DỮ LIỆU MÁY QUÉT

Tóm tắt:

Dữ liệu máy quét mang lại nhiều cơ hội trong việc cải thiện tính chính xác cho số liệu chỉ số giá tiêu dùng (CPI). Sử dụng dữ liệu máy quét giúp ích cho việc cải thiện tính chính xác của dữ liệu giá tiêu dùng được sử dụng trong tính toán CPI thông qua tính các giá trị đơn vị của các sản phẩm đồng nhất, đồng thời giúp cải thiện cỡ mẫu cho điều tra giá tiêu dùng. Bài viết đưa ra những cơ hội và thách thức chính của dữ liệu máy quét trong việc tính toán CPI.

1. Giới thiệu

Dữ liệu máy quét (Scanner data) mang lại nhiều cơ hội trong việc cải thiện tính chính xác cho số liệu CPI. Các bộ dữ liệu máy quét cũng bao gồm toàn bộ danh mục và số lượng mặt hàng được bán bởi các nhà bán lẻ tại các điểm bán hàng. Sử dụng dữ liệu máy quét giúp cho việc cải thiện tính chính xác của dữ liệu giá tiêu dùng được sử dụng trong tính toán CPI thông qua tính các giá trị đơn vị của các sản phẩm đồng nhất, đồng thời giúp cải thiện cỡ mẫu cho điều tra giá tiêu dùng. Bên cạnh đó, dữ liệu máy quét còn cung cấp các thông tin về doanh thu/số lượng bán hàng giúp cho việc tính toán quyền số trở lên tốt hơn. Bên cạnh những cơ hội mà máy quét đem lại cũng có nhiều thách thức cần được các cơ quan thống kê quốc gia (NSO) giải quyết trước khi sử dụng trong việc biên soạn CPI.

Những nội dung dưới đây sẽ chỉ ra những cơ hội và thách thức chính của dữ liệu máy quét, các cân nhắc cơ bản mang tính thực tiễn, đồng thời bài viết cũng sẽ đưa ra các giải pháp và lời khuyên trong việc sử dụng dữ liệu máy quét để tính CPI.

2. Thu thập dữ liệu máy quét

Dữ liệu máy quét đã tồn tại một vài thập kỷ nên rất có giá trị trong việc tính toán các chỉ tiêu thống kê theo thời gian. Một trong những thách thức đối với NSO là thu thập các tập dữ liệu máy quét. Có hai lựa chọn mang tính khả thi, là NSO có thể tìm kiếm dữ liệu máy quét từ các doanh nghiệp bán lẻ hoặc từ một nhà cung cấp dữ liệu thứ ba. Cả hai lựa chọn trên đều mang lại lợi ích cũng như thách thức.

Một số NSO đã tiến hành thành công các cuộc đàm phán về việc chia sẻ dữ liệu máy quét với các doanh nghiệp bán lẻ và đã sử dụng các dữ liệu được cung cấp để biên soạn CPI[1]. Việc thu thập dữ liệu trực tiếp từ các cửa hàng doanh nghiệp bán lẻ có một số lợi ích tiềm năng có thể có trong phương án đàm phán:

- Việc cung cấp các tập dữ liệu không mất phí (hoặc rất ít chi phí);
- Phạm vi các mặt hàng trong tập dữ liệu;
- Thời gian cung ứng dữ liệu đáp ứng tính kịp thời tính toán CPI;
- Mức độ tích hợp của các mặt hàng để đảm bảo tính đồng nhất của thông tin;

[1] *Úc, Hà Lan, New Zealand, Thụy Điển và Thụy Sĩ là các cơ quan thống kê quốc gia sử dụng dữ liệu máy quét để tính CPI. Danh sách đầy đủ các quốc gia sử dụng dữ liệu máy quét mã vạch nêu trong phụ lục A.*

- Đảm bảo tính kịp thời;
- Xác định chính xác người nắm giữ các tập dữ liệu trong doanh nghiệp và tiến hành liên lạc, trao đổi trực tiếp với họ;

Thu thập dữ liệu máy quét thông qua đàm phán trực tiếp với các doanh nghiệp cũng có những thách thức nhất định. Thách thức cơ bản nhất là cuộc đàm phán song phương đòi hỏi nhiều nỗ lực giữa các bên. Kinh nghiệm của Hà Lan, Thụy Điển và Thụy Sĩ cho thấy mất tới 6 tháng để đạt được các thoả thuận. Nội dung đàm phán liên quan đến nhiều lĩnh vực: Từ hệ thống công nghệ thông tin đến các mối quan tâm về việc bảo mật. Một số thỏa thuận đạt được giữa cơ quan thống kê và các doanh nghiệp được cam kết dưới dạng biên bản ghi nhớ (hoặc tương tự)[2].

Một số cơ quan thống kê tiếp cận dữ liệu máy quét theo cách khác thông qua các công ty trung gian hoặc các công ty nghiên cứu thị trường như Nielsen và GfK. Lợi ích cơ bản của cách tiếp cận này là chỉ cần đàm phán với một số ít các nhà cung cấp dữ liệu, các cơ quan thống kê đã có thể tiếp cận với nhiều nguồn dữ liệu của nhiều nhà cung ứng khác nhau.

