PHƯƠNG PHÁP THAY THẾ GIÁ ĐƯỢC TRÍCH XUẤT TỪ TRANG WEB

Matthew Mayhew

Tóm tắt:

Mất giá là một vấn đề của chỉ tiêu giá, các chỉ tiêu được tính toán từ nguồn dữ liệu giá thu thập nhờ công cụ trích xuất dữ liệu (Web scraper), vì vậy tìm ra cách giải quyết hiệu quả vấn đề là một điều cần thiết. Imputation là một phương pháp giúp khắc phục tình trạng mất giá, mặc dù có nhiều kỹ thuật khác nhau được lựa chọn. Một nghiên cứu cho thấy việc tiến hành chọn giá trị thay thế giá hiện hành là phương pháp tốt nhất nhằm tối thiểu hóa sai số. Có hai tác động của phương pháp thay thế giá đối với chỉ số giá GEKSJ được tính toán từ giá mặt hàng thu thập thông qua công cụ Web scrapper, đó là khác biệt nhỏ trong chỉ số và giảm sự biến động bất thường do tác động của việc mất giá.

1. Giới thiệu

Cơ quan Thống kê Anh (ONS) đã tiến hành thử nghiệm tính toán các chỉ số giá dựa trên thông tin về giá hàng được thu thập từ các trang web bán hàng bằng công cụ Web scraper mỗi ngày, các chỉ số được tính toán đều đặn hơn phương pháp tính chỉ số giá tiêu dùng (CPI) truyền thống. Một số loại giá không thể thu thập, nguyên nhân là do sản phẩm bị hết hàng, hoặc do không thể trích xuất dữ liệu giá mặt hàng đó, tương tự với trường hợp bất thường trong tính toán CPI truyền thống. Những giá bị mất gây ảnh hưởng tới các chỉ số vì việc tính toán các các chỉ số này không còn đúng. Có hai cách giải quyết vấn đề trên, một là bỏ mặt hàng bị mất giá khỏi mẫu điều tra kể cả với những ngày có dữ liệu giá để tính toán chỉ tiêu, cách này được hiểu là việc làm phù hợp mẫu, hoặc cách khác là thay thế giá bị mất. Bài viết này tập trung vào phương pháp thay thế giá bị mất (Imputation) để giải quyết vấn đề mất giá, đồng thời khai thác nhiều phương pháp thay thế khác nhau, đánh giá ảnh hưởng của phương pháp thay thế đến chỉ số giá và đưa ra các khuyến nghị.

2. Các phương pháp thay thế giá (Imputation methods)

Có nhiều phương pháp thay thế giá khác nhau, tuy nhiên trong số đó chỉ có 3 phương pháp đã được kiểm định, là:

(1) Thay giá hiện hành bằng giá ở thời điểm liền trước thời điểm hiện hành

$$\hat{p}_i^t = p^{t-1}$$

(2) Phân loại giá trị trung bình theo cửa hàng hoặc theo loại mặt hàng, sử dụng:

(a)Trung bình cộng

$$\hat{p}_i^t = \frac{1}{n-1} \sum_{\substack{j \in C \\ j \neq i}} p_j^t$$

(b) Trung bình nhân

$$\hat{p}_i^t = \left(\prod_{\substack{j \in C \\ j \neq i}} p_j^t \right)^{\frac{1}{n-1}}$$

(c) Trung bình điều hòa

$$\hat{p}_i^t = \frac{n-1}{\sum_{\substack{j \in C \\ j \neq i}} \frac{1}{p_j^t}}$$

Trong đó: C là phân loại, chẳng hạn cửa hàng hoặc mặt hàng

(3) Gán tỷ lệ: Lấy bình quân tốc độ phát triển của các mặt hàng nhân với giá mặt hàng đó tại thời điểm liền trước:

(a)Trung bình cộng

$$\hat{p}_i^t = p_i^{t-1} \times \frac{\sum_{\substack{j \in I \\ j \neq i}} \frac{p_j^t}{p_j^{t-1}}}{n-1}$$

(b) Trung bình nhân

$$\hat{p}_i^t = p_i^{t-1} \times \left(\prod_{\substack{j \in I \\ j \neq i}} \frac{p_j^t}{p_j^{t-1}} \right)^{\frac{1}{n-1}}$$

