

付3 ウェブスクレイピングによる品目別価格指数の算出

ウェブスクレイピングとは、ウェブサイトから情報を抽出するコンピュータソフトウェア技術のことであり、この技術を活用することにより、ウェブサイトにおいて販売されている商品（財・サービス）情報を効率的に収集することが可能である。また一般に、ウェブサイトの情報は、販売実績のほとんどない商品（財・サービス）のデータなどが含まれるが、ウェブサイトを運営する企業から販売実績に関する詳細な情報を得ることにより、それらのデータノイズを除去し、物価動向の把握において前提とする同品質の商品価格を追跡することが可能になった。

上記の状況及び近年のネット購入の増加を踏まえ、航空運賃、宿泊料及び外国パック旅行費については、2020年基準からウェブスクレイピングを活用して消費者物価指数を作成する。これにより、価格収集期間（これまでの「各月の特定日の価格」の収集から「毎日の価格」の収集へ改善）や採用プラン数などが飛躍的に増加し、消費者物価指数の統計精度の向上に寄与する。

航空運賃、宿泊料及び外国パック旅行費は、次の手順に沿って価格指数を作成する。

1 航空運賃

(1) 予約時期及び運賃

航空会社における運賃の割引体系⁵⁵や消費者の購入時期等に合わせ、普通、各種割引運賃の各月全日の価格を採用する。ただし、サービスを充実させた上位クラス航空券、子供向け航空券、シニア向け航空券など特定者向けの航空券は除外する。

(2) 路線

旅客数の多い代表的な15路線を選定し、各区間で運行する全ての発着便を選定する。

(3) 航空会社

航空会社の旅客状況から、旅客数シェアが高い航空会社を優先的に選定する。

(4) 指数の算出方法

次の①、②の手順に沿って、価格指数を算出する。

- ① 路線別、航空会社別、運賃の種類別、搭乗日別、便別の最低価格を算出し、便数で単純平均し、路線別、航空会社別、運賃の種類別、搭乗日別の平均価格を算出する。また、平日（休前日除く。）・平日以外別の取扱数割合（ $q_{0,d}$ ）を用いて加重平均し、路線別、航空会社別、運賃の種類別の平均価格を算出する。

⁵⁵ 搭乗日の28日～44日前、45日～54日前、55日～74日前、75日以上前の購入時期に対応した割引運賃体系のこと。

$$P_{t,a,b,c,d} = \frac{\sum_e P_{t,a,b,c,d,e}}{n_{t,a,b,c,d}}$$

$$P_{t,a,b,c} = \frac{\sum_d P_{t,a,b,c,d} q_{0,d}}{\sum_d q_{0,d}}$$

t : 比較時、 0 : 基準時
 c : 運賃の種類
 a : 路線
 d : 搭乗日
 b : 航空会社
 e : 便

② 路線別、航空会社別、運賃の種類別の取扱数割合 ($Q_{0,a,b,c}$) を用いて加重平均し、平均価格を算出する。最後に、基準時価格で除して、価格指数を算出する。

$$P_t = \frac{\sum_{a,b,c} P_{t,a,b,c} Q_{0,a,b,c}}{\sum_{a,b,c} Q_{0,a,b,c}}$$

$$I_t = \frac{P_t}{P_0} \times 100$$

t : 比較時、 0 : 基準時
 a : 路線
 b : 航空会社
 c : 運賃の種類

2 宿泊料

(1) 予約時期及び宿泊プラン

旅行会社における宿泊プランの発売時期や消費者の購入時期等に合わせ、旅館は「和室・1泊2食付き」、ホテルは「洋室・1泊朝食付き」のプランの各月全日の価格を採用する*。ただし、次のとおりの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（又はセールにより極端に安い。）料金設定のプランは除外する。

(2) 宿泊施設

「宿泊旅行統計調査」（観光庁）の旅行目的地（都道府県）別宿泊者数や収容人数の施設規模等を基に、都道府県別に代表的な宿泊施設を400程度選定する。

(3) 旅行会社

主要旅行会社の旅行取扱額の状況を踏まえ、取扱額シェアが高い旅行会社を優先的に選定する。

(4) 指数の算出方法

次の①～④の手順に沿って、価格指数を算出する。算出には当月 (t) と前月 ($t-1$) の2か月分のデータセットを用いる。

① 次の手順に沿って外れ値を除外する。

* 原則として、宿泊日の2か月前の月初めに収集した価格を用いる。

- (a) 予約サイト別(s)、宿泊日別(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。

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

- (b) 予約サイト別、宿泊日別、宿泊施設別の平均価格と標準偏差を算出する。 $(N_{s,a,b}$ はプラン数)

$$Y_{s,a,b} = \frac{1}{N_{s,a,b}} \sum_{c=1}^{N_{s,a,b}} Y_{s,a,b,c}$$

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

- (c) 予約サイト別、宿泊日別、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。

$$|Y_{s,a,b,c} - Y_{s,a,b}| > 3\sigma_{s,a,b}$$

- ② 外れ値を除外した個別価格について、予約サイト別、宿泊日別、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)。

$$Y'_{s,a,b} = \frac{1}{N'_{s,a,b}} \sum_{c=1}^{N'_{s,a,b}} Y_{s,a,b,c}$$

- ③ 次の手順に沿って欠測値を補完する。

- (a) ②で集計したデータテーブルを用いて、価格 $Y'_{s,a,b}$ を被説明変数、予約サイト、宿泊日、宿泊施設を説明変数(ダミー変数)とする回帰分析を行う⁵⁶。

$$Y'_{s,a,b} = \alpha + \beta_s \cdot x_s + \beta_a \cdot x_a + \beta_b \cdot x_b + \varepsilon$$