Dữ liệu máy quét mã vạch có được từ các doanh nghiệp cho thấy một số thách thức. Nhìn chung dữ liệu này được các NSO mua lại. Chi phí được bù đắp bằng việc giảm thiểu chi phí thu thập dữ liệu như phương pháp thu thập dữ liệu truyền thống, đó là cử điều tra viên tới từng cửa hàng bán lẻ thu thập giá bán, trong khi đó dữ liệu máy quét luôn được giấu kín.

Kinh nghiệm của các cơ quan thống kê quốc gia trong việc sử dụng dữ liệu máy quét mã vạch để tính toán CPI cho thấy cách thu thập dữ liệu trực tiếp từ các nhà bán lẻ thường được yêu thích hơn vì các lý do như đã trình bày. Tuy nhiên, tiếp cận dữ liệu từ các công ty nghiên cứu thị trường sẽ hữu ích hơn trong trường hợp các dữ liệu máy quét mã vạch không đảm bảo hoặc các nguồn dữ liệu không có sẵn trong việc đàm phán cung cấp dữ liệu song phương.

3. Truy cập và chuẩn bị dữ liệu máy quét mã vạch để sử dụng

Nếu như các cơ quan thống kê quốc gia đã thành công trong việc tiếp cận các tập dữ liệu scanner thì thách thức tiếp theo đối với các cơ quan thống kê này là làm sao chuyển đổi các tập dữ liệu đó thành các thông tin hữu ích và có thể sử dụng để tính toán chỉ số giá tiêu dùng CPI. Để đạt được các mục tiêu trên, các cơ quan thống kê quốc gia cần vượt qua một số thách thức sau.

3.1. Phát triển hệ thống công nghệ thông tin (IT)

Dữ liệu máy quét với các đặc điểm của nó còn được gọi là dữ liệu lớn. Các NSO cần phải có một hệ thống máy tính/IT có thể đáp ứng việc lưu trữ, xử lý nguồn dữ liệu lớn này nếu muốn sử dụng các thông tin để tính CPI. Hệ thống IT cần đáp ứng và xử lý được các tập dữ liệu có cấu trúc, định dạng, nội dung khác nhau do các doanh nghiệp bán lẻ (và các nhà cung cấp dữ liệu trung gian) thường xây dựng các hệ thống phục vụ cho báo cáo trong nội bộ. Đây có thể là thách thức đối với các cơ quan thống kê cũng như yêu cầu về việc phát triển các nguồn lực IT đòi hỏi nhiều chi phí về thời gian và tiền bạc. Một số NSO đã đưa ra các tài liệu về những thách thúc này (Bird et al., 2014; Böttcher and Sergeev, 2014). Giải pháp cụ thể phụ thuộc vào điều kiện của từng quốc gia.

[2] *Biên bản ghi nhớ là các quy định và cam kết bắt buộc của mỗi bên nhằm đảm bảo cho việc cung cấp dữ liệu máy quét cho các cơ quan thống kê diễn ra liên tục và đảm bảo tính kịp thời.*

Như vậy, rõ ràng các cơ quan thống kê cần xây dựng một hệ thống IT phù hợp mới có thể sử dụng dữ liệu máy quét mã vạch để biên soạn CPI, bất kể nhà cung cấp dữ liệu là ai.

3.2. Phân loại dữ liệu máy quét

Các tập dữ liệu máy quét của các cửa hàng bán lẻ thường có cách phân loại khác nhau và độc lập, cơ quan thống kê sẽ nhận được các tập thông tin phân loại mặt hàng khác nhau, việc phân loại cần liên kết với bảng phân loại danh mục các mặt hàng thuộc rổ hàng hóa tiêu dùng. Phân loại các tập dữ liệu máy quét chiếm nguồn lực đáng kể tại NSO. Vì vậy, NSO cần đầu tư nhiều cho công tác phân loại dữ liệu cơ sở mà họ nhận được; tuy nhiên cũng cần có nguồn lực để phân loại những mặt hàng mới xuất hiện trong tập dữ liệu.

Thách thức trong việc phân loại các mặt hàng trong tập dữ liệu máy quét theo phân loại mặt hàng trong rổ hàng hóa tiêu dùng hiện nay đã và đang được thực hiện bởi NSO theo nhiều cách. Tất cả các cơ quan thống kê đều đang cố gắng tìm ra giải pháp phù hợp trong hoàn cảnh cụ thể tại đất nước của họ. Chẳng hạn, cơ quan thống kê Thụy Sĩ đã tiến hành phân loại danh mục mặt hàng dữ liệu máy quét theo danh mục rổ hàng CPI bằng cách mua lại siêu dữ liệu nghiên cứu thị trường (Muller, 2010). Cơ quan Thống kê Hà Lan đã kết hợp phân loại danh mục mặt hàng máy quét theo danh mục được cung cấp với các thông tin nghiên cứu thị trường hình thành một quy trình xử lý (de Haan et al., 2010). Một số cơ quan thống kê khác, vì nhiều lý do, tự phân loại danh mục mặt hàng máy quét theo danh mục CPI (Howard et al, 2015).