(c) Trung bình điều hòa

$$\hat{p}_i^t = p_i^{t-1} \times \frac{n-1}{\sum_{\substack{j \in I \\ j \neq i}} \frac{p_j^{t-1}}{p_j^t}}$$

Nhược điểm phương pháp này là có thể xuất hiện sai số trong kết quả, sai số của giá mặt hàng i tại thời điểm t được tính theo công thức:

$$B_i^t = p_i^t - \hat{p}_i^t.$$

Ví dụ sai số của giá bằng 0,5 bảng, thì ảnh hưởng của sai số đến mặt hàng có giá 0,1 bảng là nghiêm trọng hơn so với mặt hàng có giá 50 bảng, vì vậy cần tính sai số tương đối của phương pháp thay thế giá, công thức tính:

$$RB_i^t = \frac{B_i^t}{p_i^t} = \frac{p_i^t - \hat{p}_i^t}{p_i^t}$$

Sai số tương đối được sử dụng để xác định phương pháp thay thế giá tối ưu. Ví dụ: Giả sử sai số của mặt hàng có giá 0,5 bảng là 0,2 và của mặt hàng 50 bảng là 0,002, phương pháp thay thế giá ảnh hưởng tới các chỉ số giá của mặt hàng thứ nhất nhiều hơn các chỉ số giá của mặt hàng thứ hai. Hướng của sai số cũng quan trọng vì nếu sai số nghiêng hẳn về một nhóm các mặt hàng thì hoàn toàn không tốt. Ví dụ giá thay thế rộng hơn so với giá thu thập thì chỉ số được tính từ giá thu thập có khả năng cao hơn so với chỉ số được tính từ giá thay thế. Mục tiêu của chúng ta là tìm ra phương pháp thay thế nào có thể tối thiểu hóa các sai số tương đối, và cho chúng ta kết quả ước lượng tốt nhất đối với giá bị mất. Giá trị tuyệt đối của các sai số tương đối, tức là lấy giá trị tuyệt đối của sai số tương đối cũng cần được kiểm tra.

3. Nghiên cứu mô phỏng

Để tìm ra phương pháp tối thiểu hóa các sai số tương đối, phương pháp sau được sử dụng: (1) Tìm kiếm một vùng trích xuất dữ liệu không có giá bị mất; (2) Bỏ một mẫu của giá; (3) Thay thế giá; (4) Tính bình quân các sai số tương đối.

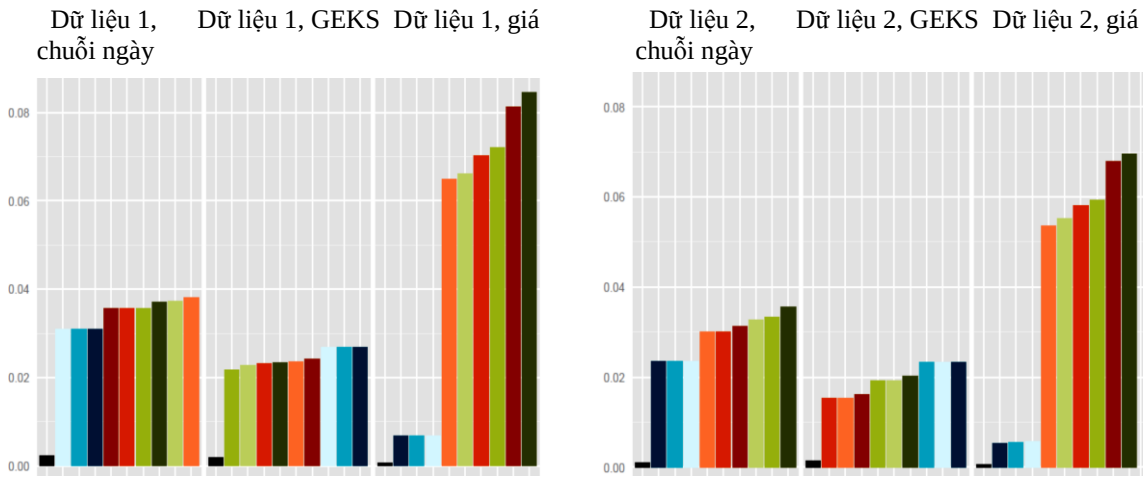
Lấy hai tập dữ liệu trong chuỗi dữ liệu giá theo thời gian, với điều kiện chuỗi thời gian đó không có giá bị mất. Thời gian của hai tập dữ liệu trong chuỗi dữ liệu giá bao gồm ba tuần tiến hành thu thập tập dữ liệu đầu tiên, từ ngày 01/6/2014 đến ngày 22/6/2014, và 4 tuần tiến hành thu thập tập

