□□□□ THỐNG KÊ QUỐC TẾ VÀ HỘI NHẬP □□□□

ỨNG DỤNG DỮ LIỆU LỚN ĐỂ TÍNH CHỈ SỐ GIÁ TIÊU DÙNG CỦA MỸ: KINH NGHIỆM VÀ NHỮNG KẾ HOẠCH

Tóm tắt:

Sự phát triển không ngừng của dữ liệu lớn giúp mở ra các cơ hội mới cho hoạt động thống kê. Trường hợp Ủy ban Thống kê Lao động Mỹ (BLS) là một trong những minh chứng điển hình của việc sử dụng dữ liệu lớn giúp nâng cao hiệu quả công tác thống kê hiện đại. Hiện nay, BLS đã tiến hành một số dự án thí điểm sử dụng các nguồn dữ liệu mới với mục đích bổ sung hoặc thay thế cho phương pháp thu thập truyền thống. Dựa trên cơ sở báo cáo "Dữ liệu lớn trong Chỉ số giá tiêu dùng của Mỹ: Kinh nghiệm & những kế hoạch" của tác giả Crystal G. Konny, Brendan K. Williams, and David M. Friedman công bố vào tháng 2 năm 2019, bài viết sẽ tóm lược một số nội dung chính liên quan đến các nguồn dữ liệu thay thế, thách thức trong việc sử dụng, kinh nghiệm thực tiễn khai thác đối với một số dự án và một số kết luận, kế hoạch trong tương lai.

1. Chỉ số giá tiêu dùng và các nguồn dữ liệu thay thế

1.1 Chỉ số giá tiêu dùng

Chỉ số giá tiêu dùng (CPI) là thước đo sự thay đổi giá bình quân theo thời gian của giỏ hàng hóa tiêu dùng trên thị trường hàng hóa và dịch vụ. CPI cũng là một chỉ số đo lường tổng hợp được biên soạn dựa trên việc kết hợp các lý thuyết kinh tế và các kỹ thuật thống kê khác.

Giá dùng trong biên soạn CPI theo phương pháp truyền thống hiện nay được thu thập từ hai kênh bao gồm: Khảo sát Hàng hóa và Dịch vụ (C&S) và Khảo sát Nhà ở được thực hiện bởi các điều tra viên của BLS.

Hạn chế lớn nhất của việc thu thập thông tin thông qua các cuộc khảo sát là chi phí thực hiện điều tra và việc tiến hành cũng ngày càng trở lên khó khăn hơn. Các khu vực ngày càng mở rộng, dẫn đến chi phí đi lại ngày càng tăng. Việc gia tăng số lượng chuỗi cửa hàng làm cho thời gian thu thập kéo dài hơn vì cần có sự chấp thuận của các doanh nghiệp để tiến hành thu thập dữ liệu tại các cửa hàng. Tỷ lệ trả lời ngày càng giảm do nhiều yếu tố: yêu cầu bảo mật mới, tăng số lượng khảo sát, tăng tỷ lệ mất lòng tin vào chính phủ, mối lo ngại về bảo mật dữ liệu và/hoặc ít tin tưởng vào độ chính xác của CPI.

Xuất phát từ thực tế trên, các nguồn dữ liệu thay thế ngày nay đã và đang mang lại cơ hội tuyệt vời giúp BLS giải quyết nhiều thách thức mà khảo sát giá tiêu dùng đang gặp phải. Thông qua các nguồn dữ liệu thay thế cho phép đo lường chính xác hơn sự thay đổi giá, giúp mở rộng cỡ mẫu, thu thập được giá giao dịch thay vì giá đề xuất, phản ánh chính xác hơn việc sử dụng các mặt hàng thay thế của người tiêu dùng, loại bỏ sự thay đổi chất lượng, giảm hoặc loại bỏ gánh nặng của người trả lời, giải quyết các vấn đề không có câu trả lời trong các khảo sát CPI và giảm chi phí thu thập trong một số tình huống.

Trong một số trường hợp, nguồn dữ liệu thay thế có thể giúp thu thập thông tin tiêu dùng kịp thời hơn. Dữ liệu có thể ở mức chi tiết hơn, số lượng mặt hàng đa dạng hơn so với mẫu hiện nay và tần suất thu thập cao hơn. Chính vì vậy, trong những năm gần đây, BLS đang chú ý tìm kiếm một quy trình thu thập mới mang lại hiệu quả hơn về mặt chi phí.

1.2 Các nguồn dữ liệu thay thế

Hiện nay, các nguồn dữ liệu thay thế được chia thành ba loại chính:

1.2.1. Dữ liệu do doanh nghiệp cung cấp là bộ dữ liệu của người trả lời cuộc khảo sát được cung cấp trực tiếp từ trụ sở của doanh nghiệp thay vì người trả lời là các chủ cửa hàng. Các bộ dữ liệu này thường được tạo ra để phục vụ mục đích quản lý. Các thành phần và cấu trúc dữ liệu được xác định bởi người trả lời và BLS phải điều chỉnh dữ liệu để phù hợp với hệ thống của mình. BLS nhận được các mức thông tin khác nhau về các tập dữ liệu - nói chung, thông tin được cung cấp là những gì các doanh nghiệp sẵn sàng cung cấp.

