて指数を作成することとしたい。

① アンケート結果から、インターネットによる予約が最も多く、インターネット販売の価格動向を捉えれば宿泊料の価格動向を適切に捉えることができると考えられること

② ウェブスクレイピングにより、各旅行予約サイトからインターネット予約価格を安定して取集できることが確認できたこと

③ 膨大な数のインターネット販売価格について、品質調整も含めて精緻に指数に反映できる見込みが立ち、また、毎日の価格を指数に反映できるなど、統計精度の向上に資すること

５．今後の価格取集方法について

これまでのウェブスクレイピングによる価格取集の状況を踏まえ、予約時期及び宿泊施設数の条件を次のとおり変更し、引き続き 2020 年基準指数の作成に向けた価格取集を行う。

（１）予約時期

これまでのウェブスクレイピングによる取集結果から、一部サイトの１か月前の取集結果において、満室のため低価格帯のプランが取集できないことにより、宿泊施設別平均値が２か月前取集に比べて異常な高価格となるケースが含まれることが分かっている。

また、2017 年８月～2018 年３月の間に実施した、30 施設に限定した長期間のウェブスクレイピングの結果によると、価格が取集できた施設数に以下の傾向が見られ、事前予約に時期的な制限があることが分かった（図表 １０）。

- ４か月先で約１割、６か月先で約半数の宿泊施設の価格が予約サイトに掲載されていないため、価格取集ができなかった。
- 特に 11 月以前について、翌年４月（網掛けセル）以降の価格の掲載はそれ以前と比べて相当少なく、年度変わりによるサイトへの価格掲載状況に断層が見られる。

図表　１０：価格取集可能施設数

	予約月										
取集月	1か月先	2か月先	3か月先	4か月先	5か月先	6か月先	7か月先	8か月先	9か月先	10か月先	11か月先
2017年8月	30	29	29	28	25	18	14	2	2	2	1
2017年9月	30	30	29	26	23	16	4	2	2	1	1
2017年10月	30	30	30	27	22	7	3	2	1	1	1
2017年11月	30	30	29	26	17	10	5	4	2	2	1
2017年12月	30	29	28	24	22	14	7	5	5	3	3
2018年1月	29	29	27	26	26	14	9	6	5	5	5
2018年2月	29	28	28	27	26	18	12	5	5	5	3
2018年3月	29	29	28	27	26	17	10	6	6	3	2
平均	30	29	29	26	23	14	8	4	4	3	2
取集割合	100%	99%	96%	89%	79%	48%	27%	14%	12%	9%	7%

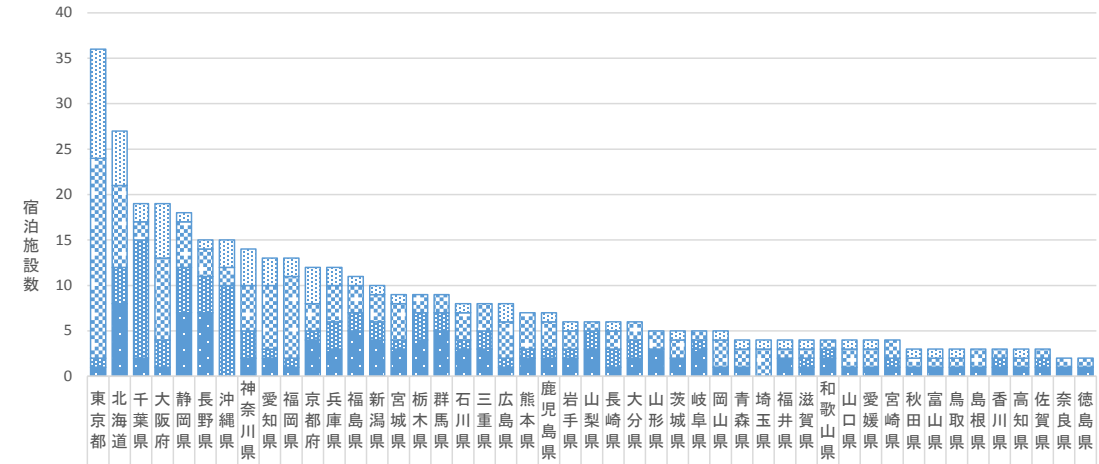
なお、特殊要因として、2020 年８月の東京五輪期間の価格は 2020 年４月以降には確実に掲載されていることが考えられるため、４月取集における８月予約日を取集するため、４か月前までの取集が必要と考えられる。