dữ liệu thứ hai từ thời điểm giữa của thời kỳ trích xuất dữ liệu, từ ngày 12/02/2015 đến ngày 12/3/2015. Tập dữ liệu 1 gồm 3.989 sản phẩm, và tập dữ liệu thứ 2 gồm 3.599 sản phẩm. Vì các tập dữ liệu có khoảng 100.000 giá nên mẫu được chọn là 10% tương ứng 10.000 giá. Số giá bị loại đối với mỗi mặt hàng và nhóm cửa hàng được tính dựa theo phương pháp phân bổ mẫu theo tỷ lệ, duy trì cấu trúc của giá bị mất trong dữ liệu cơ sở. Điều này có ý nghĩa vì các mặt hàng có nhiều loại giá và nhiều chủng loại hơn thường dễ mua, tuy nhiên các mặt hàng này có thể hết hàng nhanh hơn do số lượng hàng dự trữ thường ít nhằm đa dạng chủng loại sản phẩm. Sau khi thực hiện thay thế, sai số tương đối của việc thay thế được tính toán.

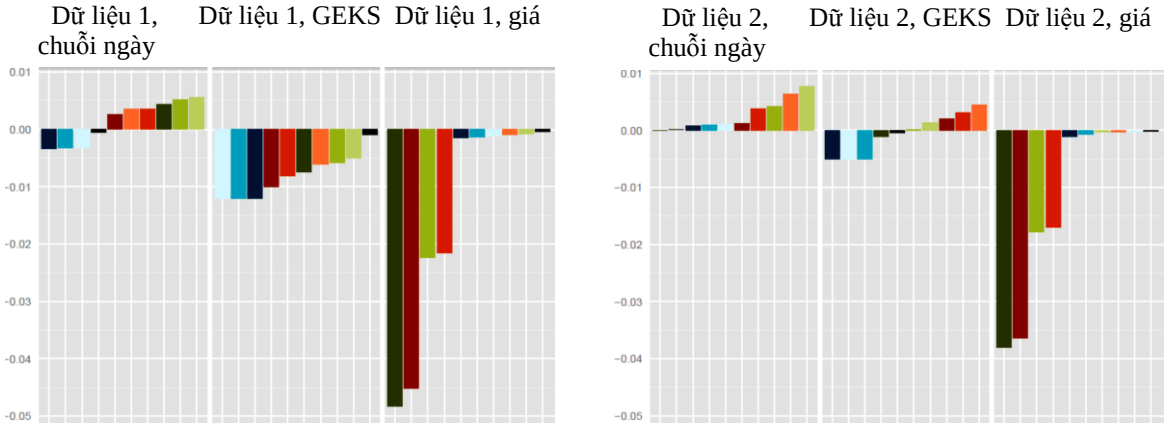
Tiếp theo tính toán hai giá trị bình quân, một là bình quân trị tuyệt đối của sai số tương đối $|\overline{RB}|$ $|\overline{RB}|$, hai là bình quân sai số tương đối $\overline{RB}$. Hai giá trị này được tính cho mỗi phương pháp thay thế đối với mỗi giá, chuỗi ngày (Daily chain) và chỉ số GEKS. Hình 1 cho thấy $|\overline{RB}|$ bình quân trị tuyệt đối của sai số tương đối với mỗi phương pháp thay thế trong cả hai tập dữ liệu. Phương pháp thay thế nào có $|\overline{RB}|$ bình quân trị tuyệt đối của sai số tương đối nhỏ nhất đối với giá và với chỉ số sẽ được sử dụng. Phương pháp thay thế tốt thứ hai phụ thuộc vào công thức chỉ số, đối với chuỗi ngày là phương pháp tốc độ phát triển bình quân, trong khi với chỉ số GEKS là phương pháp thay thế trung bình lớp, mặc dù lớp tốt nhất phụ thuộc vào thời gian. Tuy nhiên, xu hướng chệch này sẽ ảnh hưởng đến tốc độ tăng của chỉ số khá rõ, do đó, thông qua quan sát hướng chệch, kết hợp sử dụng bình quân sai số tương đối, sẽ hỗ trợ tốt hơn cho việc ra quyết định lựa chọn phương pháp nào. Hình 2 chỉ ra điều này.