1.2.2. Dữ liệu nguồn thứ cấp (bộ dữ liệu của bên thứ ba) được biên soạn bởi bên thứ ba, chứa giá cho hàng hóa hoặc dịch vụ từ nhiều cơ sở mà BLS cần phải mua hoặc được cung cấp miễn phí từ bộ dữ liệu tổng hợp. Người tổng hợp dữ liệu thực hiện một số công tác chuẩn hóa cấu trúc dữ liệu giữa các cơ sở kinh doanh.

1.2.3. Dữ liệu trích xuất từ các web là dữ liệu được nhân viên BLS thu thập tự động bằng phần mềm trích xuất dữ liệu để thu thập giá cả và đặc điểm sản phẩm từ các trang web. Một số cơ sở cung cấp giao diện lập trình ứng dụng (API) hoặc mã nguồn mở cho phép các đối tác truy cập thông tin về giá. Việc thu thập dữ liệu thông qua API thường dễ dàng và đơn giản hơn so với việc duy trì mã quét web theo thời gian.

2. Thách thức sử dụng nguồn dữ liệu thay thế

2.1 Thách thức liên quan phương pháp biên soạn CPI

Trước hết là thách thức liên quan đến sự phù hợp giữa dữ liệu thay thế với phương pháp tính CPI hiện tại. Trở ngại chính trong việc xử lý dữ liệu giao dịch trong CPI là xử lý các hiệu ứng vòng đời sản phẩm, ví dụ khi các sản phẩm thể hiện xu hướng giá một cách hệ thống trong vòng đời. Đối với một số hàng hóa nhất định như hàng may mặc và các mặt hàng điện tử, một sản phẩm thường được chào bán mức giá cao trên thị trường và dần được giảm giá theo thời gian. BLS nhận thấy các phương pháp chỉ số giá đa phương được thiết kế để giải quyết chuỗi trôi không khắc phục được sự trôi xuống liên quan đến vòng đời sản phẩm. Các phương pháp Hedonic thông thường cũng không giải quyết được các hiệu ứng vòng đời sản phẩm. Đối với từng trường hợp cụ thể, BLS đã phát triển phương pháp sử dụng thay đổi giá hàng năm để tránh các hiệu ứng vòng đời.

Nhiều nguồn dữ liệu thay thế có thể được thu thập thay cho việc thu thập dữ liệu các cửa hàng nằm trong mẫu thuận tiện, điều này giúp cho việc thu thập dữ liệu dễ dàng hơn. Khi các cửa hàng thuộc doanh nghiệp không đủ mức thị phần trên thị trường sẽ dẫn đến việc tính đại diện trong một bộ dữ liệu thay thế không được đảm bảo, do đó có khả năng gây lỗi trong biên soạn CPI. Trong các trường hợp khác, chẳng hạn như với dữ liệu trích xuất, nguồn dữ liệu được trích xuất đảm bảo tính đại diện sẽ phản ánh tốt hơn thị trường thực tế, CPI có thể xử lý dữ liệu được thu thập bằng cách chọn mẫu ngẫu nhiên và ước lượng phương

sai. Mặc dù dữ liệu lớn cho thấy vấn đề đối với việc ước lượng phương sai, nhưng nó lại có khả năng giải quyết được vấn đề cỡ mẫu nhỏ của truyền thống và giảm bớt sai số chọn mẫu.

Những thách thức còn lại về mặt phương pháp liên quan đến mức độ chi tiết được cung cấp bởi nguồn dữ liệu thay thế. Các doanh nghiệp có thể không sẵn sàng hoặc không thể cung cấp dữ liệu với mức độ chi tiết giống như dữ liệu điều tra được thu thập bởi các điều tra viên. Định nghĩa của doanh nghiệp về một mặt hàng duy nhất có thể không phù hợp với định nghĩa BLS đưa ra, điều này làm cho việc định giá cùng một mặt hàng theo thời gian trở nên khó khăn. Thông tin hạn chế về các tính năng của sản phẩm và mô tả mặt hàng không có cấu trúc đòi hỏi phải có cách tiếp cận mới đối mô hình chỉ số và việc điều chỉnh chất lượng trong CPI. Hầu hết các nguồn dữ liệu thay thế cũng bỏ qua thông tin thuế bán hàng và có thể không cung cấp đủ thông tin để xác định thẩm quyền thuế mà CPI cần áp dụng thuế suất.