Thách thức của việc phân loại danh mục hàng hóa từ dữ liệu máy quét theo rổ hàng truyền thống tăng lên khi các tập dữ liệu máy quét được bảo mật trực tiếp bởi các doanh nghiệp. Việc đàm phán với các công ty nghiên cứu thị trường cho phép NSO tiếp cận trực tiếp dữ liệu máy quét đã được phân loại theo danh mục rổ hàng truyền thống của các cơ quan thống kê của mình. Quan sát của một số cơ quan thống kê quốc gia nhận thấy lợi ích thực tế từ việc thu thập dữ liệu máy quét từ các công ty nghiên cứu thị trường.

3.3. Đảm bảo chất lượng của các tập dữ liệu máy quét

Dữ liệu máy quét là một nguồn dữ liệu mới có thể sử dụng trong việc biên soạn CPI. Trong trường hợp xuất hiện sự thay đổi trong nguồn dữ liệu, người tính toán các chỉ tiêu thống kê nên tiến hành một loạt các phép kiểm tra nhằm đảm bảo nguồn dữ liệu mới cung cấp đúng những yêu cầu cơ sở đối với việc sản xuất các số liệu phục vụ cho mục đích thống kê. Việc kiểm tra dữ liệu máy quét được chia thành hai loại, kiểm tra *tổng quát* và kiểm tra *chi tiết.*

Kiểm tra *tổng quát* liên quan đến việc đo lường mở rộng và thường được áp dụng khi các cơ quan thống kê bắt đầu nhận được dữ liệu. Loại kiểm tra này nhằm đảm bảo dữ liệu mà các cơ quan thống kê nhận được khớp với dữ liệu mà họ đã nhận được trước đó. Việc kiểm tra có thể liên quan đến định dạng tập dữ liệu, tổng số mặt hàng trong tập dữ liệu, và tổng doanh thu bán hàng. Kiểm tra toàn bộ có thể giúp phát hiện những lỗi điển hình của tập dữ liệu.

Kiểm tra *chi tiết* thường được áp dụng ở cấp sản phẩm hoặc nhóm sản phẩm. Việc kiểm tra này nhằm phát hiện những thay đổi nổi bật trong doanh số bán hàng, doanh thu và giá của các sản phẩm trong tập dữ liệu. Kiểm tra chi tiết thường liên quan đến công tác biên tập dữ liệu giá.

Cả kiểm tra tổng quát và kiểm tra chi tiết nên thực hiện tự động và báo cáo lại cho các nhân viên thống kê. Việc kiểm tra có thể cần sự tương tác với bên cung cấp dữ liệu, cũng như tham chiếu với các nguồn thông tin giá thay thế (như các tờ rơi quảng cáo hay giá tiêu dùng online).

2.3.4 Mức độ chi tiết dữ liệu sản phẩm phục vụ cho việc biên soạn CPI

Dữ liệu máy quét có thể cung cấp cho người sử dụng dữ liệu giá của các sản phẩm tương đồng. Điều này rất quan trọng vì nó đảm bảo cho việc liệu CPI có phản ánh đúng xu hướng giá tiêu dùng thay đổi theo thời gian hay không. Những thay đổi trong kết cấu sản phẩm đã bán và chất lượng của chúng thực tế không được phản ánh trong CPINSO nên tập trung vào vấn đề này như một phần trong nội dung đàm phán nhằm đảm bảo chất lượng cho dữ liệu máy quét nhận được từ các nhà cung cấp. Nội dung trên được thảo luận chi tiết trong mục 4.3 dưới đây.

4. Thực hiện - từ tranh luận đến các phương pháp mới

4.1. Cơ hội và thách thức của việc sử dụng dữ liệu máy quét

Việc sử dụng các thông tin trong dữ liệu máy quét để biên soạn chỉ số giá tiêu dùng CPI có thể mang lại sự thay đổi đáng kể trong công tác thu thập dữ liệu truyền thống hiện nay thường được thực hiện bởi các nhân viên của NSO. Điều này cho thấy sự thay đổi cần được giám sát cẩn thận, cả ảnh hưởng trong các hoạt động thống kê lẫn quan hệ với người sử dụng và các bên liên quan chính.

Kinh nghiệm thu thập dữ liệu máy quét và các phương pháp chỉ số giúp cho việc sử dụng các thông tin trong dữ liệu máy quét trở lên tốt hơn. Dữ liệu máy quét có thể giúp tăng cường độ chính xác cho CPI theo những cách tốn ít chi phí hơn. Các tập dữ liệu máy quét có thể được sử dụng để: (i) Đối chiếu dữ liệu; (ii) Thay thế giá thu thập theo phương pháp truyền thống; (iii) Mở rộng kích thước mẫu; (iv) Quyền số sản phẩm mức thấp nhất trong CPI và phản ánh mức độ quan trọng của chỉ số này trong nền kinh tế; (v) Thực hiện các phương pháp mới[3] nắm bắt được đặc tính chỉ số giá và cho phép tự động hóa quy trình.

