

「ウェブスクレイピングを用いた価格指数 の推計に関する調査研究」 報告書

内閣府経済社会総合研究所

2021年3月

目 次

I 調査研究の概要	3
1-1 本調査研究の背景と目的	4
1-2 調査研究の概要	5
II 価格情報等の収集	6
2-1 収集対象先の抽出及び選定	7
2-2 収集した情報のデータベース作成	7
2-3 収集スケジュール	8
III 収集情報を用いた分析	9
3-1 既存文献の整理	10
3-2 分析方法	12
3-3 分析結果	13
IV 文献調査および有識者ヒアリング	24
4-1 文献調査	25
4-2 有識者ヒアリング	26
V ウェブスクレイピングによる価格指数推計のための課題整理	27
5-1 本調査研究で得られた結果	28
5-2 データ収集に関する課題	29
5-3 分析に関する課題	30

(注) 本報告書は令和元年度委託調査の報告書の概要である。

I. 調査研究の概要

- 1-1 本調査研究の背景と目的
- 1-2 調査研究の概要

1-1 本調査研究の背景と目的

○ 背景

- シェアリング・エコノミーなどデジタルエコノミーに関連した新たな経済活動が急速に普及・拡大しており、GDP統計においてこうした活動を適切に捉えることが重要になっている。
- シェアリング・エコノミーは、財やサービスの新しい提供形態であることに加え、提供者（及び価格設定者）は主に個人であることから、その提供価格の設定方法や、代替財・サービス（民泊の例でいえばビジネスホテル宿泊サービスなど）価格変動との相関など、価格動向に関する分析も十分な蓄積が行われていない。
- シェアリング・エコノミーで提供される財・サービスの価格に関する詳細な分析と実質化手法の検討への要請がある一方、公的な統計調査では、価格に関する情報が得られていない。中長期的には、こうした財やサービスの価格についても公的な統計調査の対象となることが望まれるが、当面の対応として、オンライン・プラットフォーム上に提示された個別サービス販売価格等を収集し、マクロレベルの価格指数を作成し、その動向を分析することが有益と考えられる。

○ 目的

- 本調査研究は上記背景を基に、シェアリング・エコノミーにおいて提供される一部のサービスに焦点を絞り、その産出額の実質化を行うことを目的としている。
- この分析により、例えば、類似サービス価格との季節パターンの違いや、価格水準のトレンドの有無などを調べることができる。また、将来的には作成した物価指数を用いたシェアリング・エコノミーの実質化までが可能となることが期待される。

1-2 調査研究の概要

- ・本調査研究では、ウェブスクレイピング¹を用いた価格収集を行い、既存の文献を参考とし価格指数作成の検討を行った。また国内外の最新の文献研究及び、国内事情に合致した推計方法等の研究を行い、価格収集のあり方や分析方法、分析結果の解釈などについて、指数理論の知見や価格指数作成の実務経験を有する有識者からヒアリングを行った。
- ・本調査研究は、三菱UFJリサーチ&コンサルティング株式会社に委託し、2019年8月～2020年3月に実施した。

調査研究項目	概要
価格情報等の収集	■ 収集先として適当な事業者を①収集が可能かどうか、②どのような情報が収集できるか、といった観点から、民泊、駐車場スペースシェア、ライドシェア、カーシェアの4分野で各1事業者を収集対象先として選定した。 ■ 上記4事業者のサイトに対して、1か月おき、2週間おきなどの一定頻度で連続してスクレイピングを実施した。 ■ 価格情報に加え、提供地域や提供形態などの属性の情報も収集した。
収集情報を用いた分析	■ 定期収集によって作成したデータベースを用い、価格指数の試算を行った。本調査研究では、総務省(2019)²、Triplett(2004)³にならい試算を行った。 ■ 上記文献に基づく手法により外れ値の処理およびヘドニック・アプローチ⁴による分析を行い、指数を作成した。
文献調査および有識者ヒアリング	■ 上記文献に加え、国内外のシェアリング・エコノミーの産出額の実質化等についての最新の文献、報告書の調査を行い、参考になるとされる文献について整理した。 ■ 本調査研究の作業内容に関連する知見を広く有する有識者4名(一橋大学阿部教授、日本大学清水教授、東京大学渡辺教授、関西大学宇都宮教授(ヒアリング調査時順))へのヒアリングを実施した。主な観点として、①本調査における手順・手法について、②データの加工・分析について、の2点のヒアリングを行った。

1. ウェブスクレイピングとは、ウェブサイトにある情報を抽出するコンピュータソフトウェアの技術のことである。

2. 総務省(2019) 総務省物価指数研究会(第13回、資料番号2)「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」総務省物価指数研究会。

3. Triplett(2004) “Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products.” OECD Science, Technology and Industry Working Papers 2004/9, OECD Publishing.

4. ヘドニック・アプローチとは、財・サービスの価格の違いを、財・サービスの持つ多様な属性によって説明する分析であり、統計的には回帰分析の応用である。

Ⅱ．価格情報等の収集

- 2-1 収集対象先の抽出及び選定
- 2-2 収集した情報のデータベース作成
- 2-3 収集スケジュール

2-1 収集対象先の抽出及び選定

2-2 収集した情報のデータベース作成

- ・シェアリング・エコノミーに係るサービスについて分野別類型別に主要なプラットフォーム事業者を抽出し、民泊、駐車場スペースシェア、カーシェア、ライドシェアについて収集先として検討を行い、4事業者を収集対象先とした。
- ・収集先の選定にあたり、①シェアリング・エコノミーのサービス提供価格の情報が掲載されているか確認し、②それらのサイトに収集ツール(統計分析ソフトRもしくはPython)でアクセスが可能か検討を行った。
- ・収集情報のデータベース化の手順は以下の通り。

① リスティングの収集

収集する財・サービスの一覧(リスティング)の収集では、個別情報が掲載されているホームページのリンクを取得することを目的としている。多くのインターネットサイトでは、トップページ以降に都道府県別やサービス属性別の一覧リストを作成していることが多い。そこで、各都道府県順にアクセスし、一覧から個別情報のリンクを収集する。なお、民泊プラットフォームでは都道府県別に個別情報が一覧化されていなかったため、「東京」を検索して表示されたリスティングを分析対象とした。