Các kết quả tương tự đối với bình quân sai số tương đối cũng như bình quân trị tuyệt đối của sai số tương đối, mặc dù độ lớn của sai số tương đối khẳng định rằng việc thay thế không ảnh hưởng đến tốc độ tăng của chỉ số vì giá trị làm tròn cũng giống nhau.

Hình 1: Bình quân trị tuyệt đối của sai số tương đối



Hình 2: Bình quân của sai số tương đối



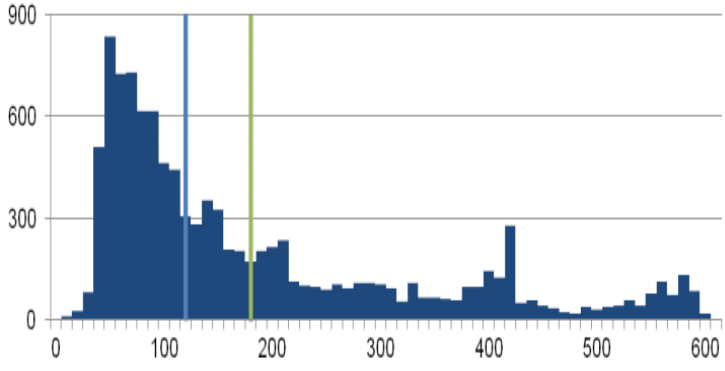
Trong Hình 1 và Hình 2: Phương pháp Inputation

- Bình quân tốc độ phát triển - cộng
- Bình quân tốc độ phát triển - nhân
- Bình quân tốc độ phát triển - điều hòa
- Tính chuyển
- Trung bình phân lớp theo mặt hàng - cộng
- Trung bình phân lớp theo mặt hàng - nhân
- Trung bình phân lớp theo mặt hàng - điều hòa
- Trung bình phân lớp theo cửa hàng - cộng
- Trung bình phân lớp theo cửa hàng - nhân
- Trung bình phân lớp theo cửa hàng - điều hòa

4. Giải thích cho giá thay thế

Hình 3 cho thấy phân bổ thời gian trung bình giữa những thay đổi của giá trong tập dữ liệu được trích xuất. Thời gian bình quân giữa những thay đổi giá được tính là tổng các mục giá hàng ngày/ số lượng giá thay đổi. Hình này không bao gồm các mặt hàng xuất hiện trong tập dữ liệu ít hơn 30 ngày.

Hình 3: Phân bố thời gian trung bình giữa những thay đổi về giá, toàn bộ các mặt hàng, dữ liệu thô từ tháng 6/2014 đến tháng 2/2016



Trung vị là 120 ngày (màu xanh nước biển); trung bình là 181 ngày (xanh lá cây). Hình 3 cho thấy đa số các loại giá không thay đổi thường xuyên, thực tế nhiều giá hoàn toàn không thay đổi trong tập dữ liệu. Điều này hỗ trợ thêm cho khuyến nghị thay thế giá trước đó.

5. Các khuyến nghị

Với các kỹ thuật thay thế giá tối ưu được tìm ra tương ứng các chức năng chính khác nhau, một số khuyến nghị sẽ được áp dụng tùy thuộc vào việc liệu giá thu thập từ hoạt động trích xuất các dữ liệu từ các trang web có được dùng để hỗ trợ cho

việc tính toán chỉ số giá CPI trong tương lai hay không, vì tính toán CPI phải tuân theo quy tắc mà Cơ quan Thống kê châu Âu Eurostat và Tổ chức lao động quốc tế ILO đưa ra. Bảng 1 cho thấy các khuyến nghị này với việc thay thế không phải giai đoạn cơ sở.