Việc cải tiến được liệt kê ở trên nhằm cải thiện tính chính xác của CPI. Điều này sẽ giải thích ở dưới đây. Một số cơ quan thống kê sử dụng dữ liệu máy quét để đạt được các mục tiêu (i), (ii) và (iii). Trong khi những sự tăng cường này có ý nghĩa, thực hiện các mục tiêu (iv) và (v) sẽ tối đa việc sử dụng dữ liệu máy quét để cải thiện chất lượng CPI. Lưu ý, một số NSO đã thực hiện lần lượt từng mục tiêu được liệt kê ở trên (ABS, 2017) trong khi các cơ quan thống kê khác chuyển từ mục tiêu (i) sang (v) (Krsinich, 2015; Chessa, 2016), cả hai cách tiếp cận đều khả thi và thường phản ánh tình hình cụ thể tại từng khu vực của cơ quan thống kê đó.

Năm mục tiếp theo mô tả lợi ích của việc sử dụng dữ liệu máy quét cho mỗi mục tiêu được liệt kê ở trên.

4.2. Sử dụng dữ liệu máy quét cho việc đối chiếu và đảm bảo chất lượng dữ liệu

NSO có thể sử dụng dữ liệu máy quét để đối chiếu số liệu và kiểm soát chất lượng dữ liệu CPI hiện nay.

Dữ liệu máy quét bao gồm số lượng và doanh thu của các sản phẩm được cung cấp bởi các nhà bán lẻ những mặt hàng này trong các khoảng thời gian xác định, thường là tuần hoặc tháng. Các thông tin này cho

[3] *Xem mục 3 của phụ lục để biết thêm chi tiết về các phương pháp này*

phép NSO tính giá từng sản phẩm riêng biệt bằng cách lấy doanh thu chia cho số lượng sản phẩm đã bán. Giá này liên quan đến *giá đơn vị* và đại diện cho mức giá trung bình mà người mua phải trả trong khoảng thời gian tuần hoặc tháng.

Đối với các sản phẩm tương đồng, giá đơn vị theo thời kỳ phản ánh được giá mà người mua phải trả sẽ chính xác hơn giá thời điểm (Balk, 1998)[4]. Giá đơn vị đã bao gồm giảm giá và ảnh hưởng của sự giảm giá này đến số lượng sản phẩm được bán ra. Việc xác định thời kỳ cho giá đơn vị được tính toán rất quan trọng vì nó đảm bảo tính chính xác của loại giá này. Diewert, Fox và de Haan (2016) đã phát hiện ra giá đơn vị được sử dụng trong việc tính CPI nên có cùng thời kỳ với các chỉ tiêu được tính toán, thay vì lấy giá thời kỳ trước đó. Tiếp cận gần nhất có thể dẫn tới xu thế tăng độ chệch đối với CPI.

Phân tích giá tiêu dùng cho phép so sánh giá tiêu dùng được thu thập với giá được tính từ dữ liệu máy quét. Những phân tích này cho biết một số giá trị chệch tiềm ẩn của giá được sử dụng để tính CPI tại thời điểm thu thập so với với đơn vị. Phân tích doanh thu và số lượng bán sản phẩm được sử dụng bởi các chuyên gia phân tích giá tại các cơ quan thống kê nhằm trả lời cho câu hỏi liệu cỡ mẫu CPI hiện nay có thể được cải thiện hay không.

4.3. Sử dụng dữ liệu máy quét thay thế giá truyền thống

Tại hầu hết các quốc gia phần lớn giá được sử dụng để tính toán CPI được thu thập bởi các điều tra viên bằng việc thu thập dữ liệu trực tiếp tại các cơ sở kinh doanh. Điều tra viên của các cơ quan thống kê quốc gia sẽ trực tiếp quan sát, thu thập giá bán tại các cửa hàng tại một thời điểm xác định, cũng như thảo luận trực tiếp về việc giảm giá, các chương trình quà tặng đặc biệt và các mặt hàng bán chạy với chủ cửa hàng. Điều tra viên sẽ ghi chép các thông tin này trong buổi phỏng vấn, sau đó nhập dữ liệu vào máy vi tính một cách thủ công. Việc tiếp cận địa bàn đều đặn giúp cơ quan thống kê nắm bắt được sự biến động của thị trường một cách chủ động và quan sát được sự thay đổi về chất lượng sản phẩm.

Sử dụng dữ liệu máy quét để thay thế giá thu thập theo phương pháp truyền thống nhìn chung giúp tiết kiệm các nguồn lực cho cơ quan thống kê quốc gia. Lý do bởi nhân viên thống kê không cần tới cơ sở kinh doanh để thu thập giá bán. Mức tiết kiệm được ảnh hưởng bởi số lượng nhân viên được giảm bớt và số lượng nguồn lực tăng cường tại cơ quan thống kê phục vụ cho việc quản lý và sử dụng dữ liệu máy quét.

Giá thu thập sử dụng cho việc thay thế cũng có một số thách thức cần được quản lý.

Để tính giá đơn vị cần sử dụng các mặt hàng đồng nhất, những mặt hàng này có đặc tính ổn định theo thời gian vì sự thay đổi trong thành phần mặt hàng và chất lượng mặt hàng sẽ không được phản ánh trong sự thay đổi giá bán (ILO, 2004, p.164). Những yêu cầu này cho thấy một số thách thức khi thay thế giá bán được thu thập bằng những thông tin lấy từ các tập dữ liệu máy quét. Việc thỏa thuận giữa NSO và bên cung cấp dữ liệu cần xác định rõ mức độ phù hợp của các nhóm sản phẩm (hoặc các sản phẩm không theo nhóm) nhằm đảm bảo việc các sản phẩm được cung cấp đáp ứng được các tiêu chuẩn mà giá đơn vị yêu cầu, từ đó mới có thể sử dụng để tính CPI.