② 個別情報の収集

個別情報の収集では、先に収集したリンク先を順番にアクセスし、必要な情報を抽出した。サイトのアドレス情報を用い、パネルデータ化を行った。

③ 前処理

前処理の段階では、分析可能な形式への変換を行っている。外れ値として、価格の平均から絶対値で3標準偏差分離れているものを外している。

④ データベースの完成

2-3 収集スケジュール

各種サイトの定期的な収集の実施については、以下に示すとおりである。駐車場スペースシェア、ライドシェア、カーシェアのプラットフォームでは安定的な収集ができた。民泊プラットフォームについては、画面仕様の変更などがたびたび発生し、あらかじめ日時を決めた収集ができなかったが、月内に1回は収集できるように対応している。

	収集開始	最終収集日	収集頻度	備考
民泊	2019年10月26日	2020年2月7日	月に一度	1回の収集では、収集日の翌々月全日分の価格収集を行っている。サイトリニューアル等、予期せぬ不具合が発生したため、厳密に収集日を設定することができなかった。詳細は「V. ウェブスクレイピングによる価格指数推計のための課題整理」を参照
駐車場スペースシェア	2019年10月1日	2020年2月1日	月に二度	各月1日収集のデータを分析に使用
ライドシェア	2019年10月1日	2020年2月15日	月に二度	各月1日と14・15日収集のデータを統合し、分析に使用
カーシェア	2019年9月30日	2020年2月21日	月に二度	各月第4金曜日収集のデータを分析に使用

Ⅲ. 収集情報を用いた分析

- 3-1 既存文献の整理
- 3-2 分析方法
- 3-3 分析結果

3-1 既存文献の整理 (1)

収集データを分析するにあたり、総務省物価指数研究会(2019)「宿泊料のウェブスクレイピングによる価格収集及び指数作成方法について」及びTriplett, J.(2004) “Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes”, を整理。

○ 総務省(2019)

- ・ 総務省物価統計室では、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するため、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収集および指数作成方法の検討を実施。
- ・ 消費者の予約の傾向として、宿泊日の1週間～3週間前、1か月前に予約するものが多いことをアンケートにより把握し、収集日の翌月、翌々月の全日価格を収集対象に選定している。
- ・ また、ウェブスクレイピングを用いる際には、サイト構造による個別の設定の必要があることを指摘し、同一サイト上にて複数の会社の販売価格を収集できる総合予約サイト等から収集することは実務上効率的としている。
- ・ 現行の総務省「小売物価統計調査」では、調査品目及び基本銘柄は、旅館を「和室・1泊2食付」、ホテルを「洋食・1泊朝食付」のプランを対象としていることより、同様のプランを対象に収集している。ただし、出張向けビジネスプランなど、消費者物価指数の対象とすべきではないプランや品質差があるプランも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収集した。なお、外れ値処理により、一般的な宿泊料金に対して極端に高い(または極端に安い)料金設定のプランは除外している。
- ・ 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点については、指数計算において、欠測値を無視した平均価格の計算では、曜日別の欠測に差があることで平均価格に偏りが生じることを指摘している。また、欠測値に対処するための手法として、回帰補完を行っている。
- ・ 指数作成にあたっては、幾何平均による価格比を算出し、それを前月の指数 I_{t-1} に乘じることで価格指数 I_t を算出している。

3-1 既存文献の整理 (2)

○ Triplett (2004)

- Triplett (2004) では物価を計測する財の品質の変化が起こるケースで、ヘドニック・アプローチを紹介している。このアプローチは、IT機器の物価を中心に各国で利用され、価格を財の品質を代理する属性に回帰し、月次や年次といった時間の効果をダミー変数で表す、ダミー変数法が用いられている。
- 例えば、家庭用パソコン i の t 期の価格を p_{it} とし、3期間 ($t = \tau, \tau + 1, \tau + 2$) の価格指数を算出することを考える。パソコン価格が速度性能と容量の二つの属性で決まると仮定すれば、ヘドニック回帰分析は以下のように定式化される。

- $\log(p_{it}) = a_0 + a_1 \log(\text{speed}_{it}) + a_2 \log(\text{memory}_{it}) + b_1(D_t^{\tau+1}) + b_2(D_t^{\tau+2}) + \varepsilon_{it}$
 speed_i はPCの速度性能、 memory_i はPCの容量を表し、 $D_t^{\tau+j}$ は $t = \tau + j$ の時に1を取るダミー変数である。この時、時間ダミーの係数 b_j が時点間の価格変化の効果を表す。

- なお、プールしたOLS推定では価格変化を表す係数が固定されてしまうため、2期間ずつのサンプリングで推定する“Adjacent-period dummy approach”(隣接2時点アプローチ)が推奨されている。

この場合、 τ 期、 $\tau + 1$ 期は、

$$\log(p_{it}) = a'_0 + a'_1 \log(\text{speed}_{it}) + a'_2 \log(\text{memory}_{it}) + b'_1(D_t^{\tau+1}) + \varepsilon_{it} \quad (t = \tau, \tau + 1)$$

$\tau + 1$ 期、 $\tau + 2$ 期は、

$$\log(p_{it}) = a''_0 + a''_1 \log(\text{speed}_{it}) + a''_2 \log(\text{memory}_{it}) + b''_2(D_t^{\tau+2}) + \varepsilon_{it} \quad (t = \tau + 1, \tau + 2)$$

によって推定結果が得られる。

3-2 分析方法

収集先によってデータ構造が異なるが、以下のような観点・方法に則り分析を行った。

データセットの構造	■ 民泊プラットフォームでは、日次単位で価格が掲載されており、すでに取引が成立している場合、価格情報は欠測値となる。駐車場シェアプラットフォームも同様に日次単位での掲載がされており、満車や利用不可の場合は欠測値となる。これら2つのサイトから構築されるデータベースは、日次パネルデータの構造となる。 ■ カーシェアプラットフォームでは、アクセスした日の当月と翌月の利用可否が掲載されている。また価格は、アクセス日の当月と翌月について一律に設定されている。そこで、毎月アクセスした時点の価格をその月の価格とみなした。そのため、データは月次レベルのパネルデータとなる。 ■ ライドシェアプラットフォームでは、同一サービスが継続的に提供されることは原則想定されていないと考えられるため、データはクロスセクションの構造となる。
外れ値の処理	■ 総務省(2019)では、個別価格と平均価格との差が標準偏差の絶対値を3倍以上超えるような場合を外れ値として認定し、サンプリングから排除している。本調査研究では同資料の方法に則り、外れ値の検出・除外を行った。
欠測値の処理	■ 総務省(2019)では、欠測データへの対策として回帰補完を行っている。本調査研究では欠測値の補完を行わず、分析する方針をとった。
分析モデル	■ 本分析では、ヘドニック・アプローチを採用し指数の作成を行った。シェアリング・エコノミーで提供される個々のサービスは相互に異質性が高く、属性をコントロールし月次の価格水準を算出することが有効と考えられる。 ■ なお、パネルデータの形状のデータベースについては、ランダム効果モデル⁵と固定効果モデル⁶での分析も行う。固定効果モデルについては、1つ1つにダミー変数を作成し、時間にかかわらず一定の異質性をコントロールすることで、リピート・セールス法⁷と同等の分析を実施したものと解釈される。
サンプル期間	■ 分析に用いるサンプルは全期間を用いたものとサンプル期間を2ヶ月おきに分けたものを用いている。隣接2期間をサンプルとする分析は、シェアリング・エコノミーに係るサービスについて、属性のパラメータが時間を通じて変化するか否か不明であるため、比較のために分析を行った。