以上のことから、宿泊料の取集期間は２か月先～４か月先とし、指数作成には３か月先の価格を用い、３か月先の価格が取集できなかった場合はその前後の価格を用いることとする。

（２）宿泊施設の選定

「宿泊旅行統計調査」（観光庁）の旅行目的地（都道府県）別宿泊者数などから、比例配分により都道府県別の施設数を設定する。都道府県別施設数は、図表　１１のとおり、全国で 400 施設となるよう配分する。都道府県内の宿泊施設の選定は宿泊施設のタイプ（シティホテル・ビジネスホテル・リゾートホテル・旅館）別に行い、収容人数などの施設規模やこれまでのウェブスクレイピングの結果などから、安定的に価格取集が可能な宿泊施設を選定する。

図表　１１：都道府県別宿泊施設数



	東京都	北海道	千葉県	大阪府	静岡県	長野県	沖縄県	神奈川県	愛知県	福岡県	京都府	兵庫県	福島県	新潟県	宮城県	栃木県	群馬県	石川県	三重県	広島県	熊本県	鹿児島県	岩手県	山梨県	長崎県	大分県	山形県	茨城県	岐阜県	岡山県	青森県	埼玉県	福井県	滋賀県	和歌山県	山口県	愛媛県	宮崎県	秋田県	富山県	鳥取県	島根県	香川県	高知県	佐賀県	奈良県	徳島県
シティホテル	12	6	2	6	1	1	3	4	3	2	4	2	1	1	1	0	0	1	0	2	0	1	1	0	1	0	0	1	0	1	1	1	1	1	0	1	1	0	1	1	1	0	0	1	0	0
ビジネスホテル	22	9	2	9	5	3	2	5	7	9	3	4	3	3	4	2	2	3	3	4	4	3	2	1	2	2	2	2	1	3	2	3	1	1	1	2	2	2	1	1	1	2	1	1	1	1
リゾートホテル	1	4	13	3	5	4	10	3	1	1	1	3	2	2	1	3	2	1	2	1	1	1	1	2	2	2	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0
旅館	1	8	2	1	7	7	0	2	2	1	4	3	5	4	3	4	5	3	3	1	2	2	2	3	1	2	3	2	3	1	1	0	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1
総計	36	27	19	19	18	15	15	14	13	13	12	12	11	10	9	9	9	8	8	8	7	7	6	6	6	6	5	5	5	5	4	4	4	4	4	4	4	4	3	3	3	3	3	3	3	2	2

（参考）宿泊施設数の設定について

現行の宿泊料調査の対象施設数は 320 に設定している一方、ウェブスクレイピングによる価格取集では、リソースの制約による対象施設数の上限を考慮する必要はない。しかしながら、インターネット販売価格を取得するためのウェブサイトへのアクセスは、サイト側への負荷を考慮すると無制限に行うことはできない。このため、適切な対象施設数を設定する必要がある。

ここでは、これまでに取集したデータテーブルを用いて幾何平均価格の標準誤差率を計算し、価格指数への影響を考慮した施設数を導出する。

（１）計算方法

① ９月に取集した 11 月予約の取集データを使用する。個別価格から、３（１）により予約サイト別、宿泊施設別、予約日別のデータクリーニング後平均価格を計算し、クロスセクションデータを作成する。

② ①で作成したデータテーブルを疑似母集団とし、予約サイト×予約日（3 サイト×30 日）別に層化する。宿泊施設数を想定したサンプル数を設定し、各層に対して設定数分を復元抽出したリサンプリング標本を作成する。