[4] *Dữ liệu về doanh thu có thể không hoàn toàn khớp với mục tiêu và nội dung của CPI quốc gia vì nó có thể bao gồm cả chi tiêu từ các hộ dân cư không thường trú và các doanh nghiệp (Fenwick, 2014).*

Một số NSO đã có kinh nghiệm trong việc sản xuất giá đơn vị từ dữ liệu giá lấy từ các tập dữ liệu máy quét. Tại một số quốc gia việc sử dụng đơn vị phân loại hàng hóa tồn kho (SKU) được chứng minh là thành công (Howard et al, 2015), trong khi việc sử dụng Mã phân loại sản phẩm toàn cầu (GTIN) và Mã vạch sản phẩm châu Âu (EAN) có thể chưa đáp ứng được mức độ chi tiết, phân biệt sản phẩm theo đặc tính, như là kiện hàng, điều này được cân nhắc là không liên quan đến người tiêu dùng. Trong khi chính dữ liệu chi tiết này mới là dữ liệu phản ánh mức độ đồng nhất của sản phẩm, vấn đề các mặt hàng biến mất hoặc tái xuất hiện thường xảy ra và thường làm cho việc tính toán chỉ số giá tiêu dùng trở lên khó khăn hơn[5].

Điều cốt lõi trong việc đo lường giá cần phải tính đến thay đổi chất lượng và chỉ ra các mặt hàng mới (ILO, 2004). Điều này được hầu hết các cơ quan thống kê giải quyết khi điều tra viên tới cửa hàng để thu thập sự thay đổi giá bán của các mặt hàng cụ thể hoặc các mặt hàng tương đương trong các giai đoạn tiếp theo, và xác định các mặt hàng mới. Do tính chất của các mặt hàng là có thể thay thế nhau nên các điều tra viên thống kê đã tiến hành thu thập các thông tin mô tả, những thông tin này cho biết ảnh hưởng của sự thay đổi chất lượng được phân chia theo sự thay đổi giá, vì vậy CPI có thể tính đến sự thay đổi thuần túy của giá.

Tính toán việc thay đổi quy cách sản phẩm là một thách thức điển hình khi sử dụng dữ liệu máy quét. Dữ liệu máy quét có xu hướng cho thấy mức biến động của mặt hàng khá cao từ tháng này sang tháng khác. Có những mẫu mã sản phẩm mới (cũng như các phiên bản mới) hiện trên thị trường và các mẫu mã cũ biến mất khỏi thị trường vì chúng đã bị thay thế (xuất hiện các mặt hàng xuất thay đổi mẫu mã mới và các mặt hàng cũ biến mất khỏi thị trường do bị thay thế). Việc tính giá cho các mặt hàng điều chỉnh quy cách vì vậy trở lên khó khăn.

Có ba kịch bản cho giá của các mặt hàng điều chỉnh quy cách từ dữ liệu máy quét, gồm:

a. Trường hợp các sản phẩm mới được bán với giá đã thu thập trong mẫu, kể cả mặt hàng thay thế

b. Trường hợp khối lượng mặt hàng thay đổi (ví dụ thay đổi về khối lượng đóng gói) và quy cách mặt hàng thay đổi

c. Trường hợp khối lượng mặt hàng thay đổi nhưng quy cách mặt hàng không thay đổi

Kịch bản đầu tiên là trường hợp đơn giản nhất và đòi hỏi tính giá ở giai đoạn trước đó cho sản phẩm mới.

Đối với kịch bản 2 và 3, nhân tố điều chỉnh quy cách được xem xét là sự thay đổi về khối lượng. Các cơ quan thống kê cần xây dựng phương pháp liên kết với các sản phẩm mới xuất hiện hoặc biến mất. Ví dụ, nếu một sản phẩm thay đổi về kích thước đóng gói, quy trình liên kết có thể sử dụng thông tin mô tả về sản phẩm, giá, doanh thu, thời gian (khi các sản phẩm xuất hiện hoặc biến mất trong danh sách các mặt hàng bày bán) và số lượng bán. Quy trình xác định, sản phẩm mới gần như có thể thay thế cho các sản phẩm đã biến mất (nhưng với quy cách phẩm cấp sản phẩm khác nhau). Việc điều chỉnh quy cách sau đó được thực hiện bởi các nhà phân tích giá dựa trên mô tả sản phẩm.

Dữ liệu máy quét có khối lượng lớn và đa dạng về cấu trúc cũng như kiểu định dạng

[5] *Ví dụ, khi sử dụng mã vạch để xác định một mặt hàng, sự thay đổi giá của sản phẩm đồng nhất, mã vạch của các sản đó thay đổi cùng thời điểm sẽ không thể đo lường được.*

đối với mỗi cửa hàng bán lẻ. Kết quả NSO cần nhiều nguồn lực để chuyển đổi các tập dữ liệu thô ban đầu thành cơ sở dữ liệu phù hợp với việc phân tích và tính CPI (Bird et al., 2014; Böttcher and Sergeev, 2014). Lưu trữ, làm sạch và mã hóa dữ liệu máy quét cũng là những thách thức lớn cần được cân nhắc kỹ bởi các NSO.