Bảng 1: Các khuyến nghị cho việc thay thế giá

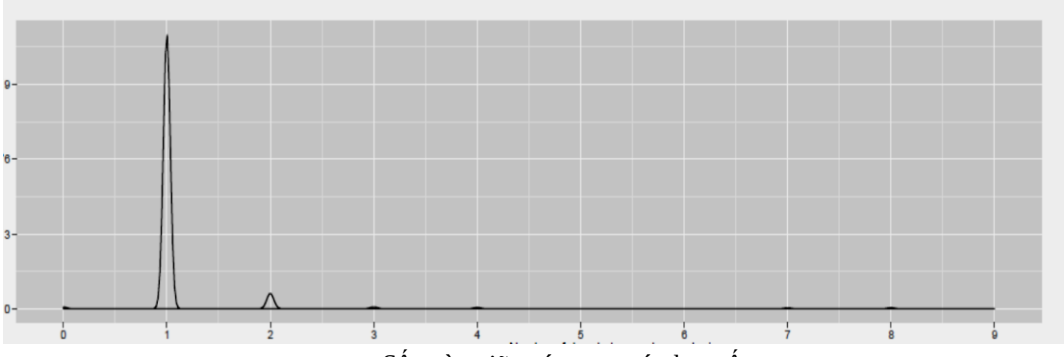
Thay thế	Dữ liệu được sử dụng để hỗ trợ tính toán CPI	Chỉ dùng trong thống kê thực nghiệm
Giá	Trung bình nhân tốc độ phát triển	Thay thế
Chuỗi hàng ngày	Trung bình nhân tốc độ phát triển	Thay thế
GEKS	Trung bình nhân phân lớp theo cửa hàng	Thay thế

6. Thay thế trong bao lâu?

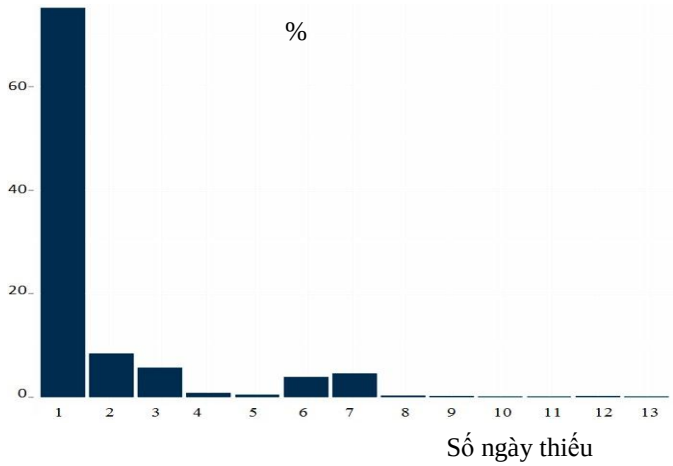
Thay thế giá là một cách tốt để giải quyết vấn đề mất giá, giúp tăng tính ổn định cho cỡ mẫu trong suốt thời kỳ quan sát, nhưng thực tế đôi khi một sản phẩm có thể đồng thời hết hàng trong thời kỳ cụ thể hoặc được bổ sung thêm hoặc biến mất khỏi thị trường. Vì thế, có thể là thiếu khôn ngoan khi tiếp tục thay thế giá trong những trường hợp này, vì nó sẽ làm cho chỉ số cố định hoặc khiến cho chỉ số không còn đại diện cho sự biến động giá thực tế. Để quyết định được số ngày phù hợp đối với việc thay thế giá, số ngày giữa các giá quan sát sẽ được tính toán, đồng thời tính phân phối Gaussian - ước lượng hàm mật độ Kernel (KDE) của phân bổ cho tất cả các mặt hàng và cho từng mặt hàng. Hình 4 cho thấy ước lượng mật độ Kernel KDE (Kernel density estimation) cho từng mặt hàng. Hình 4 cho thấy KDE của tất cả các mặt hàng trong tập dữ liệu đã được làm sạch.