5. ランダム効果モデルとは、属性を表す説明変数とは相関しない、個別の財・サービス固有の効果の存在を仮定したモデルである。

6. 固定効果モデルとは、属性を表す説明変数と相関のある、個別の財・サービス固有の効果の存在を仮定したモデルである。

7. リピート・セールス法とは、複数回売買される財・サービスについて、異時点の取引価格を時間ダミー変数で回帰させることで推計する方法である。

3-3 分析結果：民泊①

データセットの構造

対象とした民泊プラットフォームの価格は、個別のホームページ上で、人数と日付を選択することで表示される。本調査研究では、大人1名が1泊するという想定の下で表示される価格を収集している。

民泊プラットフォームに対しては、収集日より2ヶ月先の月の価格を利用している。例えば、10月に収集を行う場合は、12月1日から12月31日までの価格を収集している。本調査研究では10月から2月に収集を行い、12月から4月までの価格を収集している。

収集した価格のデータセットは下記のような構造(数値は例)になっている。month列は月次ダミーとして扱われる。

下図の例では、IDが1の物件の12月2日の価格は欠測であるが、これはオーナーが貸し出さない時やすでに予約が入っている場合を表す。したがって、この物件の12月2日の実際の取引価格は観察・集計できず、それ以外に表示されている価格のみが集計対象となる。

民泊のデータセットのイメージ

ID	date	month	price	...
1	12月1日	12月	3000	...
1	12月2日	12月	N.A.	...
1	12月3日	12月	3200	...
...	...	...	...	...
2	12月1日	12月	2000	...
2	12月2日	12月	2100	...
...	...	...	...	...
1	1月1日	1月	4000	...
...	...	...	...	...

3-3 分析結果：民泊②

- 本分析に用いる推計式は以下の通りである。
- まず、収集した全期間のサンプルを一括した、ヘドニック・アプローチによるモデルは以下のように定義される。

$$\ln P_{it} = \alpha + \sum_{j=1}^n \beta_j x_{ijt} + \sum_{k=201912}^{202004} \delta_k d_k + u_{it}$$

P_{it} は部屋 i の日次 t の価格を表す。 x_{ijt} は、部屋 i の日次 t における属性 j を表す。また、 d_k は k 時点(年×100+月で表示)における月次のダミー変数であり、日次 t が k 時点に含まれる時1を取るダミー変数である。なお、価格指数作成のため、12月を基準とし、1月から4月の月次ダミー変数が作成される。なお、 α は定数項、 β_j は属性 j の係数、 δ_k は k 時点の係数を表す。 u_{it} は誤差項である。

- 次に、収集したサンプルを2ヶ月ごとに分けた、隣接2時点でのモデルは以下の通りとなる。

$$\ln P_{it} = \alpha + \sum_{j=1}^n \beta_j x_{ijt} + \delta_T d_T + u_{it}$$

例えば、12月～1月をサンプルの期間とする場合、 d_T は1月のデータであるとき、1を取る変数である。

- 本分析では、アンバランスなパネルデータを考慮し、すべてのデータをプールした最小二乗法(Pooling OLS)と、ランダム効果を考慮したランダム効果モデル(Random Effect)、固定効果を取った固定効果モデル(Fixed Effect)の3種類の分析を行っている。

3-3 分析結果：民泊③

本分析では t 期の価格指数 I_t は以下のように算出される。 $\hat{\delta}_t$ は分析から得られた時点 t の時間ダミーの係数の推定値である。なお、今回の分析で前期の価格指数が得られていない場合については、前期を100として算出する。

$$I_t = \exp(\hat{\delta}_t) \times I_{t-1}$$

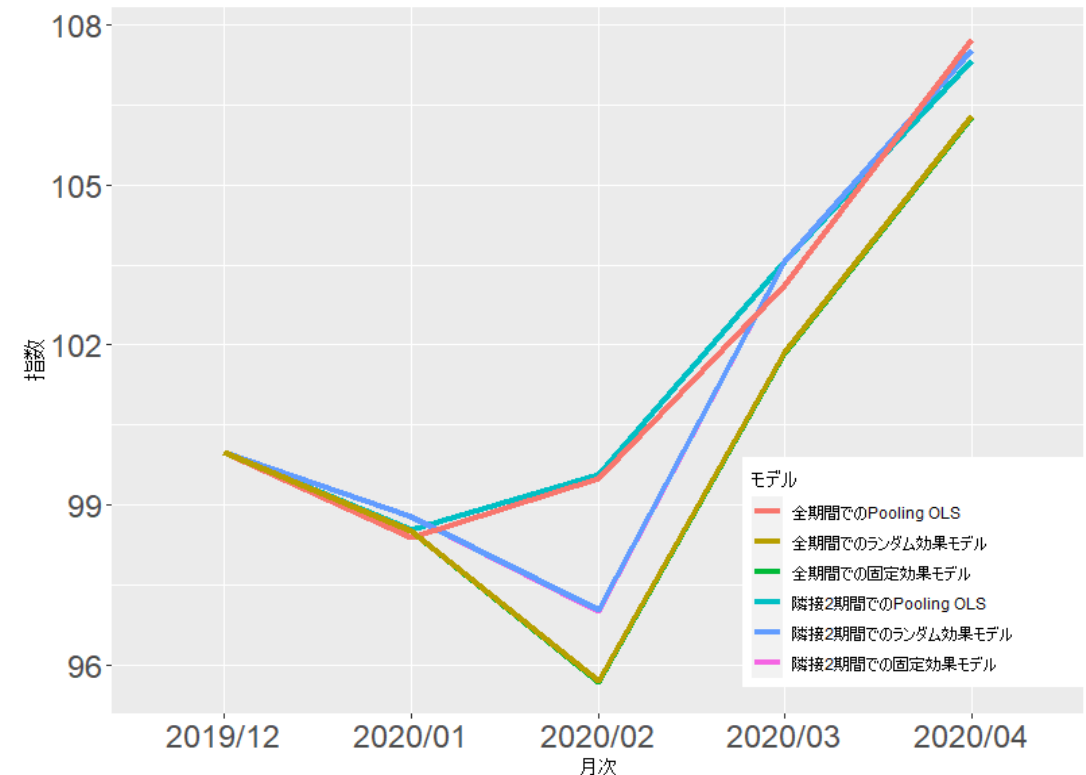
以下は月次指数の推計結果を示す。月次ダミーの有意性について、OLSによる推定では、2月は全期間・隣接2期間ともに有意な結果とはならなかった。指数として、OLSとランダム効果・固定効果モデルと大きく乖離していることもわかる。