4.4. Sử dụng dữ liệu máy quét để cập nhật mẫu điều tra giá

Việc thu thập các điểm giá theo phương pháp truyền thống bằng việc thu thập giá bán lẻ trực tiếp tại các cửa hàng bán lẻ là rất tốn nguồn lực. Việc thu thập toàn bộ giá bán các mặt hàng mỗi kỳ là không thực tế, vì vậy cần tiếp cận giá bán thông qua điều tra chọn mẫu. Ví dụ, các sản phẩm trong rổ hàng CPI thu thập bởi các điều tra viên thuộc các NSO được thực hiện thông qua hình thức điều tra chọn mẫu. Điều tra viên chính là những người trực tiếp hỏi người bán mặt hàng nào được bán với số lượng lớn, và trực tiếp kiểm tra kệ hàng bày bán sản phẩm để đưa ra quyết định về mức độ quan trọng tương ứng của loại mặt hàng đó. Mục đích của điều tra viên là thu thập giá bán của các mặt hàng đại diện thuộc rổ hàng hóa. Đây cũng chính là mục tiêu của cuộc điều tra chọn mẫu. Chọn mẫu có mục đích là phương pháp vẫn được sử dụng từ trước đến nay vì dàn mẫu các mặt hàng được bán không có sẵn đồng thời thiếu dữ liệu về số lượng bán, doanh thu bán hàng, những thông tin được sử dụng trong việc đo lường mức độ quan trọng của mặt hàng trong nền kinh tế. Tuy nhiên, việc chọn mẫu có mục đích có thể gây ra sự chệch khi các mặt hàng được chọn không đủ tính đại diện cho tổng thể các mặt hàng.

Chọn mẫu truyền thống có thể được thay thế bởi nhiều phương pháp chọn mẫu khác do sự sẵn có của dữ liệu máy quét. Dữ liệu máy quét có thể được sử dụng như dàn mẫu để cập nhật mẫu giá. Một mẫu giá thường bao gồm 2 chiều là kết hợp của một mẫu các cửa hàng và mẫu danh mục các mặt hàng. Nếu toàn bộ các cửa hàng trong chuỗi bán lẻ được tiếp cận, dữ liệu thu được có thể được sử dụng làm dàn mẫu cho các cửa hàng và danh mục sản phẩm.

Tỷ lệ doanh thu của từng mặt hàng (hoặc sự kết hợp sản phẩm/cửa hàng) có thể được xác định cụ thể đối với mỗi mặt hàng trong nhóm. Các sản phẩm được lựa chọn để thu thập dữ liệu nằm trong rổ hàng hóa CPI căn cứ vào tỷ lệ doanh thu lấy từ mẫu hoặc điểm cắt mẫu (de Haan, Opperdoes and Schut, 1999).

Tuy nhiên, theo thời gian các sản phẩm trong mẫu có thể biến mất hoặc ngừng bán. Trong trường hợp thay thế sản phẩm cần duy trì sự liên quan đến mẫu. Kiểm tra tương quan có thể được sử dụng để phát hiện mặt hàng nào trong mẫu không phù hợp và đánh giá mức độ phù hợp của các mặt hàng được sử dụng làm mặt hàng thay thế.

Nguyên tắc cơ bản của phép kiểm định tính liên quan là tỷ suất doanh thu của các sản phẩm phải ổn định (ví dụ tỷ suất doanh thu cố định để so sánh với các sản phẩm khác) trong nhóm hàng hóa CPI. Những nhóm hàng này được liên quan đến giá sơ cấp "EA" (Elementary Aggregate) trong CPI (Chapter 20 of ILO, 2004). Tỷ suất doanh thu ổn định là điều thực sự quan trọng, vì có các mặt hàng có được bán rộng rãi trên thị trường do đây là mặt hàng mới lạ hoặc đang được giảm giá, nhưng sau một thời gian doanh thu lại không đáng kể. Do vậy những mặt hàng như thế này không thể là mặt hàng đại diện cho thị trường.

Để giải quyết vấn đề này, yêu cầu đặt ra đối với mặt hàng thay thế là doanh thu của

mặt hàng phải ổn định và cụ thể tại những khoảng thời gian nhất định (ví dụ khoảng thời gian từ 3 đến 6 tháng) trước khi chúng được coi như là một phần của mẫu giá. Các chuyên gia phân tích CPI nên kiểm tra một cách thủ công toàn bộ các mặt hàng thay thế được chọn và các mặt hàng chọn từ một danh sách sắp xếp theo doanh thu hàng tháng những tháng trước đó.

Các mặt hàng thực phẩm và đồ dùng gia đình rất đa dạng về chủng loại sản phẩm đều có thể là các mặt hàng tương đồng nếu không xác định được sự biến động giá của chúng. Chẳng hạn, cùng một nhãn hiệu cá ngừ đóng hộp có nhiều loại hương vị khác nhau và người tổng hợp CPI sẽ nhận ra giá của những hộp cá ngừ có mùi vị khác nhau của cùng một hãng là tương tự nhau, chúng được bán ở cùng thời điểm và thay đổi giá cũng cùng thời điểm. Việc chỉ đưa một loại hương vị vào mẫu vẫn sẽ đảm bảo tính đại diện cho tỷ lệ biến động giá trên thị trường.