説明変数

予約サイト	$x_s = (x_{s,1}, \dots, x_{s,S-1})$	S: 予約サイト数
宿泊日	$x_a = (x_{a,1}, \dots, x_{a,A-1})$	A: 前月と当月の合計日数
宿泊施設	$x_b = (x_{b,1}, \dots, x_{b,B-1})$	B: 宿泊施設数

⁵⁶ 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、当月に新たに収集できた価格や、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。

- (b) 推定された回帰モデルに基づき、欠測値となった予約サイト、宿泊日、宿泊施設の組合せにおいて、その属性情報（予約サイト： x_s' 、宿泊日： x_a' 、宿泊施設： x_b' ）を用いて価格の推計値 \widehat{y}_{mis} を計算し、補完値として代入する。

$$\widehat{y}_{\text{mis}} = \widehat{\alpha} + \widehat{\beta}_s \cdot x_s' + \widehat{\beta}_a \cdot x_a' + \widehat{\beta}_b \cdot x_b'$$

- ④ 補完後のデータセットから、当月(t)と前月($t-1$)のそれぞれにおける幾何平均価格を算出する。これらの価格比を前月の価格指数に乗じることで、当月の価格指数を算出する。

$$\begin{aligned} P_t &= \left(\prod_{s,a,b} P_{t,s,a,b} \right)^{\frac{1}{N_t}} = \exp \left[\frac{1}{N_t} \sum_{s,a,b} \log(P_{t,s,a,b}) \right] \\ &= \exp \left[\frac{1}{N_t} \sum_{s,a,b} Y'_{t,s,a,b} \right] \\ I_t &= I_{t-1} \times \frac{P_t}{P_{t-1}} \end{aligned}$$

3 外国パック旅行費

(1) 予約時期及びツアープラン

旅行会社におけるプラン発売時期や消費者の購入時期等に合わせ、観光ツアーが付属していないツアープラン（フリープラン）の各月全日の価格を採用する**。

(2) 旅行先

日本人訪問者数の多い代表的な地域（アジア、北米、ヨーロッパ、オセアニア）・都市から選定する。また、パック旅行費は地域情勢によって影響を受けやすい等の理由から、毎月の価格動向を可能な限り安定的に捉えるため、原則、各地域から2都市以上を選定する。

(3) 旅行会社

主要旅行会社の旅行取扱額の状況を踏まえ、取扱額シェアが高い旅行会社を優先的に選定する。

(4) 航空会社及び宿泊施設等

各旅行会社の取扱い状況を踏まえ、輸送サービスを提供する航空会社や宿泊施設の

** 原則として、出発日の2～7か月前の各月の月初めに収集した価格を用いる。

グレード等については、各地域において安定的に供給されており取扱数量の多いものを選定する。

(5) 指数の算出方法

次の①～③の手順に沿って、価格指数を算出する。

- ① ツアープラン数で単純平均し、旅行先別、旅行会社別、出発日別の平均価格を算出する。

$$P_{t,a,b,c} = \frac{\sum_d P_{t,a,b,c,d}}{n_{t,a,b,c}}$$

t : 比較時、 0 : 基準時 a : 旅行先 b : 旅行会社 c : 出発日 d : ツアープラン

- ② 当月の日数で単純平均し、旅行先別、旅行会社別の月平均価格を算出する。また、基準時価格で除して、旅行先別、旅行会社別の指数を算出する。

$$P_{t,a,b} = \frac{\sum_c P_{t,a,b,c}}{n_{t,a,b}}$$

$$I_{t,a,b} = \frac{P_{t,a,b}}{P_{0,a,b}} \times 100$$

t : 比較時、 0 : 基準時 a : 旅行先 b : 旅行会社 c : 出発日

- ③ 各旅行先における旅行会社別取扱高割合($w_{0,a,b}$)を用いて加重平均し、旅行先別の指数を算出する。最後に、旅行先別の取扱高割合($w_{0,a}$)を用いて加重平均し、価格指数を算出する。

$$I_{t,a} = \frac{\sum_b I_{t,a,b} w_{0,a,b}}{\sum_b w_{0,a,b}}$$

$$I_t = \frac{\sum_a I_{t,a} w_{0,a}}{\sum_a w_{0,a}}$$

t : 比較時、 0 : 基準時 a : 旅行先 b : 旅行会社