サンプル期間	(1)OLS	(2)ランダム効果	(3)固定効果	(4)OLS	(5)ランダム効果	(6)固定効果
2019/12 ~ 2020/4						
1月ダミー	-0.0162***	-0.0151***	-0.0150***	-0.0147***	-0.0123***	-0.0122***
2月ダミー	-0.00494	-0.0440***	-0.0441***			
3月ダミー	0.0308***	0.0184***	0.0183***			
4月ダミー	0.0744***	0.0610***	0.0609***			

サンプル期間	(7)OLS	(8)ランダム効果	(9)固定効果	(10)OLS	(11)ランダム効果	(12)固定効果
2020/1 ~ 2020/2						
1月ダミー						
2月ダミー	0.0105	-0.0179***	-0.0181***			
3月ダミー				0.0392***	0.0652***	0.0654***
4月ダミー						

サンプル期間	(13)OLS	(14)ランダム効果	(15)固定効果
2020/3 ~ 2020/4			
1月ダミー			
2月ダミー			
3月ダミー			
4月ダミー	0.0359***	0.0378***	0.0376***

(注) 主要な変数の係数のみ表示。標準誤差は省略している。*** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$



3-3 分析結果：民泊④

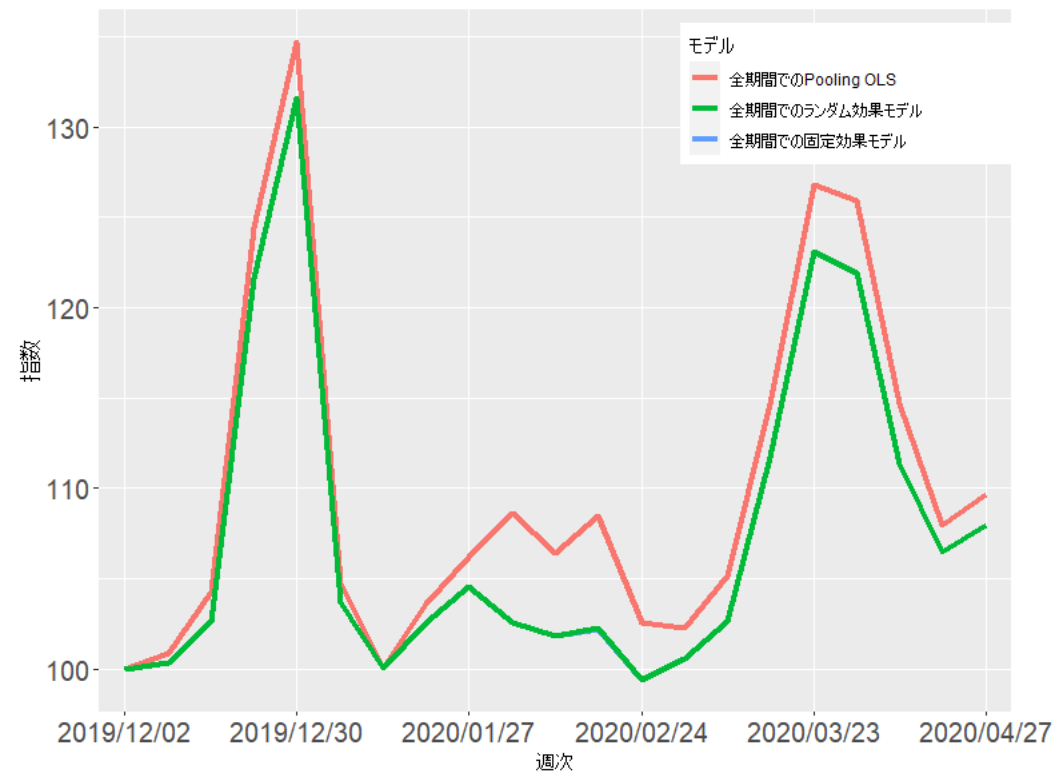
週次指数の作成

民泊プラットフォームのデータでは、週次ダミーを作成し、週次指数の推移も確認している。期間は12月2日始まりの週を基準とし、以降4月20日始まりの週までを対象として、ダミーを設定している。推定結果は以下に示す通りである。

Pooling OLSによる推定結果ではランダム効果を考慮したモデルや固定効果モデルよりも指数が上振れることがわかる。また、1月の後半から2月にかけてはこうした乖離が大きいことがわかる。

	(1)OLS	(2)ランダム効果	(3)固定効果
2019/12/09ダミー	0.00875	0.00373	0.00372
2019/12/16ダミー	0.0418***	0.0260***	0.0260***
2019/12/23ダミー	0.219***	0.196***	0.196***
2019/12/30ダミー	0.298***	0.275***	0.275***
2020/01/06ダミー	0.0467***	0.0362***	0.0362***
2020/01/13ダミー	0.000532	0.000485	0.000523
2020/01/20ダミー	0.0357***	0.0255***	0.0255***
2020/01/27ダミー	0.0601***	0.0448***	0.0448***
2020/02/03ダミー	0.0827***	0.0252***	0.0250***
2020/02/10ダミー	0.0623***	0.0184***	0.0182***
2020/02/17ダミー	0.0811***	0.0220***	0.0219***
2020/02/24ダミー	0.0253*	-0.00591	-0.00600
2020/03/02ダミー	0.0227**	0.00595	0.00590
2020/03/09ダミー	0.0501***	0.0264***	0.0263***
2020/03/16ダミー	0.137***	0.111***	0.111***
2020/03/23ダミー	0.237***	0.208***	0.208***
2020/03/30ダミー	0.230***	0.198***	0.198***
2020/04/06ダミー	0.137***	0.108***	0.108***
2020/04/13ダミー	0.0769***	0.0630***	0.0629***
2020/04/20ダミー	0.0925***	0.0764***	0.0763***

(注) 主要な変数の係数のみ表示。標準誤差は省略している。*** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$



(注) ランダム効果モデルと固定効果モデルの係数の推定値は近い値を取っているため、グラフ上は重なってしまっている。

3-3 分析結果: 駐車場①

データセットの構造

駐車場シェアのサイトにおいて価格を収集する場合、個別の駐車場を表示させると、各駐車場のページには、当日の価格と当日から少なくとも7日先までの利用可否と価格情報が掲載されている。

表示される価格について、駐車場ごとに掲載されている日数が異なることが分かった。今回の分析では収集日から7日間の価格データを分析の対象としている。

駐車場シェアのデータセットのイメージ

ID	date	month	price	...
1	11月1日	11月	1000	...
1	11月2日	11月	N.A.	...
1	11月3日	11月	1000	...
1	11月4日	11月	1000	...
1	11月5日	11月	1100	...
1	11月6日	11月	1100	...
1	11月7日	11月	1000	...
2	11月1日	11月	2000	...
...	...	...	...	...
2	11月7日	11月	2300	...
...	...	...	...	...