Quy trình chọn mẫu phải đảm bảo các sản phẩm được chọn là sản phẩm đại diện. Các mặt hàng thay thế cần được chuyên gia lựa chọn thủ công từ danh sách xếp hạng của các sản phẩm tiềm năng và đáp ứng được các tiêu chuẩn bắt buộc. Mẫu tiếp cận từ dữ liệu máy quét đòi hỏi cần thêm nhiều nguồn lực phân tích CPI, tuy nhiên bù lại thì số lượng nhân lực thu thập dữ liệu sẽ được giảm bớt.

4.5. Sử dụng dữ liệu máy quét cập nhật cấu trúc chỉ số và áp dụng các quyền số

Các mẫu giá truyền thống thường nhỏ. Nguồn nhân lực phân tích CPI thực sự được bù đắp bằng việc giảm số lượng điều tra viên thu thập dữ liệu, NSO có thể quyết định mở rộng mẫu mà không cần thay đổi công thức tính chỉ số giá ở cấp địa bàn (EA) hoặc quy trình chọn mẫu.

Điều này thực sự xứng đáng, tuy nhiên, NSO cần phải cân nhắc cấu trúc chỉ số và quy trình chọn mẫu khi thu thập dữ liệu máy quét trực tiếp từ các chuỗi cửa hàng bán lẻ. Theo truyền thống, một chỉ số EA được tính từ giá được thu thập tại các cửa hàng thuộc các chuỗi bán lẻ khác nhau (hoặc các cửa hàng độc lập). Trong khi NSO muốn sử dụng nhiều thông tin giá từ nhiều chuỗi các cửa hàng bán lẻ hơn trước thì dường như nên coi việc kết hợp theo chuỗi EA như là tầng dữ liệu trong quy trình biên soạn chỉ số là điều cần thiết.

Thực tế khi NSO quyết định sử dụng hệ thống phân loại của nhà bán lẻ, dường như cấu trúc chỉ số cũng cần phải thay đổi: Mức thấp nhất của phân tầng nên được phân chia theo EA (chuỗi chi tiết). Điều này dẫn tới một số vấn đề, thứ nhất là liệu các cửa hàng thuộc chuỗi có nên được coi như là các cửa hàng riêng lẻ hay không, thứ hai là việc tính giá đơn vị cho tất cả các cửa hàng thuộc chuỗi cũng có thể hữu ích (Ivancic and Fox, 2013). Một số NSO không có lựa chọn, họ nhận được dữ liệu ở cấp độ chuỗi cửa hàng.

Vấn đề tiếp theo là thủ tục chọn mẫu hiện nay phải thay đổi. Cho rằng NSO vẫn chọn mẫu theo tỷ lệ doanh thu của mặt hàng lấy từ dữ liệu máy quét thì phương pháp này cũng có thể được sử dụng để chọn mẫu các mặt hàng từ các chuỗi EA cụ thể, tiêu chuẩn để xác định mặt hàng chi tiết (và tính các giá trị đơn vị) cấp cửa hàng hoặc chuỗi cửa hàng. Nếu NSO muốn tăng kích thước mẫu để sử dụng phần lớn các thông tin giá từ dữ liệu máy quét thu thập được, quy trình chọn mẫu cần được cân nhắc.

Một vấn đề khác là làm sao để tích hợp chỉ số giá EA chuỗi chi tiết từ dữ liệu máy

quét với thông tin giá từ các nguồn khác. Bởi những EA này khác với EA trong cấu trúc chỉ số giá truyền thống, chỉ số giá từ dữ liệu máy quét phải được tổng hợp ở cấp độ chi tiết nhất của chỉ số giá được NSO công bố hiện nay. Nói cách khác, việc tổng hợp gồm 2 bước: Chỉ số tổng hợp chuỗi EA mức chi tiết hơn, và tổng hợp các chỉ số dữ liệu máy quét với các chỉ số giá ở mức liên quan đến các chuỗi cửa hàng bán lẻ và các cửa hàng độc lập.

Dữ liệu doanh thu mang lại cơ hội cho các NSO trong việc tính toán các quyền số sử dụng để tính chỉ số giá một cách kịp thời và đều đặn hơn. Điều này có được theo nhiều cách, phụ thuộc vào sự tiếp cận dữ liệu máy quét của NSO tại các chuỗi cửa hàng. Điều đó cho thấy quyền số sử dụng các chỉ số giá từ dữ liệu máy quét được cập nhật hàng năm, sử dụng dữ liệu doanh thu từ 12 tháng liền trước. Sự kết hợp các chỉ số tính từ dữ liệu máy quét với các chỉ số được tổng hợp từ các nguồn khác đòi hỏi dữ liệu tiêu dùng của các chỉ số lân cận, các chỉ số này khó có được hoặc khó ước lượng được.

