NGHIÊN CỨU SỬ DỤNG NGUỒN DỮ LIỆU SCANNER TRONG BIÊN SOẠN CHỈ SỐ GIÁ TIÊU DÙNG TẠI VIỆT NAM

*TS. Nguyễn Trí Duy, ThS. Nguyễn Thị Minh Ánh**

Tóm tắt:

Sự phát triển nhanh chóng của khoa học công nghệ trong thời gian qua đã tạo ra sự bùng nổ về thông tin, dẫn đến kỉ nguyên mới của dữ liệu. Sự ra đời của những nguồn dữ liệu mới là nguồn tài nguyên quan trọng cho sự phát triển của nhiều lĩnh vực khoa học, trong đó có thống kê. Dữ liệu máy quét mã vạch (Scanner data) cũng là một trong những nguồn dữ liệu mới mang lại tiềm năng to lớn trong công tác thống kê giá truyền thống tại nhiều cơ quan và tổ chức thống kê trên thế giới.

Trong phạm vi bài báo này, nhóm nghiên cứu sẽ đưa ra các cơ sở lý luận về scanner data. Dựa trên cơ sở nghiên cứu kinh nghiệm quốc tế và thực trạng biên soạn chỉ số giá tiêu dùng theo phương pháp truyền thống tại Việt Nam, nhóm nghiên cứu tiếp tục đánh giá khả năng áp dụng nguồn dữ liệu mới này trong công tác thống kê giá tiêu dùng và đưa ra các khuyến nghị quan trọng về mặt kỹ thuật trong quá trình khai thác nguồn dữ liệu mới này. Cuối cùng, nhóm nghiên cứu sẽ tiến hành đề xuất một danh mục mặt hàng tiêu dùng cụ thể có thể thu thập thông tin từ nguồn dữ liệu scanner data.

1. Khái niệm scanner data

Trước hết cần hiểu về scanner data là gì? Hiện nay có khá nhiều khái niệm về dữ liệu máy quét (scanner data). Theo INSEE, dữ liệu máy quét là dữ liệu được các nhà bán lẻ ghi lại khi người tiêu dùng mua hàng tại một thời điểm nhất định. Thông tin trong scanner data bao gồm lượng mặt hàng được bán và giá bán của mỗi mặt hàng được bán. Những dữ liệu này được sử dụng trong việc tổng hợp Chỉ số giá tiêu dùng hoặc nghiên cứu về chỉ số này.[1] Trong khi đó, theo Cơ quan Thống kê Liên bang Đức, dữ liệu máy quét là dữ liệu giao dịch kỹ thuật số được ghi lại tại máy tính tiền của các cửa hàng bán lẻ, cung cấp thông tin về doanh thu, doanh số bán hàng và chủng loại mặt hàng bán ra. Dữ liệu máy quét có khả năng số hóa các bộ thống kê khác nhau và đảm bảo cũng như cải thiện chất lượng dữ liệu. Dữ liệu máy quét có thể được sử dụng trong thống kê giá tiêu dùng hoặc để xác định chênh lệch giá theo khu vực.[2] Một số định nghĩa khác được đưa ra bởi trang *thebusinessprofessor*: Dữ liệu máy quét chỉ là dữ liệu được ghi lại thông qua máy quét thanh toán để theo dõi hoạt động mua hàng

* *Viện Khoa học Thống kê*

[1] *https://www.insee.fr/en/metadonnees/definition/c2159#:~:text=Scanner%20data%20are%20the%20data,sold%20 and%20the%20sales%20price.*

[2] *https://www.destatis.de/EN/Service/EXSTAT/Datensaetze/scanner-data.html#:~:text = Scanner%20data%20are%20 digital%20transaction,the%20types%20of%20items%20sold.*

của khách hàng. Khi họ thanh toán hóa đơn tại cửa hàng, giao dịch mua hàng đó sẽ được ghi lại và tạo ra lượng dữ liệu khổng lồ.[3]

Như vậy, hiện nay có khá nhiều cơ quan tổ chức đưa ra các khái niệm về scanner data. Tuy nhiên, các khái niệm đều có chung quan điểm về dữ liệu máy quét là loại dữ liệu giao dịch kỹ thuật số được ghi lại tại máy tính tiền của các cửa hàng bán lẻ, bao gồm các thông tin về giá bán, lượng bán, doanh thu,…của mặt hàng được bán.

- Ưu điểm và hạn chế của scanner data

Nguồn dữ liệu scanner data có ưu điểm là mặt hàng vô cùng phong phú, khối lượng và tốc độ dữ liệu lớn đảm nên bảo tính đại diện và tính kịp thời cho dữ liệu. Thay vì phải di chuyển xuống từng địa bàn để thu thập giá bán, thông qua việc sử dụng scanner data cho phép cơ quan thống kê truy cập trực tiếp/ thu thập dữ liệu giá bán, khối lượng bán của toàn bộ các mặt hàng tiêu dùng tại cửa hàng/ chuỗi siêu thị bán lẻ mà không cần điều tra trực tiếp. Ngoài ra, sử dụng nguồn dữ liệu máy quét có thể giảm bớt công sức điều tra thu thập dữ liệu, tiết kiệm thời gian và nhiều nguồn lực kinh phí như chi phí thiết kế, chi phí xây dựng phương án điều tra, chi phí thu thập dữ liệu cho điều tra viên hiện nay của phương pháp điều tra giá truyền thống.

Bên cạnh những ưu điểm, nguồn dữ liệu này cũng có một số hạn chế cụ thể như sau. Trước hết, việc tiếp cận các tập dữ liệu scanner data đòi hỏi sự thỏa thuận thống nhất giữa bên cung cấp và bên thụ hưởng. Dữ liệu quét được bảo mật trực tiếp từ các doanh nghiệp bán lẻ. Để tiếp cận được nguồn dữ liệu này, các cơ quan thống kê (người thụ hưởng) cần có chính sách đàm phán hoặc có sự hỗ trợ về mặt cơ chế pháp lý đối với các đơn vị cung cấp dữ liệu trong khi hiện nay tại Việt Nam, dữ liệu scanner data chưa phải là nguồn dữ liệu thống kê chính thức. Ngoài ra, việc sử dụng scanner data phục vụ tính CPI cũng đòi hỏi những công cụ và cách thức mới so với tiếp cận theo phương pháp truyền thống, đặc biệt trong công tác phân loại và xử lý dữ liệu.