2.2 Những thách thức trong quy trình thu thập và xử lý

Mặc dù tốc độ là một trong những ưu điểm của dữ liệu lớn, nhưng vấn đề thời gian của cả bộ dữ liệu thứ cấp và sơ cấp cũng có thể là vấn đề. Yêu cầu của BLS đối với chỉ số hàng tháng không phải là ưu tiên hàng đầu của các nhà cung cấp dữ liệu. BLS phải kiểm soát tất cả quy trình dữ liệu thu thập theo truyền thống đồng thời cần có thêm nhiều quy trình kiểm soát chất lượng tổng thể đối với các nguồn dữ liệu được thu thập và sử dụng để tính toán chỉ số giá tiêu dùng CPI.

Độ sạch của dữ liệu cũng có thể là một rủi ro với dữ liệu của nhà cung cấp. Dữ liệu mô tả không được thu thập và việc so sánh dữ liệu theo thời gian không được đảm bảo. Ngoài ra, rủi ro cũng có thể phát sinh từ việc sử dụng nguồn dữ liệu không ổn định và có thể biến mất mà không có bất kỳ cảnh báo nào.

Để một nguồn dữ liệu thay thế được sử dụng kết hợp trong đo lường CPI, dữ liệu phải được gắn tương ứng vào phân loại danh mục và cấu trúc địa lý của CPI. Điều này đơn giản khi một tập dữ liệu bao gồm danh mục các sản phẩm trong CPI. Tuy nhiên, với một số trường hợp nhất định, dữ liệu giao dịch thu được bao gồm nhiều loại mặt hàng và BLS phải khớp với danh mục mặt hàng dựa trên các phân loại và mô tả mặt hàng của các doanh nghiệp. Một hệ thống học máy đã được phát triển nhằm hỗ trợ các phân loại này, kết quả đã cải thiện đáng kể khả năng xử lý các bộ dữ liệu lớn với hàng trăm ngàn mục.

Khi đã có được nguồn dữ liệu, giải quyết được mọi vấn đề về phương pháp luận và tiến hành kết hợp nguồn dữ liệu mới vào việc tính toán CPI, BLS vẫn phải xử lý việc tích hợp dữ liệu vào các hệ thống công nghệ thông tin hiện tại với giả định dữ liệu được cấu trúc theo quy trình thu thập dữ liệu khảo sát. Về cơ bản, có hai cách để thực hiện việc này mà không cần sửa đổi toàn bộ các hệ thống CPI. Hoặc là thay thế một quan sát giá riêng lẻ trong CPI, hoặc là thay thế một chỉ số thành phần bằng một chỉ số có nguồn gốc từ dữ liệu thay thế. Thay thế các quan sát về giá riêng lẻ sẽ hoạt động tốt khi kết hợp dữ liệu được khảo sát và dữ liệu thay thế trong các mục. Tuy nhiên, hệ thống hiện tại không được thiết kế để tạo ra các quan sát giá mới, vì vậy chiến lược hiện tại của BLS là khớp giá ước tính hoặc thay đổi giá với quan sát giá hiện tại đã được chọn để lấy mẫu.

2.3 Những thách thức về pháp lý, chính sách và ngân sách

Đối với các nguồn dữ liệu thứ cấp, các thách thức về pháp lý, chính sách và ngân sách thường tập trung vào việc đàm phán hợp đồng phù hợp với quy định của Luật Liên bang và đáp ứng nhu cầu của cả hai bên.

Trở ngại chính của BLS là việc đảm bảo tính hợp pháp cho nguồn dữ liệu được trích xuất. Những lo ngại liên quan đến việc quét web đã phát sinh cả trong nội bộ và từ những người được hỏi. Để đảm bảo tất cả dữ liệu thay thế được sử dụng trong nghiên cứu hoặc sản xuất được bảo vệ theo đạo luật thống kê, BLS phải cung cấp cho các cơ sở, bao gồm cả những dữ liệu thu thập trực tuyến, cho dù là thủ công hay tự động, cam kết bảo mật sẽ sử dụng thông tin cho mục đích thống kê. Trong trường hợp nguồn dữ liệu thứ cấp, một điều kiện của hợp đồng có thể là nhà cung cấp được công nhận công khai. Trong trường hợp quét web, BLS không thể tiến hành mà không có sự cho phép của cơ sở. Hơn nữa, các thỏa thuận trong điều khoản dịch vụ cho các trang web và các mã nguồn mở API thường có các khía cạnh gây rắc rối cho các cơ quan thống kê.

Cuối cùng, cần phải đảm bảo rằng việc chuyển đổi sang các nguồn dữ liệu thay thế không làm tăng ngân sách chung, tức là công việc này vẫn duy trì ít nhất là ngân sách trung lập nếu không thực sự tiết kiệm chi phí tổng thể.