3-3 分析結果：駐車場②

- 駐車場シェアにおいて、分析に用いるモデル式は以下の通りである。
- まず、収集した全期間のサンプルを一括した、ヘドニック・アプローチによるモデルは以下のように定義される。

$$\ln P_{it} = \alpha + \sum_{j=1}^n \beta_j x_{ijt} + \sum_{k=201911}^{202002} \delta_k d_k + u_{it}$$

P_{it} は駐車場 i の日次 t の価格、 x_{ijt} は、駐車場 i の日次 t における属性 j を表す。また、 d_k は k 月における月次ダミーであり、日次 t が k 月に含まれる時1を取るダミー変数である。なお、価格指数作成のため、10月を基準とし、11月から2月の月次ダミー変数が作成される。なお、 α は定数項、 β_j は属性 j の係数、 δ_k は k 時点の係数を表す。 u_{it} は誤差項である。

- 次に、収集したサンプルを2ヶ月ごとに分けた、隣接2時点でのモデルは以下の通りとなる。

$$\ln P_{it} = \alpha + \sum_{j=1}^n \beta_j x_{ijt} + \delta_T d_T + u_{it}$$

例えば、12月～1月をサンプルの期間とする場合、 d_T は1月のデータであるとき、1を取る変数である。

- 分析では、アンバランスなパネルデータを考慮し、すべてのデータをプールした最小二乗法 (Pooling OLS) と、ランダム効果を考慮したランダム効果モデル (Random Effect)、固定効果を取った固定効果モデル (Fixed Effect) の3種類の分析を行っている。

3-3 分析結果：駐車場③

推計結果

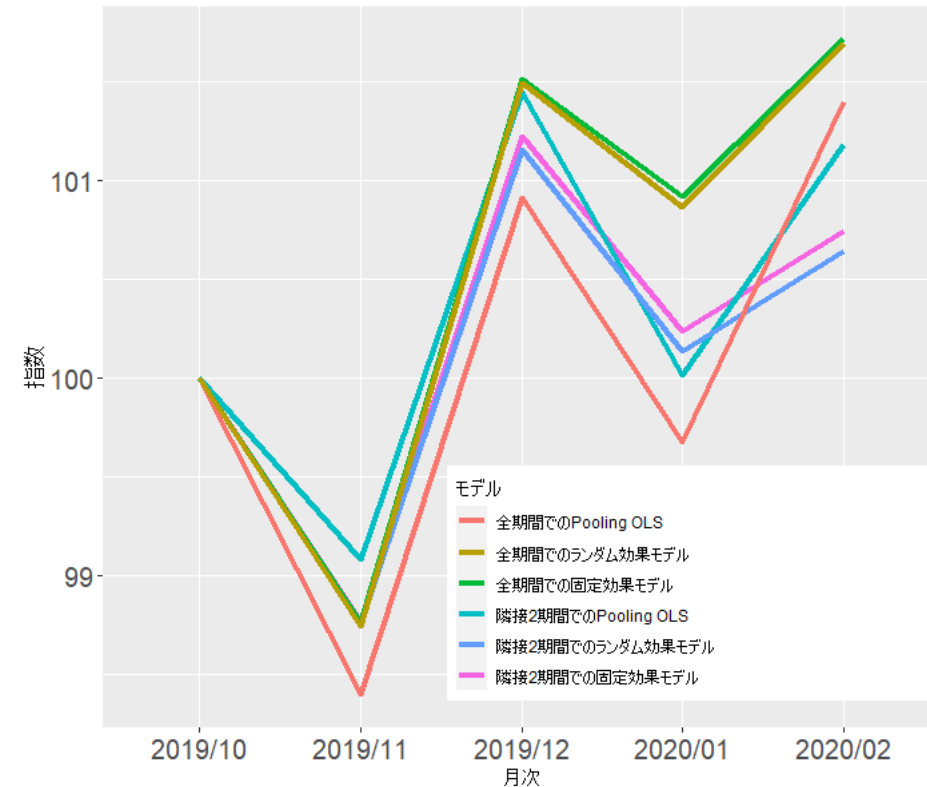
月次ダミーについて、OLSでは、全期間では有意な差は見られず、隣接2期間のOLSでは12月のみ有意となった。一方、全期間を対象とするランダム効果・固定効果モデルでは、4つのダミー変数がすべて有意となった。また、隣接2期間をサンプル期間とするランダム効果・固定効果モデルでも2月を除いて有意となった。

民泊プラットフォームと比較して指数の推移幅は小さいものの、いずれのモデルでも同じような傾向を持って推移していることがわかる。

期間	(1)OLS 2019/10 ~ 2020/2 (全期間)	(2)ランダム効果 2019/10 ~ 2020/2 (全期間)	(3)固定効果 2019/10 ~ 2020/2 (全期間)	(4)OLS 2019/10 ~ 2019/11	(5)ランダム効果 2019/10 ~ 2019/11	(6)固定効果 2019/10 ~ 2019/11
11月ダミー	-0.0162	-0.0126***	-0.0124***	-0.00924	-0.0125***	-0.0123***
12月ダミー	0.00908	0.0148***	0.0150***			
1月ダミー	-0.00328	0.00864*	0.00916*			
2月ダミー	0.0139	0.0168***	0.0170***			

期間	(7)OLS 2019/11 ~ 2019/12	(8)ランダム効果 2019/11 ~ 2019/12	(9)固定効果 2019/11 ~ 2019/12	(10)OLS 2019/12 ~ 2020/1	(11)ランダム効果 2019/12 ~ 2020/1	(12)固定効果 2019/12 ~ 2020/1
11月ダミー						
12月ダミー	0.0236**	0.0239***	0.0245***			
1月ダミー				-0.0142	-0.0101**	-0.00982**
2月ダミー						