Để khai thác được nguồn dữ liệu tiềm năng này, nhóm nghiên cứu tiến hành tìm hiểu kinh nghiệm quốc tế về việc sử dụng scanner data trong biên soạn chỉ số giá tiêu dùng. Dưới đây là kinh nghiệm cụ thể của một số cơ quan và tổ chức thống kê uy tín trên thế giới.

2. Kinh nghiệm quốc tế về việc sử dụng scanner data trong biên soạn chỉ số giá tiêu dùng

Nhóm làm việc quốc tế về dữ liệu máy quét

Cơ quan thống kê của Liên hợp quốc đã thành lập Nhóm làm việc chuyên trách về dữ liệu máy quét cho thống kê nhà nước của Ủy ban thống kê Liên hợp quốc thành lập. Nhóm bao gồm các thành viên đến từ 21 quốc gia[4], 3 tổ chức bao gồm Eurostat, IMF, UNSD và đại học Đại học Graz của Áo. Trách nhiệm chính của nhóm là khám phá việc sử dụng các nguồn dữ liệu mới như là dữ liệu máy quét (scanner) và dữ liệu trích xuất từ web (web scraped data) giúp tăng cường việc sử dụng các nguồn dữ liệu, cải thiện hiệu quả làm việc. Nhóm hoạt động dựa trên 3 mục tiêu chính của nhóm bao gồm[5]: Mục tiêu 1: Hướng dẫn sử dụng các nguồn dữ liệu thay thế (ADS) tính toán chỉ số giá tiêu dung; Mục tiêu 2: Phân loại dữ liệu máy quét; Mục tiêu 3: Đào tạo nâng cao năng lực

[3] *https://thebusinessprofessor.com/en_US/principles-of-marketing/scanner-data-explained*

[4] *Các quốc gia bao gồm: Australia, Austria, Belgium, Brazil, Canada, Denmark, Finland, Germany ,Italy, Malaysia, Mexico, Netherlands, New Zealand, Norway, Poland, Switzerland, Thailand, Turkey, United Kingdom, United States*

[5] *https://unstats.un.org/bigdata/task-teams/scanner/index. cshtml*

Cơ quan Thống kê Đan Mạch

Bắt đầu tiếp nhận dữ liệu scanner data từ năm 2011, Cơ quan thống kê Đan Mạch đã thực hiện việc phân loại và tính toán thử nghiệm chỉ số giá tiêu dùng từ scanner data. Năm 2016, dữ liệu này trở thành một phần của CPI/HICP bên cạnh dữ liệu truyền thống. Cho tới nay, các kết quả nghiên cứu cho thấy, dữ liệu scanner data có thể bổ sung và thay thế khoảng 30% danh mục mặt hàng truyền thống hiện nay tại Đan Mạch.

Quá trình thử nghiệm đến lúc vận hành hoạt động sử dụng dữ liệu máy quét để biên soạn chỉ số giá tiêu dùng về cơ bản bao gồm 2 giai đoạn: Giai đoạn 1 (giai đoạn cơ sở) tiến hành xây dựng biểu mẫu thu thập, chương trình phân loại dữ liệu, các danh mục mặt hàng), chương trình tính toán và thiết lập các báo cáo. Giai đoạn 2 (giai đoạn cập nhật, bổ sung) thực hiện việc theo dõi, đánh giá, hoàn thiện và chỉnh sửa. Cho đến nay, chỉ số giá tiêu dùng tính từ scanner là kênh bổ sung chính thức cho bộ dữ liệu truyền thống.

Về phương pháp thu thập dữ liệu giá phục vụ công tác tính chỉ số giá tiêu dùng, hiện nay, cơ quan thống kê Đan Mạch sử dụng kết hợp 3 phương pháp: (1) Thu thập giá trực tiếp tại các cửa hàng bán lẻ; (2) Thu thập giá thông qua gọi điện thoại; (3) Sử dụng dữ liệu máy quét sản phẩm được cung cấp bởi các siêu thị tại Đan Mạch.

Về quy trình tính toán chỉ số giá tiêu dùng CPI sử dụng scanner data được Thống kê Đan mạch thực hiện theo 3 bước: (1) Kiểm tra dữ liệu; (2) Gán dữ liệu; (3) Kết hợp dữ liệu scanner data với Biên soạn chỉ số giá tiêu dùng kết. Trong đó, tại bước 1 của quy trình sẽ bao gồm công tác kiểm tra làm và sạch dữ liệu. Tại bước 2, hoạt động gán dữ liệu được thực hiện theo 4 phương pháp:

- *Phương pháp 0 (Gắn nhãn thủ công hoặc xác thực nhãn dự đoán)* Phương pháp này áp dụng với tập dữ liệu máy quét có kích thước nhỏ hoặc trung bình, được thực hiện một cách thủ công đối với các từ khóa đơn giản.

- *Phương pháp 1 (phân loại dựa trên thuộc tính sản phẩm):* áp dụng khi dữ liệu có cấu trúc, các biến trong tập dữ liệu của nhà bán lẻ ổn định theo thời gian và không quá nhiều.

- *Phương pháp 2 (phân loại dựa trên mã barcode của sản phẩm:* Dựa trên mã barcode sản phẩm (trường hợp Đan Mạch thường sử dụng các mã EAN), cơ quan Thống kê Đan Mạch sẽ tiến hành phân loại mã sản phẩm và khớp với mã trong danh mục một cách tự động hơn.

- *Phương pháp 3 (sử dụng học máy)* Sau khi thực hiện các phương pháp 0,1,2, cơ quan Thống kê Đan Mạch kết hợp sử dụng phương pháp học máy với các kỹ thuật học máy không giám sát để tiến hành phân loại sản phẩm.

Kết quả dữ liệu scanner data được phân loại sẽ có cấu trúc bao gồm các trường thông tin: Ngày tháng, mã số GTIN, doanh thu, khối lượng bán, đơn vị tính, lượng trên mỗi đơn vị, mô tả sản phẩm.