3. Kinh nghiệm khai thác các nguồn dữ liệu thay thế

3.1 Kinh nghiệm của BLS trong việc khai thác các bộ dữ liệu doanh nghiệp

Doanh nghiệp X

Tháng 5/2016, doanh nghiệp X đã bắt đầu cung cấp cho BLS dữ liệu hàng tháng về giá trung bình và doanh thu bán hàng của mỗi sản phẩm được bán cho mỗi cửa hàng thuộc doanh nghiệp X trong các khu vực địa lý bao gồm trong CPI. (Trước tháng 5/2016, BLS có được dữ liệu không được phép sử dụng và sau đó doanh nghiệp X đã tái cơ cấu cơ sở dữ liệu của mình và quyết định cung cấp dữ liệu cho BLS.) Tuy nhiên, dữ liệu chỉ bao gồm các mô tả giới hạn về các mặt hàng được bán. Không có dữ liệu có cấu trúc về các tính năng của sản phẩm và mô tả biến là tương đối ngắn. Việc thiếu dữ liệu mô tả này cản trở việc xây dựng hồi quy hedonic hoặc đưa ra quyết định sáng suốt về khả năng so sánh tương đối của các mặt hàng mới với các mặt hàng hiện hành, hạn chế khả năng áp dụng các phương pháp điều chỉnh chất lượng và thay thế thông thường của CPI. Dữ liệu được đánh giá trong khoảng hai năm nhằm thay thế cho hơn 1000 giá được sử dụng trong CPI.

Phân tích nội bộ cho thấy xu hướng các chỉ số mô hình phù hợp giảm nhanh chóng. Một số loại mặt hàng cho thấy giảm hơn 90% trong vòng chưa đầy hai năm. Hầu hết những sự sụt giảm này có thể được cho là kết quả của chiến lược giá nhà bán lẻ. Sản phẩm được giới thiệu với giá cao và giảm giá theo thời gian.

BLS đã phát triển một phương pháp ngắn hạn mô phỏng theo các quy trình CPI hiện tại để bắt đầu kết hợp dữ liệu từ nhà bán lẻ này vào chỉ số giá tiêu dùng hiện tại. Phương pháp chọn một mẫu trên cơ sở tỷ lệ doanh số bán hàng trong các bộ dữ liệu do phía X cung cấp và tính toán giá tương quan mô hình cho các mặt hàng được chọn trong suốt một năm. Các chỉ số mô hình đối chiếu

này thường cho thấy xu hướng giảm. Sau mười hai tháng, một mẫu sản phẩm mới thuộc cùng loại mặt hàng được chọn và giá tương đối được xây dựng là giá trung bình của tất cả các sản phẩm mới trong danh mục mặt hàng so với giá trung bình của sản phẩm trong danh mục 12 tháng trước. Tỷ lệ giữa đơn giá của mẫu mới và mẫu cũ thường dương và bù cho giá giảm trong năm do vòng đời sản phẩm.

Để kết hợp dữ liệu từ phía X vào CPI, BLS cũng phải phát triển cách lập bản đồ phân loại mặt hàng. Nhà bán lẻ sẽ cung cấp các mô tả ngắn và thông tin phân loại cho từng mặt hàng được bán tại các cửa hàng của mình trong các khu vực địa lý được nêu trong CPI. Kết hợp thủ công từng mặt hàng, theo thứ tự từ một đến hàng trăm nghìn, với danh mục sản phẩm CPI là không khả thi. Dựa trên các phương pháp được phát triển tại BLS cho việc mã hóa tự động, BLS đã sử dụng công cụ học máy để phân loại các mặt hàng theo cấu trúc CPI dựa trên các mô tả. Nhân viên CPI phân loại mã hóa bằng tay cho một phân đoạn của các mục trong dữ liệu doanh nghiệp để tạo ra một tập dữ liệu đào tạo. Sau đó, sử dụng cách tiếp cận "những chiếc túi ngôn ngữ" dựa trên tần suất xuất hiện của các từ trong mục mô tả. Hồi quy logistic sau đó được sử dụng để ước tính xác suất của từng mục. Sau khi xác thực kết quả và xem xét dự đoán độ tin cậy, BLS sử dụng phương pháp này với mỗi tập dữ liệu hàng tháng để phân loại các mục mới.

Doanh nghiệp Y

Một chuỗi siêu thị khác (gọi tắt là CorpY) đã đồng ý cung cấp cho BLS dữ liệu về các loại thuốc kê đơn tại trụ sở doanh nghiệp. Tháng 2/2012, doanh nghiệp Y từ chối tham gia các mẫu mới do gánh nặng đặt lên các hiệu thuốc trong cửa hàng. Các cuộc thảo luận đã diễn ra giữa văn phòng thống kê của khu vực và phía doanh nghiệp về việc cung cấp dữ liệu của doanh nghiệp làm sao vừa có thể sử dụng dữ liệu cho biên soạn CPI vừa và đáp ứng được các vấn đề bảo mật của doanh nghiệp Y. Kể từ tháng 3/2015, phía doanh nghiệp Y đã cung cấp cho cơ quan điều tra giá bộ dữ liệu giá giao dịch thuốc theo toa tại cửa hàng của họ trung bình 2 tháng/lần.