期間	(13)OLS 2020/1 ~ 2020/2	(14)ランダム効果 2020/1 ~ 2020/2	(15)固定効果 2020/1 ~ 2020/2
11月ダミー			
12月ダミー			
1月ダミー			
2月ダミー	0.0117	0.00503	0.00506



3-3 分析結果:ライドシェア①

データセットの構造

ライドシェアの価格は、収集時点で募集しているライドシェアの価格が掲載されている。ドライバーが指定した日時、指定した場所間の移動に対する1人あたり金額である。本サイトの特徴として、定員に達していても募集金額が表示されており、価格の欠測は発生しない。一方、一部では0円でのオファーもある。

ライドシェアプラットフォームのデータセットのイメージ

	ID	day	Price		ID	day	Price
	1	10月1日	500	→	1	10月1日	500
10月収集時点→	2	11月4日	1000	→	2	11月4日	1000
11月収集時点→	2	11月4日	1000		...	...	...
	...	...	...				

推計式

収集した全期間のサンプルを一括した、ヘドニック・アプローチによるモデルは以下のように定義される。

$$\ln P_i = \alpha + \sum_{j=1}^n \beta_j x_{ij} + \sum_{k=201911}^{202002} \delta_k d_k + u_i$$

P_i はライドシェア i の価格を表す。 x_{ij} は、ライドシェア i の属性 j を表す。また、 d_k は k 月における月次ダミーであり、

ライドシェア i の出発日が k 月に含まれる時1を取るダミー変数である。なお、価格指数作成のため、10月を基準とし、11月から2月の月次ダミー変数が作成される。なお、 α は定数項、 β_j は属性 j の係数、 δ_k は k 時点の係数を表す。 u_{it} は誤差項である。

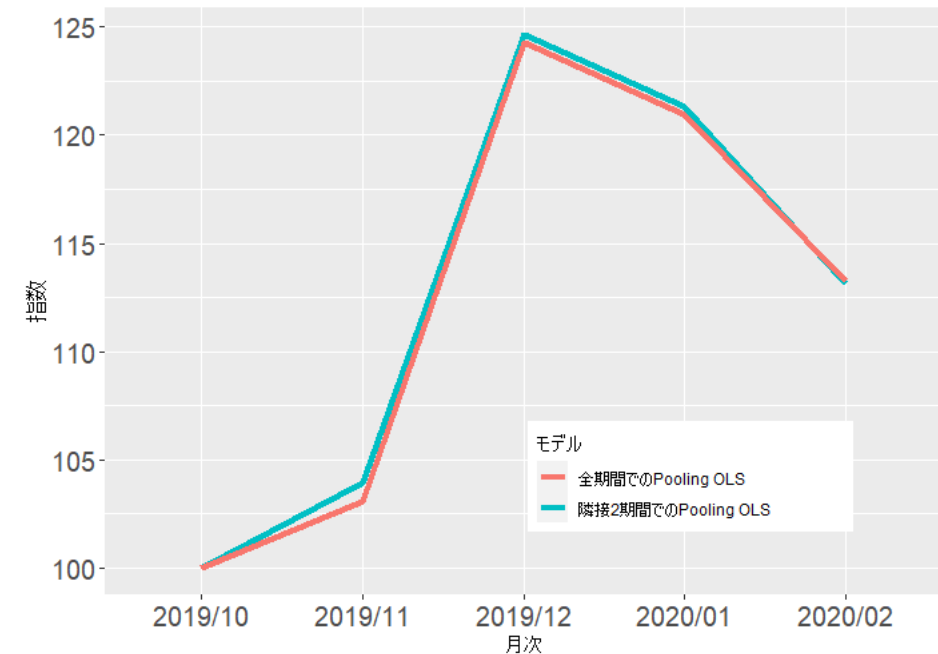
収集したサンプルを2ヶ月ごとに分けた、隣接2時点でのモデルでも分析を行った。

3-3 分析結果:ライドシェア②

- 全期間をサンプルとした場合、11月は10月と比較して係数値は有意な結果ではなかったものの、12月、1月、2月はいずれも有意となっている。ただし、隣接2期間でサンプルを区分けした場合、12月のみ有意となった。
- 指数をみると10月・11月と比べ、12月・1月は大きい値を取っている。また、民泊と比較してもこの値は大きく出ている。

期間	(1)OLS 2019/10 ~ 2020/02 (全期間)	(2)OLS 2019/10 ~ 2019/11	(3)OLS 2019/11 ~ 2019/12
11月ダミー	0.0303	0.0389	
12月ダミー	0.217***		0.182***
1月ダミー	0.190**		
2月ダミー	0.125*		

期間	(4)OLS 2019/12 ~ 2020/01	(5)OLS 2020/01 ~ 2020/02
11月ダミー		
12月ダミー		
1月ダミー	-0.0269	
2月ダミー		-0.0700



3-3 分析結果:カーシェア①

- カーシェアの個別ページでは、アクセス日の当月と翌月の利用可能な日付が表示される。
- 価格情報は、アクセス時点での平日・休日の価格が収集可能である。収集日時点の平日・休日それぞれの価格をその月の価格として収集している。

カーシェアプラットフォームのデータセットのイメージ

ID	day	month	Price
1	10月平日	10月	2000
1	10月休日	10月	2500
...	...	...	...
10	10月平日	10月	1000
...	...	...	...
1	11月平日	11月	1000
1	11月休日	11月	1000
...	...	...	...

推計式

収集した全期間のサンプルを一括した、ヘドニック・アプローチによるモデルは以下のように定義される。

$$\ln P_{it} = \alpha + \sum_{j=1}^n \beta_j x_{ijt} + \sum_{k=201910}^{202002} \delta_k d_k + u_{it}$$

P_{it} は自動車 i の収集時点 t での価格を表す。 x_{ijt} は、自動車 i の t 時点における属性 j を表す。また、 d_k は k 時点における月次ダミーであり、サンプルの収集時点 t が k 時点に含まれる時1を取るダミー変数である。なお、価格指数作成のため、9月を基準とし、10月から2月の月次ダミー変数が作成される。なお、 α は定数項、 β_j は属性 j の係数、 δ_k は k 時点の係数を表す。 u_{it} は誤差項である。

民泊同様、収集したサンプルを2ヶ月ごとに分けた、隣接2時点でのサンプリング分けを行い、アンバランスなパネルデータを考慮したOLS (Pooling OLS) と、ランダム効果を考慮したランダム効果モデル (Random Effect)、固定効果を取った固定効果モデル (Fixed Effect) の3種類の分析を行っている。