Dữ liệu scanner data sau khi được làm sạch sẽ kết hợp với dữ liệu giá được thu thập bằng phương pháp truyền thống đưa vào tính toán chỉ số giá tiêu dùng. Hiện nay, cơ quan Thống kê Đan Mạch áp dụng tính toán chỉ số hoài hòa giá HICP, bộ quyền số được tính toán từ các dữ liệu điều tra mức sống dân cư.

Về danh mục mặt hàng và thông tin tiến hành thu thập, hiện nay, Cơ quan thống kê Đan Mạch sử dụng dữ liệu scanner chủ yếu đối với các nhóm mặt hàng liên quan đến hàng ăn các mặt hàng lương thực, thực phẩm, đồ uống, và thuốc lá, sau đó tiếp tục bổ sung thêm một số mặt hàng tiêu dùng khác thuộc nhóm đồ gia dụng như giấy vệ sinh, thực phẩm cho mèo, que thử thai….

Dữ liệu được thu thập từ 4 chuỗi siêu thị lớn chiếm khoảng 80% doanh thu trên thị

trường với số lượng chi nhánh phân bổ rộng khắp Đan Mạch. Số lượng mặt hàng được cung cấp xấp xỉ 4.500 mặt hàng, tương đương khoảng 20 triệu GB dữ liệu mỗi tuần.

Các trường dữ liệu thu thập bao gồm:

- Ngày (date): bao gồm 4 chữ số và bao gồm năm (2 chữ số), tuần (2 chữ số);

- Mã số cửa hàng (store) : Mỗi cửa hàng được đánh một mã số riêng, không trùng lặp;

- Mã sản phẩm EAN[6] (EAN number): mỗi sản phẩm được xác định bằng mã sản phẩm được gọi là Mã đánh số bài viết châu Âu (EAN) hoặc Mã tra cứu sản phẩm (PLU[7]).

- Doanh thu (turn-over): Doanh thu từ việc bán mặt hàng;

- Số lượng (volume): lượng mặt hàng được bán;

- Giá của mặt hàng: Giá được tính bằng cách chia doanh thu hàng tuần cho khối lượng hàng tuần cho mỗi số EAN. Cuối cùng, mã số sản phẩm có thể được sử dụng để phản ánh thứ bậc sản phẩm của chuỗi siêu thị. Hệ thống phân cấp sản phẩm này là không thể thiếu khi liên kết số EAN với COICOP. Đối với mỗi EAN có một mô tả sản phẩm do chuỗi siêu thị tạo ra;

- Đơn vị tính (unit): đơn vị tính của mặt hàng được bán;

- Khối lượng theo đơn vị (Quanity per unit): Khối lượng tương ứng với từng đơn vị mặt hàng;

- Mã số sản phẩm (product number): mã được đánh cho mặt hàng, thông thường do siêu thị quy định thứ tự đánh mã;

- Mô tả sản phẩm (product description): bao gồm các thông tin mô tả về quy cách, phẩm cấp của mặt hàng.

[6] *https://thegioimavach.com/ean-code-la-gi-ean-8-va-ean-13-cung-nhung-dieu-ban-can-biet*

[7] *https://fas.tdtu.edu.vn/tin-tuc/2018/cach-nhan-biet-ma-so-tren-tem-hoa-qua-nhap-khau*

- Dưới đây là hình ảnh minh họa cấu trúc các thông tin thu thập từ scanner data:

Date	Store	EAN number	Turn-over	Vol-ume	Unit	Quantity per unit	Product number	Product description
1104	7894	2920080800007	3402,70	211	Gram	300	910076003	Sliced bacon 2x150 G.
1104	7895	2920080800007	2119,65	163	Gram	300	910076003	Sliced bacon 2x150 G.
1104	7896	2920080800007	1516,05	108	Gram	300	910076003	Sliced bacon 2x150 G.
1104	7897	2920080800007	1478,13	105	Gram	300	910076003	Sliced bacon 2x150 G.
1104	7214	2921056000005	302,50	14	Gram	200	911056001	Chicken Fillet
1104	7215	2921056000005	102,50	5	Gram	200	911056001	Chicken Fillet

Nguồn: Extending the Danish CPI with scanner data – A stepwise analysis, Jonas Mikkelsen JOM@DST.dk, Statistics Denmark, Prices and Consumption

Đối với hệ thống công nghệ thông tin và phần mềm xử lý dữ liệu, Cơ quan thống kê Đan Mạch đã xây dựng một hệ thống tự động tiếp nhận thông tin, đồng thời xây dựng mẫu biểu cụ thể cho các siêu thị thực hiện việc báo cáo dữ liệu. Các siêu thị có trách nhiệm hoàn thiện các mẫu biểu được cung cấp đồng thời phân loại danh mục mặt hàng theo các mã barcode được thống nhất trước khi báo cáo lại cơ quan Thống kê. Hệ thống công nghệ thông tin hỗ trợ duy trì liên kết hàng tuần giữa các mã trong dữ liệu máy quét với các mã của COICOP. Hệ thống này dựa trên 2 mức độ can thiệp vào dữ liệu máy quét:

- Tự động phân loại các danh mục mặt hàng theo bảng COICOP dựa trên các kết quả đã được gán mã một cách thủ công. Việc phân công thủ công này chỉ được thực hiện một lần và chỉ thay đổi nếu chuỗi siêu thị quyết định thay đổi cách phân loại của mình.

- Quá trình tìm kiếm từ bắt đầu bằng việc theo dõi doanh thu từ các sản phẩm đã bán trong các nhóm còn lại và tỷ lệ của chúng với doanh thu của các nhóm COICOP 6 chữ số đã được chỉ định tổng hợp ở cấp độ 4 chữ số. Bằng cách đó, các nhóm còn lại được đánh giá và ưu tiên cho nhóm sản xuất (xem hình minh họa bên dưới).

Mặt khác, liên quan đến phần mềm xử lý dữ liệu, cơ quan thống kê Đan Mạch sử dụng

phần mềm R trong quá trình xử lý dữ liệu và tính toán chỉ số giá tiêu dùng.

Như vậy, hiện nay trên thế giới đã có nhiều cơ quan thống kê khai thác và sử dụng thành công scanner data. Tại Việt Nam, công tác thu thập giá tiêu dùng phục vụ tính toán CPI đã và đang được Tổng cục Thống kê thực hiện. Tuy nhiên, để hiểu rõ hơn về thực trạng của công tác thu thập giá truyền thống, trong phần tiếp theo, nhóm nghiên cứu sẽ chỉ ra những hạn chế hiện nay.