Với các phương pháp thu thập truyền thống, CPI xác định một mặt hàng duy nhất thuộc danh mục mã thuốc quốc gia và tiến hành theo dõi theo thời gian mã, số lượng, nhà cung cấp và lập kế hoạch đảm bảo và giá thành. Bằng cách giữ các biến này không đổi, CPI có thể kiểm soát thay đổi giá không phải là do thay đổi chất lượng thuốc. Các điều tra viên gán mã cho dược phẩm theo mã dược phẩm quốc gia và đưa ra các thông tin chi tiết về nhà sản xuất và liều lượng dùng. Điều tra viên cũng ghi lại giá niêm yết tại các nhà bán lẻ thuốc theo đơn.

Khi thuốc chính hiệu mất quyền bảo vệ bằng sáng chế và thuốc của các đối thủ cạnh tranh tham gia vào thị trường, việc kinh doanh sẽ chậm hơn. Điều tra viên sẽ yêu cầu dược sĩ cung cấp tỷ lệ phần trăm của thuốc bán nói chung so với thuốc có thương hiệu, dựa trên các tỷ lệ mẫu của thương hiệu hoặc tỷ lệ chung đó để tiếp tục xác định giá. Nếu một thuốc không có thương hiệu được chọn, sự thay đổi giá giữa thuốc có tên thương hiệu và thuốc đó được phản ánh trong CPI.

Do phía doanh nghiệp lo ngại về tính bảo mật và gánh nặng báo cáo, BLS cần phải thỏa hiệp và nhận bộ dữ liệu với tần suất hai tháng một lần. Doanh nghiệp Y tính giá bình quân của mặt hàng có nhãn hiệu và không nhãn hiệu. Khi người tiêu dùng sử dụng thay thế giữa thuốc có nhãn hiệu và thuốc không nhãn hiệu, giá trung bình sẽ thay đổi. Mặc dù thường có sự khác biệt lớn về chi phí tự trả giữa thuốc thương hiệu và

thuốc chính hãng, nhưng chúng thường được coi là tương đương và hiệu quả như nhau và do đó, giá trung bình được coi là chấp nhận được bởi CPI.

3.2 Kinh nghiệm khai thác với các nguồn dữ liệu thứ cấp

Một số nhà cung cấp tiến hành tổng hợp và bán dữ liệu của họ. Các bộ dữ liệu này thường được sử dụng bởi các bên tiếp thị và thường được xây dựng tập trung vào cấp độ bán hàng hơn là cung cấp mức độ chi tiết của sản phẩm. Hầu hết các bộ dữ liệu bao gồm nhiều mặt hàng hơn so với số mặt hàng thuộc danh mục CPI. BLS đã mua một số bộ dữ liệu và nghiên cứu sử dụng chúng để thay thế cho các thành phần CPI. Nguồn dữ liệu thứ cấp cho thấy các vấn đề tương tự như những vấn đề được tìm thấy trong dữ liệu của doanh nghiệp.

Dữ liệu thường thiếu chi tiết mô tả so với thông tin được điều tra viên thực hiện trong khảo sát C&S thường không có sự minh bạch từ các nguồn thứ cấp về mức độ sẵn sàng chia sẻ đầy đủ các phương pháp của họ với BLS. Dưới đây là kinh nghiệm sử dụng các nguồn dữ liệu thứ cấp.

3.3 Dữ liệu các phương tiện mới

Để giải quyết gánh nặng của người trả lời, tỷ lệ phản hồi thấp, giá ước tính của đại lý và chi phí thu gom cao, BLS mua dữ liệu giao dịch từ nhà cung cấp J.D. Power. Ngoài việc giải quyết các vấn đề của phương pháp thu thập truyền thống, dữ liệu của J.D. Power có thể cung cấp thông tin chất lượng cao hơn bao gồm giá giao dịch và chi tiêu theo thời gian thực. Dữ liệu cho phép đo lường chi phí sinh hoạt tốt hơn so với chỉ số hiện tại.

J.D. Power cung cấp cho BLS dữ liệu giao dịch bao gồm khoảng một phần ba doanh số bán xe mới ở Hoa Kỳ. Phân tích chỉ ra rằng thị phần của các phương tiện giao thống trong CPI và Dữ liệu của J. Power Power tương tự nhau. Mỗi hồ sơ chứa thông tin về cấu hình xe, giá giao dịch và bất kỳ khoản tài chính nào được thiết lập bởi các đại lý. Mã định danh có sẵn trong bộ dữ liệu J.D. Power được sử dụng để xác định một mục duy nhất, đặc biệt là các tùy chọn cụ thể được bán với một giao dịch nhất định.