Quan sát các phân bổ trong hình 4 nhận thấy sự khác biệt, mật độ ngày 1 cao nhất tiếp theo là ngày 2 và thấp hơn ở ngày thứ 3. Sự khác biệt mỗi ngày có nghĩa giá được liên tục quan sát qua các ngày. Sau khi loại bỏ dữ liệu giá được quan sát liên tục hàng ngày trung bình số ngày giữa các giá quan sát là 2,7 ngày, do đó có thể khuyến nghị nên thay thế giá 3 ngày sau khi một mặt hàng bị loại bỏ. Nếu gián đoạn trích xuất dữ liệu nhiều hơn 3 ngày liên tiếp vẫn tiếp tục thay thế cho tới khi việc trích xuất được thực hiện lại, trừ khi việc gián đoạn nhiều hơn một tuần thì dừng thay thế giá. Hình 5 cho thấy giá trị của 3 ngày và 7 ngày không phải ngẫu nhiên, bởi 3 ngày tỷ lệ các sản phẩm có giá bị mất là 89% và 7 ngày là 99%, vì vậy việc tính cho một tuần không bao gồm hầu hết sự biến mất.

Hình 4: Sự sai khác ngày KDE đối với tất cả các mặt hàng



Hình 5: Phần trăm các sản phẩm theo số ngày mất giá



Quy tắc thay thế 7 ngày đối với các khoảng trống trích xuất cũng được xác định bằng cách quan sát số ngày xuất hiện khoảng trống trích xuất dữ liệu. Ở Bảng 2 ta thấy, phần lớn các khoảng trống dữ liệu ít hơn một tuần, việc thay thế tối đa 7 ngày có thể chấm dứt sự bất thường trong chuỗi chỉ số, nguyên nhân bất thường xuất phát từ những giá bị mất.

Bảng 2: Độ dài khoảng trống trích xuất theo siêu thị từ tháng 6/2014 đến 04/2016

Độ dài khoảng trống (ngày)	Siêu thị			
	Sainsbury	Tesco	Waitrose	Lab Failure
1	22	15	16	12
2	1	2	1	1
3	1	2	2	2[1]
4	1	0	0	0
6	1	1	0	0
7	1	1	0	0
26	1	0	0	0
34	1	1	1	1

[1] *Số này lớn hơn số liệu của ba ngày không trích xuất được dữ liệu từ siêu thị Sainsbury vì khoảng trống thử nghiệm ba ngày là một phần của khoảng trống dài hơn đối với trường hợp trích xuất dữ liệu từ siêu thị Sainsbury.*

7. Sự thay thế có ảnh hưởng tới các chỉ số?

Phần này xem xét chỉ số GEKSJ, đóng vai trò như một chỉ tiêu bị ảnh hưởng bởi việc thay thế, khi mà việc thay thế đã được thực hiện trong toàn bộ thời gian thu thập. Quan sát các kết quả cho thấy, có hai loại ảnh hưởng khác nhau, bao gồm:

1. Các chỉ số tính toán từ việc sử dụng dữ liệu được thay thế gần giống với các chỉ số được tính toán bằng các dữ liệu không phải là dữ liệu thay thế.