3. Thực trạng thu thập dữ liệu giá tiêu dùng ở Việt Nam

Hiện nay, thông tin thống kê về chỉ số giá tiêu dùng được thu thập từ cuộc điều tra giá tiêu dùng do Tổng cục Thống kê triển khai ở cả 63 tỉnh, thành phố và được công bố hằng tháng vào các ngày cuối tháng. Trong phương án điều tra giá tiêu dùng giai đoạn 2020-2025, ngành Thống kê thực hiện cuộc điều tra giá tiêu dùng bằng phương pháp chọn mẫu đối với rổ hàng hóa, dịch vụ gồm 754 (tương ứng 11 nhóm hàng cấp 1, 32 nhóm cấp 2, 86 nhóm cấp 3 và 290 nhóm cấp 4). Hàng hóa và dịch vụ trong rổ hàng hóa được chia làm 3 nhóm chính: Nhóm thứ nhất chỉ điều tra 1 lần trong tháng và sẽ điều tra vào ngày 10 hàng tháng; nhóm thứ hai sẽ điều tra 3 lần trong tháng vào các ngày 1, 10, 20 hàng tháng; nhóm thứ ba theo số lần phát sinh trong tháng.

Mặc dù hoạt động điều tra giá tiêu dùng truyền thống hiện nay là nguồn cung cấp thông tin về giá tiêu dùng kịp thời và đáng tin cậy. Tuy nhiên, công tác điều tra giá tiêu dùng còn tồn tại một số khó khăn hạn chế nhất định như sau:

Thứ nhất, công tác thu thập tại địa bàn gặp nhiều khó khăn. Với số lượng 754 mặt hàng và dịch vụ đại diện cho thời kỳ 2020 - 2025, mỗi điều tra viên phụ trách thu thập giá khoảng 90 mặt hàng, mỗi khu vực điều tra cần 8 - 10 điều tra viên. Đặc biệt khi thời gian thu thập đúng vào những ngày nghỉ Lễ, Tết, thời gian địa phương thực hiện giãn cách xã hội… khiến phần lớn các cơ sở kinh doanh không mở cửa bán hàng và giá cả hàng hóa, dịch vụ thường có sự biến động lớn tại những thời điểm này cũng là một hạn chế cần phải xử lý.

Thứ hai, vẫn còn phát sinh sai số phi chọn mẫu trong quá trình triển khai thu thập.

Thứ ba, khó khăn trong việc xử lý đối với những hàng hóa và dịch vụ có chu kỳ sống ngắn hạn, không tồn tại vào thời điểm điều tra và nhiều hàng hóa mới phát sinh trong kỳ điều tra.

Thứ tư, những vấn đề chọn mẫu và tính toán số lượng các điểm bán hàng bình ổn ở địa phương, việc xử lý các cửa hàng trong mẫu điều tra ngày càng thu hẹp kinh doanh, thị phần giảm sút … cũng là những khó khăn khi triển khai thu thập số liệu theo phương pháp điều tra truyền thống.

Thứ năm, số lần đến thu thập thông tin giá tại điểm điều tra quá nhiều và lặp lại thường xuyên trong tháng gây lên tình trạng phiền hà cho các điểm điều tra và dễ xảy ra việc kê khai giá qua loa, đại khái nên không chính xác. Chẳng hạn như đối với những mặt hàng lấy giá tại nhiều điểm trong cùng một khu vực điều tra, phần lớn ít có sự chênh lệch về mức giá giữa các điểm nên dễ dẫn đến tình trạng điều tra viên chỉ thu thập giá tại một điểm và sao chép cho các điểm còn lại.

Trước những hạn chế còn tồn tại trong công tác điều tra giá tiêu dùng truyền thống hiện nay, việc cải tiến, nâng cao chất lượng hoạt động điều tra, thu thập các thông tin về giá tiêu dùng ngày càng trở lên cần thiết. Để sử dụng được nguồn dữ liệu scanner data trong công tác thống kê giá tiêu dùng tại Việt Nam, cần có các nghiên cứu đánh giá tính khả thi của nguồn dữ liệu này.

4. Nghiên cứu nguồn scanner data trong thống kê giá tiêu dùng ở Việt Nam

Trước hết, sự phát triển của thị trường bán lẻ tại Việt Nam đã cho thấy tiềm năng của việc sử dụng scanner data trong thống kê giá tiêu dùng.

Hiện nay, thị trường bán lẻ của Việt Nam được World Data Lab's Việt Nam sẽ tăng 8 bậc trong top 30 thị trường tiêu dùng lớn nhất thế giới trong thập kỷ này và chiếm 1,1% dân số thuộc tầng lớp tiêu dùng thế giới vào năm 2030. Nghiên cứu của Mordor Intelligence chỉ ra, thị trường bán lẻ Việt Nam được dự báo là sẽ chứng kiến tốc độ tăng trưởng kép hàng năm (CAGR) giai đoạn 2023-2028 đạt 12,05%, gấp khoảng 2 lần tốc độ tăng trưởng GDP ước tính trong cùng giai đoạn. Tốc độ tăng trưởng của ngành bán lẻ Việt Nam trong 5 năm tới cũng được dự báo sẽ vượt xa so với các thị trường Đông Á, Đông Nam Á và mức tăng trưởng bình quân của thế giới.[8] Cũng theo thống kê từ Vụ Thị trường trong nước (Bộ Công Thương) năm 2019, cả nước hiện có khoảng 8.660 chợ, 800 siêu thị quy mô lớn, 168 trung tâm thương mại các loại và hơn 1 triệu cửa hàng quy mô hộ gia đình.[9] Tuy nhiên, kênh bán lẻ hiện đại mới đáp ứng được 25% nhu cầu của người dân, 75% còn lại phụ thuộc vào kênh phân phối truyền thống.