Nếu không có dữ liệu máy quét, các dữ liệu tiêu dùng chi tiết phân theo mặt hàng (hoặc các gói hàng) sẽ không có sẵn hoặc nếu có sẵn cũng sẽ không đều đặn. Vì vậy, phần lớn các các cơ quan thống kê vẫn áp dụng các phương pháp chỉ số không dùng quyền số ở mức thấp nhất của CPI: Giá hoặc thay đổi giá của các mặt hàng được chọn mẫu từ một chuỗi EA được kết hợp không cần quyền số gián tiếp của các mặt hàng dựa trên tầm quan trọng của mặt hàng trong nền kinh tế. Trong hầu hết các trường hợp, công thức chỉ số Jevons được sử dụng bởi NSO.

Các tập dữ liệu máy quét bao gồm dữ liệu doanh thu ở hầu hết cấp độ chi tiết. Những dữ liệu này có thể được sử dụng để chọn mẫu tỷ lệ các mặt hàng theo doanh thu của chúng, như đã đề cập ở trên, nhưng tăng thêm một số mục. Bao gồm các xác suất được coi như các quyền số gián tiếp. Đó là, chỉ số giá EA thực tế sẽ là một chỉ số có quyền số gián tiếp và xác suất đưa vào sẽ tương ứng với chỉ số mục tiêu/tổng thể đang nhắm đến (Balk, 2005). Hơn thế nữa, phân bổ doanh thu mục mặt hàng trong dữ liệu máy quét thường bị lệch. Do đó, chọn mẫu tỷ lệ thuận với doanh thu có khả năng chọn một số mặt hàng có doanh thu cao với xác suất bằng 1. Cho rằng các cơ quan thống kê thử ước lượng chỉ số mục tiêu có quyền số theo công thức bình quân nhân sử dụng chỉ số Jevons dựa trên-mẫu (không quyền số). Các mặt hàng có doanh thu nhỏ sẽ có một quyền số ẩn (*implicit weight*) 1, nhưng các mặt hàng có doanh số cao sẽ không có quyền số, điều này hiển nhiên không phải là giải pháp tốt, các mục sau nên là quyền số ẩn. Chỉ số giá có quyền số phản ánh mức độ quan trọng trong nền kinh tế thường được yêu thích hơn các chỉ số không quyền số gồm các xác suất tiềm ẩn. Các phương pháp quyền số đối với dữ liệu máy quét sẽ được thảo luận cụ thể và chi tiết trong tiểu mục 4.6 và phần tiếp theo của tài liệu này.

4.6. Sử dụng các tập dữ liệu máy quét tính CPI theo phương pháp mới

Các tiếp cận trong mục 4.2-4.5 cho phép NSO tiếp tục sử dụng các phương pháp chọn mẫu cơ sở để tính toán CPI. Việc cải thiện tính chính xác của CPI có thể thực hiện vì các loại giá (ví dụ giá trị đơn vị) có tính đại diện cao hơn cho mức tiêu dùng thực tế của người tiêu dùng; các mặt hàng được chọn mẫu phản ánh khối lượng bán; và quyền số sử dụng để đo lường sự thay đổi giá cập nhật hơn với tần suất đều đặn hơn.

Thách thức chính mà NSO gặp phải liên quan đến sự gia tăng nhu cầu các nguồn lực (Bird et al., 2014). Duy trì cỡ mẫu cơ bản, đặc biệt khi các mẫu giá được mở rộng, đòi hỏi sự hỗ trợ một cách thủ công bởi *doanh thu sản phẩm* có thể lớn[6].

Cách tốt nhất, NSO sẽ sử dụng tất cả các thông tin có sẵn trong các bộ dữ liệu máy quét thay vì chọn mẫu. Quy trình xử lý thủ công toàn bộ các tập dữ liệu máy quét cực kỳ tốn kém, và không thể đáp ứng được lịch biên soạn CPI. Vì vậy, quy trình tổng hợp CPI tự động được đặt ra.

Đồng thời, khi sử dụng tổng toàn bộ các sản phẩm, không chọn mẫu, công thức chỉ số có quyền số nên được sử dụng. Doanh thu sản phẩm đặt ra một vấn đề quan trọng. Nhằm tối đa hóa lượng liên kết trong dữ liệu, chuỗi liên kết ở tần suất lớn là điều cần thiết. Tuy nhiên có thể dẫn tới chuỗi (drift) trôi trong chỉ số. Các phương pháp tính chỉ số giá đa phương được xây dựng cho chuỗi tự do là phù hợp nhất giúp xử lý toàn bộ sản phẩm trong dữ liệu máy quét.

Minh Ánh (dịch)

Nguồn: Charp 10, Scanner data, pp 2-11.

---

[6] *Cơ quan Thống kê Hà Lan lần đầu tiên giới thiệu việc sử dụng dữ liệu máy quét từ các siêu thị để tính CPI, chỉ số Lowe được sử dụng (Schut et al., 2002). Ý tưởng giống như các phương pháp truyền thống và xử lý mẫu khoảng 10.000 mã mặt hàng (mã sản phẩm) từ các chuỗi siêu thị. Tiếp cận này là cần thiết trong điều kiện các lựa chọn thủ công các mặt hàng thay thế hoặc biến mất và trong trường hợp điều chỉnh chất lượng được coi như là cần thiết.*