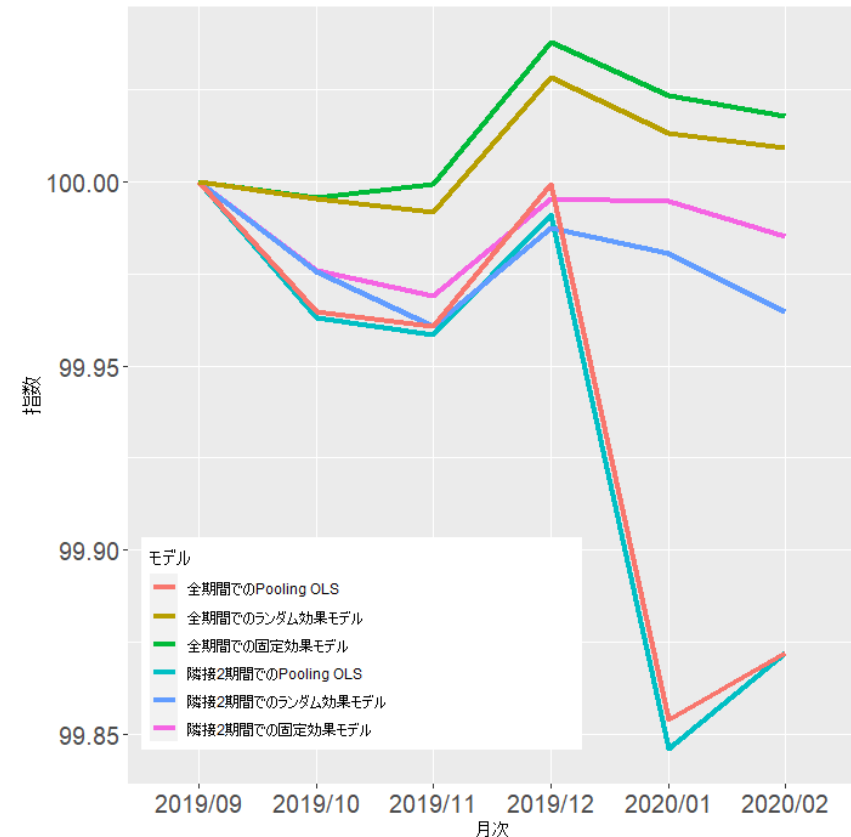
3-3 分析結果:カーシェア②

- 月次ダミーについて、10月ダミーは隣接2期間ランダム効果・固定効果モデルで有意、12月はサンプルの期間に関わらず、ランダム効果・固定効果モデルで有意、1月ダミーはサンプルの期間に関わらずOLSで有意、2月ダミーは全期間をサンプルとするOLSでのみ有意であった。モデルによって有意性にばらつきがみられた。
- 指数の取りうる範囲は、民泊プラットフォームの結果と比べてかなり狭いものである。

期間	(1) OLS 2019/9 ~ 2020/2 (全期間)	(2) ランダム効果 2019/9 ~ 2020/2 (全期間)	(3) 固定効果 2019/9 ~ 2020/2 (全期間)	(4) OLS 2019/9 ~ 2019/10	(5) ランダム効果 2019/9 ~ 2019/10	(6) 固定効果 2019/9 ~ 2019/10
10月ダミー	-0.000352	-4.55e-05	-0.0000442	-0.000370	-0.000243**	-0.000239*
11月ダミー	-0.000393	-8.11e-05	-0.0000604			
12月ダミー	-4.73e-06	0.000284*	0.000380**			
1月ダミー	-0.00146**	0.000132	0.000248			
2月ダミー	-0.00128*	9.22e-05	0.00018			

期間	(7) OLS 2019/10 ~ 2019/11	(8) ランダム効果 2019/10 ~ 2019/11	(9) 固定効果 2019/10 ~ 2019/11	(10) OLS 2019/11 ~ 2019/12	(11) ランダム効果 2019/11 ~ 2019/12	(12) 固定効果 2019/11 ~ 2019/12
10月ダミー						
11月ダミー	-4.51e-05	-0.000149	-7.16e-05			
12月ダミー				0.000326	0.000266**	0.000263**
1月ダミー						
2月ダミー						

期間	(13) OLS 2019/12 ~ 2020/1	(14) ランダム効果 2019/12 ~ 2020/1	(15) 固定効果 2019/12 ~ 2020/1	(16) OLS 2020/1 ~ 2020/2	(17) ランダム効果 2020/1 ~ 2020/2	(18) 固定効果 2020/1 ~ 2020/2
10月ダミー						
11月ダミー						
12月ダミー						
1月ダミー	-0.00145**	-6.91e-05	-4.69e-06			
2月ダミー				0.000262	-0.000160	-9.48e-05



IV. 文献調査および有識者ヒアリング

- 4-1 文献調査
- 4-2 有識者ヒアリング

4-1 文献調査

- 国内外のシェアリング・エコノミーの産出額の実質化に関して、最新の文献研究、特に国際機関や、シェアリング・エコノミーに係る先進的な取組が進められている先進各国における報告書や先行研究等のサーベイを行った。

文献	概要
OECD (2017)	■ 2017年夏にOECD加盟国間で実施された価格指数に対するデジタル経済の影響の測定に関するIMF / OECDアンケートの結果を整理している。アンケートの中でシェアリング・エコノミーの活動の記録の有無、価格変化の把握方法を確認している。 ■ 結果として、AirbnbやUberなどのシェアリング・エコノミーの経済活動はほとんどのOECD諸国の消費者物価指数には含まれていないことが分かった。また、オーストラリア、ドイツ、アメリカにおいてのみ、シェアリング・エコノミーに関する経済活動の一部を消費者物価指数内で記録していることが分かった。
Dolft and Konijn (2018)	■ オランダの統計局やデンマークの統計局がどのようにAirbnbを扱っているかを説明している。オランダの統計局ではAirbnbを宿泊サービスとして扱い、所有者が居住する住宅サービスがその増分だけ差し引かれている。一方、デンマークの統計局ではAirbnbは住宅サービスとして扱い、テナントとして一部を貸し出すものとみなしている。 ■ 価格指数について、オランダの統計局ではAirbnbの宿泊施設の料金に関して、ホテルの宿泊料の価格指数を代理指標として用いている。この方法はAirbnbとホテルの宿泊料が類似の季節性を示せば合理的と考えられるが、Airbnbが様々な種類の短期宿泊施設を対象としていることから、別荘なども含めた宿泊施設をより幅広くカバーした指標が必要だと指摘している。またホテルを代理指標として用いることが長期的にも適切といえるのか、決定的な証拠が不足していることも指摘している。 ■ 宿泊サービスとAirbnbに支払われるサービス料の変化を算出するためには、Airbnbに支払われるサービス料が、基礎的な宿泊サービスの価値と一定の比率の積によって算出されることを考慮する必要がある。価格と取引量の計算にあたっては、Eurostatが2016年に公表したハンドブックの手法に従うことができるとしている。
IMF (2018)	■ シェアリング・エコノミーによる低価格の代替製品の登場が生活費へ与える影響を推定する必要性を指摘すると共に、Uberのように、従来のサービスに対して非常に近い代替品とみなせる製品の場合には、消費者物価指数にサービスで支払われる平均的な価格変化を反映させることが合理的であるとしている。