③ リサンプリング標本の幾何平均価格を計算する。

④ ②、③をリサンプリング回数分（1000 回）繰り返す。

⑤ 得られた幾何平均価格の平均値と標準偏差から、リサンプリングによる標準誤差率を計算する。

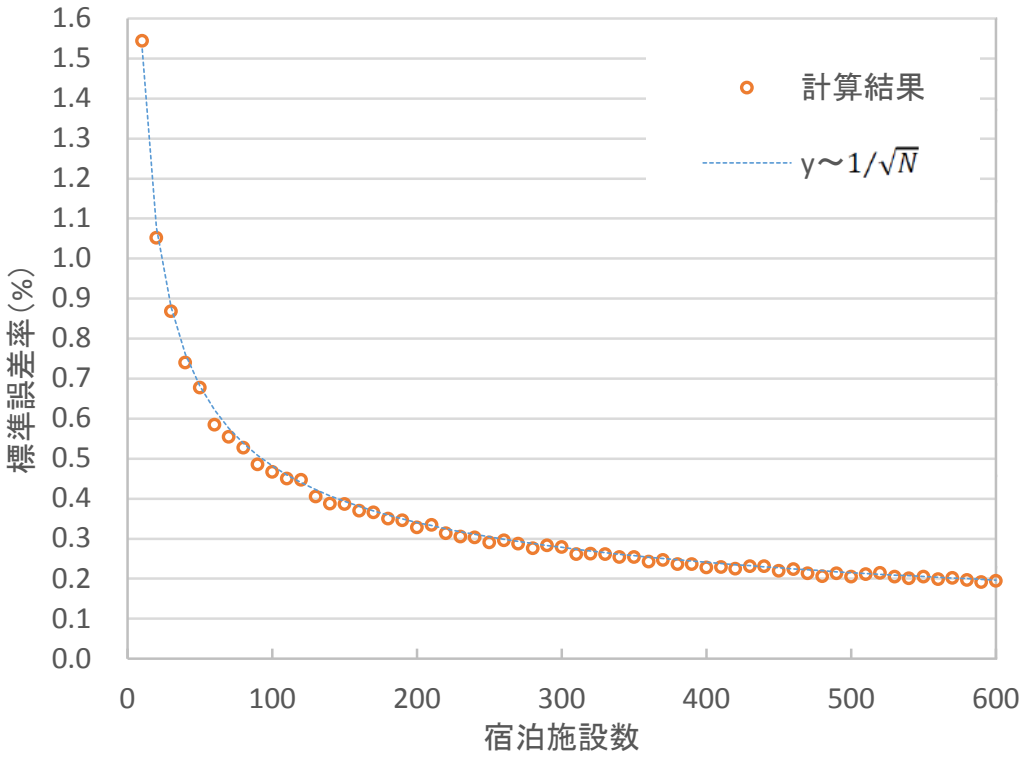
⑥ 施設数を変化させるよう②のサンプル数を設定し、10～600 施設を 10 施設おき（全体のサンプルサイズは 90×10～90×600）に変化させて②～⑤を繰り返す。（90＝３サイト×30 日）

（２）計算結果

縦軸を「標準誤差率（%）」、横軸を「宿泊施設数」としたグラフを図表 １２に示す。計算結果は$1/\sqrt{N}$に比例したグラフとなっており、現行の宿泊料調査の施設数 320 では誤差率 0.262%、施設数 400 では誤差率 0.227%となった。また、施設数が 400 を超えたあたりでは施設数の増加に対する標準誤差率の減少はほぼなくなり、横ばいとなっている。

この結果を踏まえ、今後のウェブスクレイピングによる価格取集では、施設数を 400 に設定することとしたい。

図表　１２：標準誤差率


1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
標準誤差率（%）
0
100
200
300
400
500
600
宿泊施設数
計算結果
y～1/√N