Về thị phần, hiện nay 15% thị phần của thị trường bán lẻ là qua trung tâm thương mại, 50% thị phần của phương thức bán lẻ qua cửa hàng tiện lợi, 10% thị phần của phương thức bán hàng qua siêu thị mini và khoảng 50% thị phần của phương thức bán lẻ không thông qua cửa hàng.[10] Báo cáo của đã Deloitte[11], công ty kiểm toán hàng đầu thế giơi nhận định chỉ ra, tính đến hết năm 2019, cả nước có khoảng 3.450 siêu thị, với tổng diện tích sàn lên đến hơn 1,6 triệu m2. Cũng theo báo cáo của Dellloite Đối với các doanh nghiệp bán lẻ trong nước, hai doanh nghiệp Saigon Co.op và Bách Hóa Xanh là hai đơn vị chiếm thị phần bán lẻ cao nhất (lần lượt là 43% và 14% thị phần)[12]. Ở phân khúc đại siêu thị với 58 điểm bán, chuỗi Big C của Thái Lan chiếm 57,6% thị phần. Saigon Co.op là tên tuổi Việt Nam duy nhất cạnh tranh với các thương hiệu đại siêu thị quốc tế như Lotte Mart, Aeon Mall và E-Mart.

Về hình thức tiêu dùng, nếu như trước kia người dân Việt Nam chủ yếu mua bán các sản phẩm tiêu dùng tại các chợ, cửa hàng tạp hóa truyền thống thì nay rất nhiều người đã mở rộng chi tiêu tại các siêu thị, cửa hàng tiện lợi, trung tâm thương mại hay các kênh bán hàng online. PSI, Công ty Cổ phần Chứng khoán Dầu khí, nhận định, chợ truyền thống và các cửa hàng tạp hoá mặc dù chiếm thị phần lớn doanh số bán lẻ tại Việt Nam, tuy nhiên, thị trường bán lẻ Việt Nam chứng kiến sự thay đổi nhanh chóng trong xu hướng tiêu dùng. Trong khi các cửa hàng tạp hoá, chợ truyền thống dần mất đi vị thế thì các hình thức bán lẻ hiện đại hơn đã tăng thị phần từ 15% năm 2015 lên 26% vào năm 2022[13]. Đối với hình thức siêu thị và cửa hàng tiện lợi, Deloitte cũng đánh giá người tiêu dùng thành thị ưa chuộng siêu thị và cửa hàng tiện lợi hơn vì sự tiếp cận dễ dàng của những mô hình này. Đại siêu thị chỉ được tận dụng với những trường hợp mua sắm lượng hàng lớn.

---

[8] *https://tapchicongthuong.vn/bai-viet/psi-thi-truong-ban-le-soi-dong-voi-cuoc-dua-gianh-thi-phan-cua-cac-ong-lon-112298.htm*

[9] *https://mof.gov.vn/webcenter/portal/ttpltc/pages_r/l/chi-tiet-tin-ttpltc?dDocName=MOFUCM154907*

[10] *https://mof.gov.vn/webcenter/portal/ttpltc/pages_r/l/chi-tiet-tin-ttpltc?dDocName=MOFUCM154907*

[11] *https://sapp.edu.vn/bai-viet-acca/gioi-thieu-tong-quan-ve-deloitte-global-va-deloite-viet-nam/#1-gioi-thieu-chung*

[12] *https://vietnamfinance.vn/sieu-thi-nao-chiem-thi-phan-cao-nhat-tai-viet-nam-20180504224240941.htm*

[13] *https://tapchicongthuong.vn/bai-viet/psi-thi-truong-ban-le-soi-dong-voi-cuoc-dua-gianh-thi-phan-cua-cac-ong-lon-112298.htm*

Như vậy, trước thực trạng phát triển đa dạng về các loại hình bán lẻ và sự thay đổi về nhu cầu tiêu dùng của người dân Việt Nam, việc bổ sung các nguồn dữ liệu mới như scanner data vào công tác biên soạn chỉ số giá tiêu dùng truyền thống là một điều cần thiết. Điều này không những giúp đảm bảo tính đầy đủ, kịp thời của dữ liệu giá tiêu dùng trong thời đại mới mà còn phù hợp với định hướng hiện đại hóa ngành Thống kê Việt Nam.

Tính khả thi của việc sử dụng scanner data trong công tác thống kê giá tiêu dùng đã được chứng minh thông qua thử nghiệm khai thác scanner data.

Vừa qua, Tổng cục Thống kê đã tiến hành thử nghiệm khai thác nguồn dữ liệu scanner data trong việc tính toán chỉ số giá tiêu dùng dưới sự hỗ trợ của cơ quan Thống kê Đan Mạch. Thử nghiệm sử dụng dữ liệu scanner biên soạn chỉ số giá tiêu dung chứng minh tính khả thi của việc sử dụng nguồn dữ liệu này trong việc hỗ trợ biên soạn chỉ số giá tiêu dùng hiện nay.

Về dữ liệu, Tổng cục Thống kê đã phối hợp và làm việc với hệ thống siêu thị Bách Hóa Xanh, một trong những đơn vị quốc gia có thị phần bán lẻ cao nhất tại Việt Nam (14% thị phần toàn quốc, theo kết quả công bố của Delloite), tiến hành thu thập dữ liệu scanner data của chuỗi siêu thị này từ 6 đến tháng 9 năm 2022. Các thông tin tiêu dùng được hệ thống siêu thị Bách Hóa Xanh cung cấp từ hệ thống máy thanh toán của chuỗi hệ thống siêu thị này. Kết quả Tổng cục Thống kê đã tiếp nhận 17,4 triệu quan sát, tương ứng khoảng 21.000 sản phẩm tiêu dùng riêng biệt dùng từ hệ thống Bách Hóa Xanh.

Nhìn chung, các kết quả tính toán đối với một số mặt hàng tiêu dùng phổ biến và ổn định tại chuỗi siêu thị cho thấy xu hướng phù hợp và đúng đắn của dữ liệu scanner so với các kết quả truyền thống.

Tuy nhiên, thử nghiệm cũng chỉ ra một số vấn đề còn tồn đọng bao gồm các vấn đề mang tính kỹ thuật trong công tác xử lý và biên soạn dữ liệu, sự cần thiết trong việc xây dựng một danh mục mặt hàng có thể khai thác thông tin từ scanner thay thế cho các mặt hàng truyền thống. Để lấp đầy những khoảng trống này, nghiên cứu sẽ đưa ra một số khuyến nghị mang tính kỹ thuật quá trình khai thác scanner data, đồng thời đề xuất một số danh mục mặt hàng có thể sử dụng scanner data thay thế cho dữ liệu truyền thống.