Doanh số bán xe mới hiển thị vòng đời sản phẩm trong đó xe được giới thiệu ở mức giá cao và sau đó được giảm giá qua năm mô hình cho đến khi chúng được thay thế bằng xe kế nhiệm. Do kết quả của mô hình này, các chỉ số giá xe mới phù hợp với mô hình cho thấy sự sụt giảm đều đặn vì chúng chỉ phản ánh sự sụt giảm giá trong năm và không tính đến bất kỳ thay đổi giá trong năm của mô hình chéo. Hành vi chỉ số này có thể gợi ý sự trôi dạt chuỗi, nhưng như trường hợp của chỉ số CorpX, sự trôi dạt chuỗi dường như không phải là một yếu tố do các phương pháp đa phương không thể làm giảm sự di chuyển xuống. Williams và Sager cho rằng giá giảm theo vòng đời của sản phẩm có thể do người bán sử dụng chiến lược phân biệt giá không phù hợp với giả định về người tiêu dung có sở thích ổn định như trong lý thuyết chỉ số giá sinh hoạt.

Đo lường giá cả qua các năm qua làm trơn các biến động tần số cao trên thị trường. Để khôi phục thông tin về hành vi ngắn hạn của thị trường xe mới, chỉ số giá tần suất hàng tháng được tính toán. Bộ lọc chuỗi thời gian được sử dụng để tách một thành phần theo chu kỳ khỏi xu hướng sai lệch của chỉ số tần số hàng tháng. Thành phần theo chu kỳ này được kết hợp với xu hướng hàng năm để tạo ra một chỉ số (YY + Chu kỳ) phản ánh cả hành vi ngắn hạn và dài hạn của giá xe mới.

3.4. Dữ liệu giá dịch vụ y tế

Hiện tại, nhóm chăm sóc y tế có tỷ lệ đáp ứng thấp nhất trong tất cả các nhóm chính trong CPI, trong đó, dịch vụ y tế và các dịch vụ của bệnh viện là quan trọng nhất. Có nhiều lý do cho phản hồi thấp này và tất cả đều rất khó khắc phục, chẳng hạn như những lo ngại về tính bảo mật được Luật quy định về trách nhiệm giải trình và trách nhiệm bảo hiểm y tế, khó khăn trong việc xác định tỷ lệ gói bảo hiểm, phân chia cho bác sĩ viện phí và vấn đề đầu vào. BLS quyết định tìm hiểu tính khả thi của việc bổ sung thu thập dữ liệu truyền thống về tiền mặt và giá dịch vụ chăm sóc sức khỏe bằng dữ liệu bảo hiểm. BLS đã mua một bộ dữ liệu y tế năm 2009 và 2010 của một hãng bảo hiểm với một mẫu nhỏ các dịch vụ y tế ở khu vực đô thị Chicago. Giá trung bình trên tất cả các giao dịch cho sự kết hợp giữa nhà cung cấp/dịch vụ y tế đã được nhận hàng tháng cùng với số lượng giao dịch được sử dụng để tạo ra mức giá trung bình. Mục tiêu nghiên cứu chính là phân tích hiệu quả của việc sử dụng dữ liệu yêu cầu bảo hiểm bị trì hoãn. Khiếu nại bảo hiểm thường mất vài tháng để được phân xử đầy đủ và việc xử lý dữ liệu của nhà cung cấp có thể mất thêm thời gian. Dữ liệu khiếu nại sẽ bị chậm trễ, từ hai đến chín tháng, trước khi có thể được gửi tới cơ quan thống kê để biên soạn CPI.

3.5 Dữ liệu giá dịch vụ điện thoại không dây

Bắt đầu vào tháng 2 năm 2018, BLS đã nghiên cứu và tận dụng một nguồn dữ liệu khảo sát hộ gia đình thứ cấp dựa vào các nhà mạng để đưa ra tỷ lệ lấy mẫu cho các dịch vụ điện thoại không dây.

BLS đã tính toán các chỉ số nghiên cứu với một nguồn thứ cấp khác có giá niêm yết cho các gói dịch vụ điện thoại không dây được thu thập từ các trang web của các nhà mạng không dây. Sự đảm bảo của các nhà cung cấp CPI là hơn 90%. Phương pháp "đối chiếu và thay thế" đã được sử dụng để tính toán các chỉ số, theo đó các gói dịch vụ trong tập dữ liệu CPI được khớp với các mô tả kế hoạch trong dữ liệu thay thế, giá được thay thế và các chỉ số được tính toán lại bằng phương pháp CPI hiện tại và phần còn lại của mẫu CPI không bao gồm dữ liệu.