4-2 有識者ヒアリング

- 価格収集のあり方や指数作成の質を高めるため、作業内容に関連する知見を広く有する有識者4名（一橋大学阿部教授、日本大学清水教授、東京大学渡辺教授、関西大学宇都宮教授（ヒアリング調査時点順））へヒアリングを実施した。
- 主な論点として、①本調査における手順・手法、②データの加工・分析、の2点のヒアリングを行った。①ではシェアリング・エコノミーという経済活動の価格指数作成にあたり考慮する点やどのような価格収集が望ましいか、本調査研究で行っている手法の妥当性はどうか、②では、データを収集、分析した結果を基に、どのような改善が可能であるか、どのような点に注意して解釈を行うべきであるか、といった論点での示唆を得た。

（主なコメント）

- 分析手法として、ヘドニック・アプローチは妥当ではあるものの、リピート・セールス法による比較も必要である。
ただし、今回の分析では固定効果モデルを作成しており、解釈としては同じものとみなすことができる。
- 実務的には、代表的な財・サービスを決め、その価格の変動と全体の分布の推移を追うことが重要である。
- ヘドニック・アプローチはモデルの安定性、係数の選択が重要。ダミー変数で整理すべきものは何か検討が必要。記述統計を確認し、全サンプルの大部分が有している属性は、分析モデルから除いて構わないだろう。
- 表示価格を集計しており、取引価格が観察できない（欠測となる）点には注意が必要である。サイト上で公開されている価格は売買が成立する前の価格だが、価格が日々変動する可能性があるため、実際に売買が成立した時点の価格と一致するか不明である。また、売買が成立したサービスは実際どのくらいの金額で売買が成立したかは観察できない。
このため、高頻度で日々収集し、取引成立直前の表示価格を取引価格とみなして分析を行うことを検討してみてはどうか。
また、欠測に意味があるような場合には、ヘドニック・アプローチでトービットモデルを用いた分析を行うことも検討するとよい。

V. ウェブスクレイピングによる価格指数推計のための課題整理

- 5-1 本調査研究で得られた結果
- 5-2 データ収集に関する課題
- 5-3 分析に関する課題

5-1 本調査研究で得られた結果

①政府統計部局におけるウェブスクレイピングによるデータ収集の実現可能性

- 政府統計部局において、ウェブスクレイピングを用いてデータを集計し、統計作成を行うというアイディアには実現可能性があると考えられる。
- 他方、属性を含めた大規模なデータをリアルタイムで収集ができるというメリットと、情報収集にはサイトへの負担を伴うことから、事業者の理解・協力が不可欠なこと、また表示価格と取引価格の違いや取引数量の把握の精度が安定しないといったデメリットの両面を踏まえ、実施を検討する必要がある。

②シェアリング・エコノミーにおける価格の実態

- 分散が大きく、分野によっては単峰の形状とはならない分布がみられる。
- ヘドニック・モデルの当てはまりは分野を問わず比較的良好であり、シェアリング・エコノミーに係るサービスの価格がその属性によって決定されている面があると考えられる。
- ヘドニック・モデルをベースに価格指数を作成すると、分野によって月次の変動が（例えば消費者物価指数の類似サービスと比較し）非常に大きくなる。
推計モデル間でも変動幅には乖離があり、やや不安定な結果がみられた。

5-2 データ収集に関する課題

①全件数が不明の場合のリスティングの収集

一部のプラットフォームでは全件数が表示されないため、検索を繰り返して収集するしか方法がない。この時、全件把握できたことを確認する手段がないほか、リスティング収集と価格収集に若干のラグが発生することから、新規及び退出したケースがサンプルから欠落する。

②欠測が多く発生するケースへの対処

一部のプラットフォームでは、月によって欠測の発生率が大きく異なる。欠測が多く発生した原因の一つは、需要ないし供給の季節性にあると考えられる。需要側の要因に対応するためには、収集時期を前倒しする必要がある。そのためには、ある程度の期間データ収集を計測し、どのような時期に欠測が多く発生しているかを把握する必要がある。

③突然の仕様変更への対応・コスト

サイトの構造が頻繁に変わる場合、その都度収集プログラムを修正し対応する必要がある。画面デザインについて、改修もさることながら、例えばA/Bテスト⁸の対象となった場合はランダムに収集できなくなる可能性がある。

④収集運用の管理

スクレイピングにおけるクラウドサーバーへのコスト、プログラムの維持管理コストを考慮しつつ、負担感のない収集方法を検討すべきと考える。

(注) ⁸ A/Bテストとは、ウェブサイトの改善等を図るため2種類のサイト画面を用意し、ランダムに異なるデザインをユーザーへ閲覧させ、どちらの画面イメージが望ましいか検証する方法のことである。

5-3 分析に関する課題

①モデルで用いる変数の選択

- 時間ダミー変数が有意とはならない場合があり、シェアリング・エコノミーに係るサービスについて、需給の動向に応じて価格が推移している状況といえるのか検討が必要である。
- 記述統計からベースモデルを検討し、どの属性情報をダミー変数としてモデルに含めるのか検討が必要である。

②類似するサービスの価格指数との比較

- 民泊サービス価格については、家賃・住宅価格の延長とみなすか、宿泊料の延長とみなすかの整理が必要である。
- データ取得が短期間であり、季節性の影響を考慮できていない。類似のサービス指数を参考に、季節性の考慮について検討する必要がある。
- 類似指数を説明する変数を参考にモデルの構築を考える必要がある。例えば、住宅価格の場合、駅前からの距離、沿線情報、都心までの距離などが主要な価格決定要因と考えられ、今回は収集対象外であるそうした情報を分析に用いる可能性を検討する。

③欠測値の発生するケースの検証

- いつリストアップされているのか、いつ取引が成立しているのかを把握することが重要である。
- 市場で予約可能な状態となってから、どのように価格が推移するのか確認する。