Trên cơ sở các nghiên cứu về kinh nghiệm quốc tế và đánh giá tính khả thi của việc sử dụng scanner data trong công tác biên soạn chỉ số giá tiêu dùng, nhóm nghiên cứu đưa ra một số khuyến nghị như sau:

5. Đề xuất, khuyến nghị

Khuyến nghị về mặt kỹ thuật

+ Đối với hoạt động thu thập dữ liệu scanner data

Các file dữ liệu khi được thu thập từ siêu thị nên định dạng dưới các dạng thức file dễ dàng đọc trên máy tính và được sắp xếp theo một cấu trúc đồng nhất về thứ tự các danh mục. Chẳng hạn, các file dữ liệu excel nên định dạng csv, những file này không những giúp giảm tải dung lượng bộ nhớ, máy tính dễ đọc mà còn giúp cho việc đọc dữ liệu ổn định và ít xảy ra lỗi hơn. Điều này nên được quy định cụ thể ngay từ bước xây dựng mẫu biểu cung cấp dữ liệu cho các siêu thị.

Trong quá trình làm việc với các siêu thị (nhưng nhà cung cấp dữ liệu) cần làm rõ cấu trúc mã vạch sản phẩm, vì mã thực tế của mặt hàng giữa các nhà cung cấp khác nhau có thể khác nhau. Đặc biệt cần đề xuất siêu thị cung cấp dữ liệu với mã mặt hàng nhất quán giữa các chủng loại. Việc nhất quán mã mặt hàng sẽ đặc biệt hữu ích cho công tác phân loại mặt hàng theo danh mục COICOP, đồng thời giúp tiết kiệm các nguồn lực về thời gian, kinh phí, nhân lực một cách hiệu quả.

+ Đối với hoạt động làm sạch dữ liệu scanner data

Hoạt động làm sạch dữ liệu cần dặc biệt chú ý đến các trường thông tin có nội dung giống hệt nhau. Những dòng tin này cần được loại bỏ. Ngoài ra cũng cần lưu ý đến số lượng và chất lượng của dữ liệu đối với từng biến được thu thập. Chẳng hạn, dữ liệu về doanh thu bán mặt hàng là luồng thông tin dễ bị bỏ sót nhất. Mặt khác, với những biến chứa nhiều giá trị bị thiếu hoặc cao/thấp bất thường (gọi chung là giá ngoại lai) cần phải xác minh với nhà cung cấp dữ liệu để đảm bảo không bị khuyết thiếu và tăng cường tính chính xác cho dữ liệu trước khi thực hiện các hoạt động làm sạch dữ liệu khác như gán giá, xóa giá trị khuyết thiếu…

+ Đối với hoạt động phân loại dữ liệu scanner data

Hoạt động phân loại dữ liệu scanner data cần lưu ý đến mã vạch sản phẩm. Trong nhiều trường hợp mã vạch sản phẩm trong dữ liệu scannner data được các siêu thị cung cấp không phải là mã vạch chính thức, chỉ là mã vạch do siêu thị tự đánh. Phân biệt được rõ vấn đề mã vạch sẽ giúp việc phân loại được rõ ràng, tránh nhầm lẫn và tiết kiệm thời gian, công sức hơn.

+ Đối với hoạt động biên soạn chỉ số giá tiêu dùng từ dữ liệu scanner data

Trong quá trình biên soạn chỉ số giá tiêu dùng từ dữ liệu sccanner data cần lưu ý tới việc tích hợp các dữ liệu scanner data với bộ dữ liệu truyền thống, kết hợp các nhóm phân loại mặt hàng để tính các chỉ số cao hơn mà vẫn đảm bảo tính hài hòa nhất quán giữa các loại dữ liệu.

Mặc dù dữ liệu scanner data có ưu điểm là đa dạng các loại mặt hàng nhưng số mặt hàng ổn định được phân nhóm phù hợp vẫn gặp phải rất nhiều vấn đề. Chẳng hạn như tính đại diện của mặt hàng trên phạm vi toàn quốc, tính thời vụ của mặt hàng…Người làm công tác phân loại dữ liệu cần có biện pháp xử lý cụ thể cho từng nhóm mặt hàng này.

Đối với bộ quyền số phục vụ công tác tính toán chỉ số giá tiêu dùng hiện nay đang được xây dựng dựa trên kết quả điều tra Mức sống dân cư do TCTK thực hiện. Tuy nhiên khi kết hợp dữ liệu scanner như một kênh bổ sung cho dữ liệu truyền thống cần cân nhắc bộ quyền số phù hợp, vấn đề nên giữ nguyên bộ quyền số cũ hay xây dựng một bộ quyền số mới và nếu xây dựng mới thì theo tiêu chí nào? Để trả lời câu hỏi trên cần có thêm những nghiên cứu chuyên sâu, đánh giá khả năng và ưu nhược điểm của bộ quyền số.

Đề xuất danh mục các mặt hàng có thể thu thập dữ liệu từ nguồn scanner data

Nghiên cứu kinh nghiệm của cơ quan thống kê Đan Mạch cho thấy, cơ quan này sử dụng dữ liệu scanner chủ yếu đối với các nhóm mặt hàng liên quan đến hàng ăn các mặt hàng lương thực, thực phẩm, đồ uống và thuốc lá, (tương ứng với các nhóm số I, II, VI của rổ truyền thống Việt Nam); sau đó thống kê Đan Mạch tiếp tục bổ sung thêm một số mặt hàng tiêu dùng khác thuộc nhóm đồ gia dụng như giấy vệ sinh, thực phẩm cho mèo, que thử thai…(tương ứng nhóm V và XI của rổ hàng truyền thống của Việt Nam).