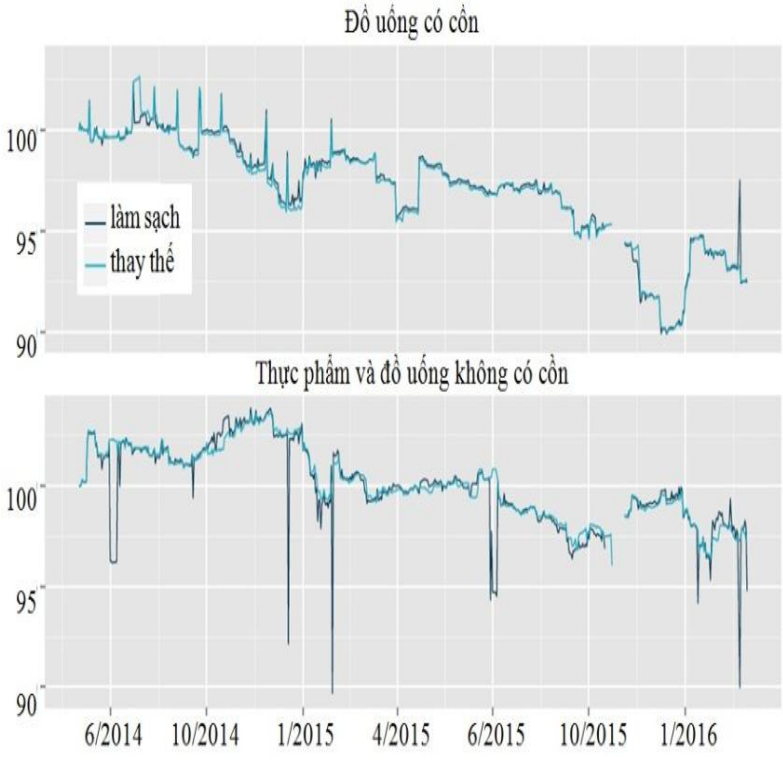
2. Các chỉ số được tính toán sử dụng dữ liệu đã được xử lý bằng cách loại bỏ những giá trị bất thường và làm trơn chuỗi.

Hình 6 chỉ ra cả hai trường hợp[2] đối với mặt hàng đồ uống có cồn, chỉ số GEKSJ tính từ các dữ liệu được thay thế gần giống với chỉ số GEKSJ tính từ các dữ liệu được làm sạch, vì vậy việc thay thế không làm thay đổi chỉ số. Mặt khác, các chỉ số của mặt hàng thực phẩm và đồ uống không cồn minh

[2] *Khoảng trống trong chuỗi do khoảng trống trích xuất lớn hơn và do vậy dẫn đến các quy tắc thay thế vẫn có dữ liệu thiếu.*

chứng cho trường hợp thứ hai, vì các chỉ số tính từ các dữ liệu không được thay thế thường xuyên xuất hiện những bất thường. Nguyên nhân do chỉ số thực phẩm là một chỉ số tổng hợp bao gồm các chỉ số cấp thấp hơn, sử dụng quyền số từ Điều tra thực phẩm và mức sống, các quyền số có tổng bằng 1, vì vậy những khoảng ngắt quãng nguyên nhân do mất giá dẫn đến mất chỉ số, hậu quả là tổng quyền số không phải là 1. Việc thay thế giá giúp các chỉ số không bị mất đi, vì vậy tổng các quyền số vẫn là 1 và do đó tốc độ tăng của chỉ số hoàn toàn do sự thay đổi về giá, không phải do sự thay đổi về mặt quyền số. Từ việc thay thế giá người tiêu dùng hiểu hơn về lạm phát, ngay cả khi có các khoảng trống trích xuất thì họ vẫn có thể mua các sản phẩm từ các trang web. Đối với các trường hợp tạm thời hết hàng, người tiêu dùng ở những vùng khác nhau vẫn có thể mua sản phẩm vì siêu thị được trích xuất dữ liệu là các siêu thị có chuỗi cửa hàng

Hình 6: Chỉ số GEKSJ của thực phẩm, đồ uống có cồn



khắp quốc gia, và việc thay đổi sản phẩm sẵn có trên trang web phụ thuộc vào các sản phẩm sẵn có tại địa phương nơi người tiêu dùng sinh sống.

8. Kết luận

Tóm lại, sử dụng phương pháp thay thế giá là một phương pháp tốt trong việc giải quyết vấn đề mất giá do mặt hàng không sẵn có trong giỏ hàng hóa và khoảng trống trích xuất dữ liệu. Điều này là do có tác động thuận lợi tới các chỉ số và ngăn chặn sự biến động bất thường nguyên nhân do quyền số thay đổi. Phương pháp thay thế tốt nhất là tiến hành thay thế giá sao cho bình quân sai số tương đối nhỏ nhất. Bước thay thế giá này được dùng trong việc cập nhật nghiên cứu về việc sử dụng dữ liệu được trích xuất từ trang web để tính các chỉ số giá.

Minh Ánh (lược dịch)

Nguồn: Imputing Web Scraped Prices, https://www.ons.gov.uk/economy/inflationandpriceindices/methodologies/imputingwebscrapedprices.