3.6 Dữ liệu giá dịch vụ viễn thông

Bắt đầu vào tháng 2 năm 2019, dựa trên dữ liệu khảo sát hộ gia đình, BLS đã đưa ra tỷ lệ lấy mẫu cho dịch vụ điện thoại cố định, dịch vụ truyền hình cáp và vệ tinh và dịch vụ internet để hỗ trợ các nhà kinh tế học trong việc lựa chọn nhiều mặt hàng đại diện hơn. Một bộ dữ liệu khác chứa giá niêm yết cho các dịch vụ viễn thông dân dụng được tổng hợp bởi một bộ dữ liệu tổng hợp từ một số kênh bán hàng. Để tính toán các chỉ số thử nghiệm, BLS đã sử dụng quyền số của cửa hàng CPI và phân bổ quyền số đó cho tất cả các mục trong tập dữ liệu như nhau. BLS đã phát triển các chỉ số mô hình phù hợp để nhân rộng phương pháp CPI. Có sự khác biệt nhiều về chỉ số giữa CPI và chỉ số thử nghiệm do các thủ tục đối với dữ liệu bị thiếu và thiếu phương pháp thay thế. Khó khăn trong việc xác định một mặt hàng duy nhất để định giá trong dữ liệu thay thế - điều tạo ra một mặt hàng duy nhất trong bộ dữ liệu không phải là cách BLS định nghĩa mặt hàng đó. Kết quả sơ bộ cho thấy việc tính toán CPI cho các dịch vụ Viễn thông dân dụng với dữ liệu thay thế là có thể và với các điều chỉnh về phương pháp cho phép truy cập vào bộ dữ liệu rộng hơn, phong phú hơn so với thu thập trường truyền thống.

3.7 Dữ liệu giá thực phẩm tiêu dùng hộ gia đình

Khoảng tám năm trước, BLS đã mua lịch sử dữ liệu máy quét Nielsen Scantrack và sử dụng dữ liệu để biên soạn các chỉ số giá. Dữ liệu bao gồm lịch sử dữ liệu năm năm kết thúc vào năm 2010 theo mã giá toàn cầu (UPC) / khu vực địa lý, và một số mô tả về sản phẩm và giá trung bình của mỗi quan sát. Dữ liệu Nielsen mà BLS mua không bao gồm toàn bộ phạm vi của các loại cửa hàng được nêu trong CPI cho các loại thực phẩm. Nó bỏ qua các cửa hàng tiện lợi, tiệm bánh, người bán thịt, cửa hàng tạp hóa nhỏ hơn, cửa hàng kho và trạm xăng[1]. Dữ liệu UPC của Nielsen phải được đối chiếu với bảng phân loại mặt hàng được sử dụng trong CPI. Khoảng 80% UPC có thể được so khớp trực tiếp vào danh mục CPI dựa trên phân loại Nielsen, nhưng 20% còn lại phải được khớp thủ công.

3.8 Dữ liệu giá thuê nhà ở

Khảo sát giá thuê nhà ở trong CPI thu thập được dữ liệu giá thuê của khoảng 47.000 quan sát được chọn trong mẫu đại diện của thị trường cho thuê nhà tư nhân. BLS đã khám phá một bộ dữ liệu thứ cấp về giá thuê nhà ở và giá thuê ước tính để đánh giá tiềm năng thay thế hoặc bổ sung dữ liệu khảo sát Nhà ở CPI. Bộ dữ liệu nguồn thứ cấp không được thiết kế dưới dạng mẫu đại diện hoặc điều tra dân số cho khu vực địa lý và mặc dù nó bao gồm giá thuê và giá thuê ước tính cho hơn 50 triệu đơn vị nhà ở, tỷ lệ trùng khớp với các đơn vị CPI chỉ khoảng 30%. Trường hợp có thể đối chiếu, BLS khớp các đơn vị trong mẫu CPI với cùng các đơn vị trong bộ dữ liệu và chỉ số được tính toán. Trong phân tích cuối cùng, BLS đã quyết định rằng sự khác biệt giữa CPI Nhà ở và bộ dữ liệu nguồn thứ cấp là rất quan trọng và các mục đích khác nhau để cho phép sử dụng nguồn thứ cấp này trong CPI tại thời điểm này.

3.9 Dữ liệu nguồn nhiên liệu động cơ

GasBuddy là một doanh nghiệp công nghệ chuyên thu thập nguồn nhiên liệu từ gần 100.000 trạm xăng ở Mỹ. Cơ quan thống kê đã có được sự cho phép của GasBuddy đối với dữ liệu trích xuất từ trang web của họ và thừa nhận chúng là nguồn dữ liệu để tính CPI. Các chỉ số dựa trên dữ liệu của GasBuddy theo dõi chặt chẽ các chỉ số giá xăng CPI mặc dù có sự khác biệt về quyền số và thông tin chi tiết về giá cả.

Phần lớn công việc nghiên cứu đã tập trung vào các so sánh giữa việc thu thập dữ liệu hiện tại của CPI cho nhiên liệu động cơ và dữ liệu được trích xuất trên web. Không giống như hầu hết các mặt hàng khác trong CPI có các loại mặt hàng riêng lẻ được lấy mẫu, cả năm loại nhiên liệu động cơ được chọn tự động tại bất kỳ nhà bán lẻ nhiên liệu động cơ nào nằm trong mẫu. Trong số năm loại nhiên liệu động cơ trong CPI, thông tin GasBuddy có thể thay thế dữ liệu thu thập được cho ba loại xăng và dầu diesel, nhưng chúng không có phạm vi đảm bảo của nhiên liệu thay thế. Hiện nay, một số trạm xăng thực sự cung cấp nhiên liệu động cơ thay thế (như sạc điện, ethanol, E85 hoặc diesel sinh học), vì vậy các quan sát về các lựa chọn thay thế nhiên liệu động cơ có thể được thu thập bình thường và đưa vào dữ liệu được quét trên web.