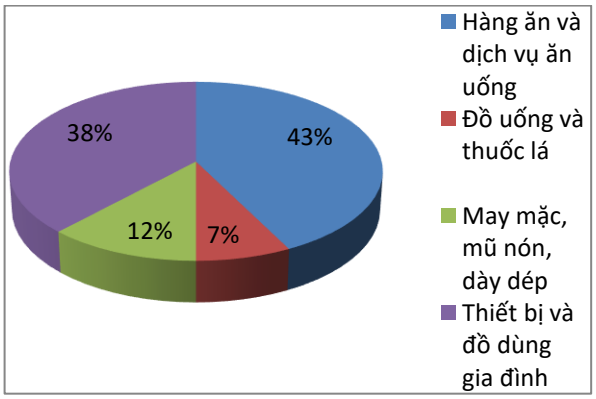
Tài liệu "Harmonised Index of Consumer Prices" (Chỉ số hài hòa giá tiêu dùng) do Ủy ban Châu Âu công bố đưa ra hướng dẫn về việc sử dụng scanner data trong biên soạn chỉ số giá tiêu dùng, đề cập đến danh mục hàng hóa sử dụng loại dữ liệu này chủ yếu bao gồm các sản phẩm về lương thực, thực phẩm, đồ uống và nhu yếu phẩm hàng ngày như thuốc, thuốc lá…(tương ứng với những nhóm hàng cấp I. Hàng ăn và dịch vụ ăn uống, II. Đồ uống và thuốc lá, V.Thiết bị và đồ dùng trong gia đình) trong rổ hàng truyền thống của Việt Nam.

Đây là những bài học kinh nghiệm quan trọng cho việc đề xuất danh mục các mặt hàng có thể thu thập dữ liệu từ nguồn scanner data.

Từ kết quả nghiên cứu kinh nghiệm quốc tế, thử nghiệm thực tiễn và rà soát danh mục rổ hàng truyền thống hiện nay, nhóm nghiên cứu thực hiện đề xuất danh mục mặt hàng tiêu dùng có thể thu thập từ nguồn dữ liệu scanner data như sau:

Các nhóm mặt hàng được đề xuất sẽ thuộc các nhóm cấp I, II, III, V bao gồm: Hàng ăn và dịch vụ ăn uống, Đồ uống và thuốc lá, May mặc, mũ nón, dày dép, Thiết bị và đồ dùng gia đình do các mặt hàng thuộc những nhóm này đáp ứng tốt tiêu chí sẵn có, ổn định, mã vạch rõ ràng, ít bị tác động bởi tính thời vụ. Trong 4 nhóm hàng cấp 1 trên lại lựa chọn ra được 134 mặt hàng cấp 4 thỏa mãn các tiêu chí. Tổng số mặt hàng cấp 4 đề xuất sẽ bao gồm 134 mặt hàng cấp 4, tương ứng với rổ hàng truyền thống. Trong tổng số 134 mặt hàng tương ứng với các mặt hàng cấp 4 của rổ hàng truyền thống được đề xuất, có tới 43% các mặt hàng thuộc nhóm I. hàng ăn và dịch vụ ăn uống, 38% thuộc nhóm V. Thiết bị và đồ dùng gia đình, 12% thuộc nhóm III. May mặc, mũ nó, dày dép và 7% thuộc nhóm II. Đồ uống và thuốc lá.

Biểu đồ: Tỷ lệ số mặt hàng phân theo các nhóm cấp 1 được đề xuất



Danh mục các mặt hàng được lựa chọn dựa trên các tiêu chí:

- Khả năng tiếp cận dữ liệu của mặt hàng tại các siêu thị;

- Sự sẵn có của các mặt hàng tại các siêu thị;

- Sự ổn định của mặt hàng, đặc biệt đối với các thông tin thu thập cần cung cấp thông tin ổn định như: lượng bán, doanh số bán, tiêu chuẩn, cách phẩm cấp của sản phẩm,...

- Sự rõ ràng của mã vạch sản phẩm (các mã EAN hoặc GTIN đồng nhất theo thời gian và trong chuỗi siêu thị);

- Sự tác động bởi tính thời vụ tới sự có mặt/biến mất của mặt hàng: mặt hàng càng ít bị tác động bởi tính thời vụ càng tốt;

- Sự phù hợp của mặt hàng đối với danh mục rổ hàng tiêu dùng truyền thống giúp thuận lợi cho việc phân loại dữ liệu theo danh mục rổ hàng COICOP và tính toán chỉ số giá tiêu dùng.

Nội dung danh mục các mặt hang chi tiết đề xuất xem tại phụ lục đính kèm.

6. Kết luận

Nghiên cứu trên đã chỉ ra các đặc điểm của nguồn dữ liệu mới scanner data cùng với kinh nghiệm quốc tế trong khai thác và ứng dụng nguồn dữ liệu scanner data đối với công tác thống kê giá. Trước thực trạng xu hướng tiêu dùng tại các siêu thị ngày càng tăng cao của người dân Việt Nam hiện nay, khả năng ứng dụng nguồn dữ liệu scanner data đối với công tác thống kê giá đã được chứng minh sau khi thử nghiệm sử dụng bộ dữ liệu scanner của Tổng cục Thống kê với một số kết quả ban đầu. Kết luận nhóm tác giả nghiên cứu cũng đưa ra một số khuyến nghị cụ thể trong quá trình khai thác nguồn dữ liệu mới scanner data và đề xuất một danh mục rổ hàng tiêu dùng có thể lấy dữ liệu từ nguồn dữ liệu mới này. Các kết quả nghiên cứu sẽ là bước đệm cơ sở cần thiết cho việc triển khai sử dụng nguồn dữ liệu mới scanner data đối với công tác thống kê giá truyền thống hiện nay.

Tài liệu tham khảo:

1. Phương án điều tra giá tiêu dùng thời kỳ 2020-2025, Tổng cục thống kê

2. Harmonised Index of Consumer Prices, EUROPEAN COMMISSION

3. Implementing scanner data in the Danish CPI - Paper to be presented at UN-Group of experts on the consumer price indices meeting, Geneve May 2014, Workshop 4- Scanner Data

Phụ lục: Danh mục các mặt hàng đề xuất lấy giá từ nguồn dữ liệu scanner data

STT	MÃ SỐ	MẶT HÀNG	STT	MÃ SỐ	MẶT HÀNG
	I	HÀNG ĂN VÀ DỊCH VỤ ĂN UỐNG	68	232402	+ Thuốc lào
1	11	LƯƠNG THỰC		III	MAY MẶC, MŨ NÓN, DÀY DÉP
2	110101	+ Gạo tẻ thường	69	312501	+ Vải các loại
3	110102	+ Gạo tẻ ngon	70	312601	+ Quần, áo cho nam (13 tuổi trở lên)
4	110103	+ Gạo nếp	71	312602	+ Quần, áo cho nữ (13 tuổi trở lên)
5	110201	+ Bột mì	72	312603	+ Quần áo cho trẻ em trai (từ 2 đến dưới 13 tuổi)
6	110202	+ Ngô	73	312604	+ Quần áo cho trẻ em gái (từ 2 đến dưới 13 tuổi)
7	110203	+ Khoai	74	312605	+ Quần áo cho trẻ sơ sinh (từ 0 đến 2 tuổi)
8	110204	+ Sắn	75	322701	+ Khăn mặt, khăn quàng
9	110301	+ Bánh mì	76	322702	+ Găng tay, thắt lưng
10	110302	+ Bún, bánh phở, bánh đa	77	322703	+ Bít tất các loại
11	110303	+ Mỳ sợi, mỳ, phở/ cháo ăn liền	78	322801	+ Mũ, nón, áo mưa
12	110304	+ Miến	79	322802	+ Mũ bảo hiểm xe máy
13	110305	+ Bột ngô	80	332901	+ Giầy dép (sandan) cho nam
14	110306	+ Ngũ cốc ăn liền	81	332902	+ Giầy, dép (sandan) cho nữ
15	110307	+ Ngũ cốc khác	82	332903	+ Dép nhựa, dép đi trong nhà người lớn
	12	THỰC PHẨM	83	332904	+ Giầy vải, thể thao người lớn
16	120401	+ Thịt lợn	84	332905	+ Giày dép trẻ em
17	120402	+ Thịt bò		V	THIẾT BỊ VÀ ĐỒ DÙNG GIA ĐÌNH
18	120403	+ Nội tạng động vật	85	514001	+ Máy điều hoà nhiệt độ
19	120404	+ Thịt gia súc đông lạnh	86	514101	+ Tủ lạnh
20	120501	+ Thịt gà	87	514201	+ Máy giặt
21	120502	+ Thịt gia cầm khác	88	514301	+ Máy hút bụi
22	120503	+ Thịt gia cầm đông lạnh	89	514302	+ Bình nước nóng nhà tắm
23	120601	+ Thịt quay, giò, chả	90	514303	+ Thiết bị gia đình lớn có động cơ
24	120602	+ Thịt hộp	91	514304	+ Máy vi tính và phụ kiện
25	120603	+ Thịt chế biến khác	92	514305	+ Máy in, máy chiếu, máy quét..
26	120701	+ Trứng tươi các loại	93	514306	+ Ổn áp điện
27	120702	+ Trứng đã chế biến	94	524401	+ Quạt điện
28	120801	+ Dầu thực vật	95	524402	+ Đèn điện thắp sáng

29	120802	+ Mỡ động vật	96	524403	+ Máy xay sinh tố, ép hoa quả
30	120901	+ Cá tươi, hoặc ướp lạnh	97	524404	+ Máy đánh trứng, trộn đa năng
31	120902	+ Tôm tươi hoặc ướp lạnh	98	524405	+ Bàn là điện
32	120903	+ Thuỷ hải sản tươi sống khác	99	524406	+ Đồ điện khác
33	121002	+ Thủy, hải sản khác chế biến	100	524501	+ Bếp gas
34	121101	+ Nước mắm, nước chấm	101	524502	+ Nồi cơm điện
35	121501	+ Mì chính (bột ngọt)	102	524503	+ Lò vi sóng, lò nướng, bếp từ
36	121502	+ Bột nêm, bột canh, viên súp	103	524504	+ Ấm, phích nước điện
37	121503	+ Muối ăn	104	524505	+ Trang thiết bị nhà bếp
38	121504	+ Đồ gia vị các loại	105	524506	+ Bếp đun không dùng điện, ga
39	121601	+ Đường	106	524601	+ Đồng hồ treo tường và để bàn
40	121602	+ Mật ong	107	524602	+ Gương treo tường
41	121701	+ Sữa tươi	108	524701	+ Giường
42	121702	+ Sữa đậu nành, sữa ngô	109	524702	+ Tủ các loại
43	121703	+ Sữa đặc	110	524703	+ Bàn, ghế, sa lông, tràng kỷ
44	121704	+ Sữa bột người lớn	111	524801	+ Đồ nhôm, inox
45	121705	+ Pho mát	112	524802	+ Đồ ăn, dao kéo làm bếp
46	121706	+ Kem	113	524803	+ Đồ kim loại khác
47	121707	+ Đậu phụ	114	524901	+ Đệm
48	121708	+ Sữa bột trẻ em	115	524902	+ Đồ dùng bằng nhựa
49	121709	+ Sữa chua	116	525001	+ Bát, đĩa
50	121801	+ Bánh quy, bánh nướng các loại	117	525002	+ Phích nước nóng
51	121802	+ Kẹo các loại	118	525003	+ Ly, cốc, lọ hoa
52	121803	+ Socola	119	525101	+ Chiếu, ga trải giường
53	121804	+ Mứt các loại	120	525102	+ Chăn, màn, gối
54	121901	+ Cà phê bột	121	525103	+ Rèm cửa
55	121902	+ Cà phê, ca cao hoà tan	122	525104	+ Thảm, tấm trải sàn
56	121903	+ Chè búp khô	123	525201	+ Xà phòng giặt
57	121904	+ Chè/trà nhúng uống liền	124	525202	+ Nước rửa bát và nước cọ sàn
58	121905	+ Các loại lá để uống khác	125	525203	+ Xà phòng tắm, nước tắm
	II	ĐỒ UỐNG VÀ THUỐC LÁ	126	525204	+ Dầu gội đầu
59	212101	+ Nước khoáng	127	525205	+ Kem đánh răng
60	212102	+ Nước giải khát có ga	128	525301	+ Công cụ cầm tay
61	212103	+ Nước quả ép	129	525302	+ Dụng cụ làm vườn
62	212104	+ Nước uống tăng lực đóng chai, lon, hộp	130	525303	+ Khoá các loại
63	222201	+ Rượu mạnh	131	525304	+ Pin, đèn pin
64	222202	+ Rượu vang	132	525305	+ Chổi
65	222301	+ Bia chai	133	525306	+ Giấy ăn
66	222302	+ Bia lon	134	525307	+ Giấy vệ sinh
67	232401	+ Thuốc lá	135	525308	+ Nến, diêm