Kết quả của cho thấy giá trung bình và chỉ số giá dựa trên dữ liệu của GasBuddy và CPI hoạt động rất giống nhau.

3.10 Giá vé máy bay

[1] *Nielsen cung cấp dữ liệu cho các cửa hàng tiện lợi, cửa hàng kho và trạm xăng nhưng BLS đã chọn không mua dữ liệu đó trong dự án nghiên cứu ban đầu này.*

Hiện nay việc thu thập giá vé của các hãng hàng không được thực hiện bởi các điều tra viên, họ thu thập giá từ các trang web của người trả lời. Giá vé trên web cho phép những người làm thống kê giá CPI theo dõi chuyến đi được xác định theo từng tháng, trong đó giá được thu thập bằng cách chỉ định từng thông số kỹ thuật cố định cho giá vé một chiều hoặc khứ hồi, xuất phát và thành phố đích, ngày khởi hành và trở về, hạng vé của vé, tình trạng đặt trước, và ngày trong tuần. Mỗi tháng, thông tin chi tiết về vé đặt trước và chi tiết trong tuần là tương tự nhau sẽ được sử dụng để thu thập giá. Phương pháp này cho phép CPI định giá một chuyến đi được xác định nhất quán mỗi tháng ngoài việc mô phỏng chính xác cách người tiêu dùng đặt vé máy bay.

Trước mắt, BLS đang nghiên cứu việc sử dụng phương pháp đối chiếu và thay thế, nghĩa là thu thập giá cho từng mục hiện có trong mẫu dựa trên các mô tả về thông số kỹ thuật và mức giá đang được sử dụng trong mẫu vé máy bay. BLS cũng đã bắt đầu tiếp cận với những người trả lời về dữ liệu được báo cáo cụ thể, cho phép sử dụng mã nguồn mở của họ hoặc cho phép trích xuất dữ liệu.

4. Kết luận và các kế hoạch trong tương lai

Dữ liệu lớn có thể cung cấp thông tin một cách kịp thời hơn phương pháp thống kê giá truyền thống. Các nguồn dữ liệu thay thế mới có khả năng giải quyết nhiều vấn đề chúng ta đã gặp bao gồm tỷ lệ phản hồi thấp hơn và chi phí thu thập cao. Sau nhiều năm làm việc trên nhiều nguồn dữ liệu thay thế khác nhau, BLS hiện đã phác thảo một tầm nhìn CPI sẽ như thế nào trong thập kỷ tới. Điều này bao gồm các mục tiêu thay thế một phần đáng kể cho dữ liệu được thu thập trực tiếp theo phương pháp điều tra bởi các nguồn dữ liệu thay thế trong 5 năm tới. Cách tiếp cận của BLS là ưu tiên dữ liệu thay thế cho các danh mục và cửa hàng dựa trên một số yếu tố gồm: mức độ quan trọng của mặt hàng, số lượng giá thay thế, chi phí thu thập, chi phí dữ liệu thay thế, độ chính xác của chỉ số mặt hàng hiện tại, mối quan hệ của người trả lời với BLS, khả năng dễ dàng thực hiện, tỷ lệ phản hồi và mức tập trung của mẫu cho một mặt hàng nhất định. BLS cũng sẽ ưu tiên hợp tác để thu thập dữ liệu của doanh nghiệp lớn và sẽ khám phá các nguồn dữ liệu thay thế mang lại lợi ích và độ chính xác cao hơn.

Mặc dù dữ liệu thay thế cho phép khám phá nhiều cải tiến về phương pháp, nhưng kinh nghiệm của BLS cho đến nay cho thấy có một số vấn đề cơ bản cần giải quyết. Các kỹ thuật đơn giản như mô hình chỉ số giá đối chiếu phù hợp không phải lúc nào cũng cho ra các kết quả có thể sử dụng được, và các phương pháp CPI hiện tại có thể không phù hợp với dữ liệu giao dịch. BLS đã phát triển các cách giải quyết vòng đời sản phẩm với các chỉ số mới sẽ sớm được công bố trên cơ sở thử nghiệm. Giải pháp ngắn hạn cho phép BLS thay thế việc thu thập dữ liệu giá thủ công từ trang web của doanh nghiệp bằng bộ dữ liệu giao dịch.

BLS sẽ tiếp tục xem xét các tài liệu mới nhất về các phương pháp chỉ số giá, đồng thời phát triển thêm các phương pháp và quy trình mới để tận dụng các nguồn dữ liệu thay thế. Ngoài ra BLS sẽ tiếp tục giới thiệu dữ liệu thay thế trong CPI, đồng thời tiếp tục chú ý đến các mục tiêu đo lường CPI cốt lõi và đáp ứng nhu cầu của cơ sở dữ liệu rộng rãi của chương trình.

Minh Ánh (dịch)

Nguồn: https://www.nber.org/chapters/c14280.pdf