Khai thác dữ liệu lớn trong việc tính chỉ số giá tiêu dùng

Hà Văn Sơn Nguyễn Thanh Bình

UEH University

KHAI THÁC DỮ LIỆU LỚN TRONG VIỆC TÍNH CHỈ SỐ GIÁ TIÊU DÙNG

Hà Văn Sơn [a], Nguyễn Thanh Bình [b]

[a] Trường Đại học Kinh tế TP.HCM, Email: hason@ueh.edu.vn
[b] Cục Thống kê TP.HCM, Email:ntbinhhcm@gso.gov.vn

TÓM TẮT

Chỉ số giá tiêu dùng (CPI) có một vị trí, vai trò rất quan trọng trong công tác quản lý và điều hành các chính sách vĩ mô của nhà nước. Hiện nay việc thu thập thông tin giá được thực hiện theo phương pháp truyền thống, điều tra viên trực tiếp đến thu thập thông tin tại các điểm bán lẻ. Phương pháp này có một số bất cập như có độ trễ trong việc công bố số liệu, khó khăn trong việc thu thập thông tin tại địa bàn, sai số phi chọn mẫu, chi phí cho điều tra viên thu thập thông tin.... Với sự phát triển của nền kinh tế số, sẽ phát sinh nguồn dữ liệu lớn có thể thay thế nguồn dữ liệu truyền thống. Bài viết tập trung vào việc khai thác dữ liệu lớn, là các thông tin giá tại các trang web trực tuyến để tính toán CPI tại Thành phố Hồ Chí Minh. Mặc dù cũng còn một số hạn chế, nhưng kết quả tính toán CPI từ dữ liệu lớn có nhiều ưu điểm, thể hiện đúng xu hướng và không có chênh lệnh nhiều so với chỉ số giá tiêu dùng truyền thống.

Từ khóa: *Chỉ số giá, Chỉ số giá tiêu dùng, Dữ liệu lớn, Khai thác dữ liệu lớn*

1. GIỚI THIỆU

Kinh tế số và chuyển đổi số là một trong những thuật ngữ được đề cập nhiều nhất trong thời gian gần đây, đặc biệt là trong bối cảnh tình hình đại dịch Covid_19 vẫn còn đang diễn biến phức tạp trên thế giới. Các chuyên gia kinh tế đều có nhận định là kinh tế số sẽ là xu thế tất yếu trong bối cảnh cuộc cách mạng công nghiệp 4.0, kinh tế số đã, đang và sẽ tác động mạnh mẽ đến kinh tế thế giới. Với dân số gần 100 triệu người, Việt Nam đang đứng trước tiềm năng lớn để phát triển nền kinh tế số và Việt Nam được xem là một trong những quốc gia có tốc độ phát triển kinh tế số khá tốt trong khu vực ASEAN.

Với xu hướng ngày càng phát triển của nền kinh tế số, các giao dịch mua bán, giới thiệu sản phẩm trực tiếp trên mạng Internet ngày càng phổ biến. Theo khảo sát của Cục Thương mại điện tử và Công nghệ thông tin thời gian gần đây, các doanh nghiệp sử dụng thương mại điện tử ngày càng

phổ biến và mở rộng khắp các địa phương trong cả nước. Nhiều doanh nghiệp đã vận dụng các mô hình kinh doanh hiện đại, áp dụng công nghệ mới vào các khâu sản xuất và lưu thông hàng hóa (Hiệp hội Thương mại Điện tử Việt Nam, 2016). Việc người dân ngày càng quan tâm và phát triển việc mua bán trên mạng ngày càng nhiều sẽ tạo ra nguồn dữ liệu vô cùng lớn. Chúng sẽ trở nên quý giá nếu chúng ta khai thác tốt, nó sẽ giúp ích rất nhiều cho công tác nghiên cứu khoa học, cho việc điều hành kinh tế vĩ mô của nhà nước và cũng rất hữu dụng cho doanh nghiệp trong việc đưa ra các giải pháp phát triển hoạt động sản xuất kinh doanh của doanh nghiệp.

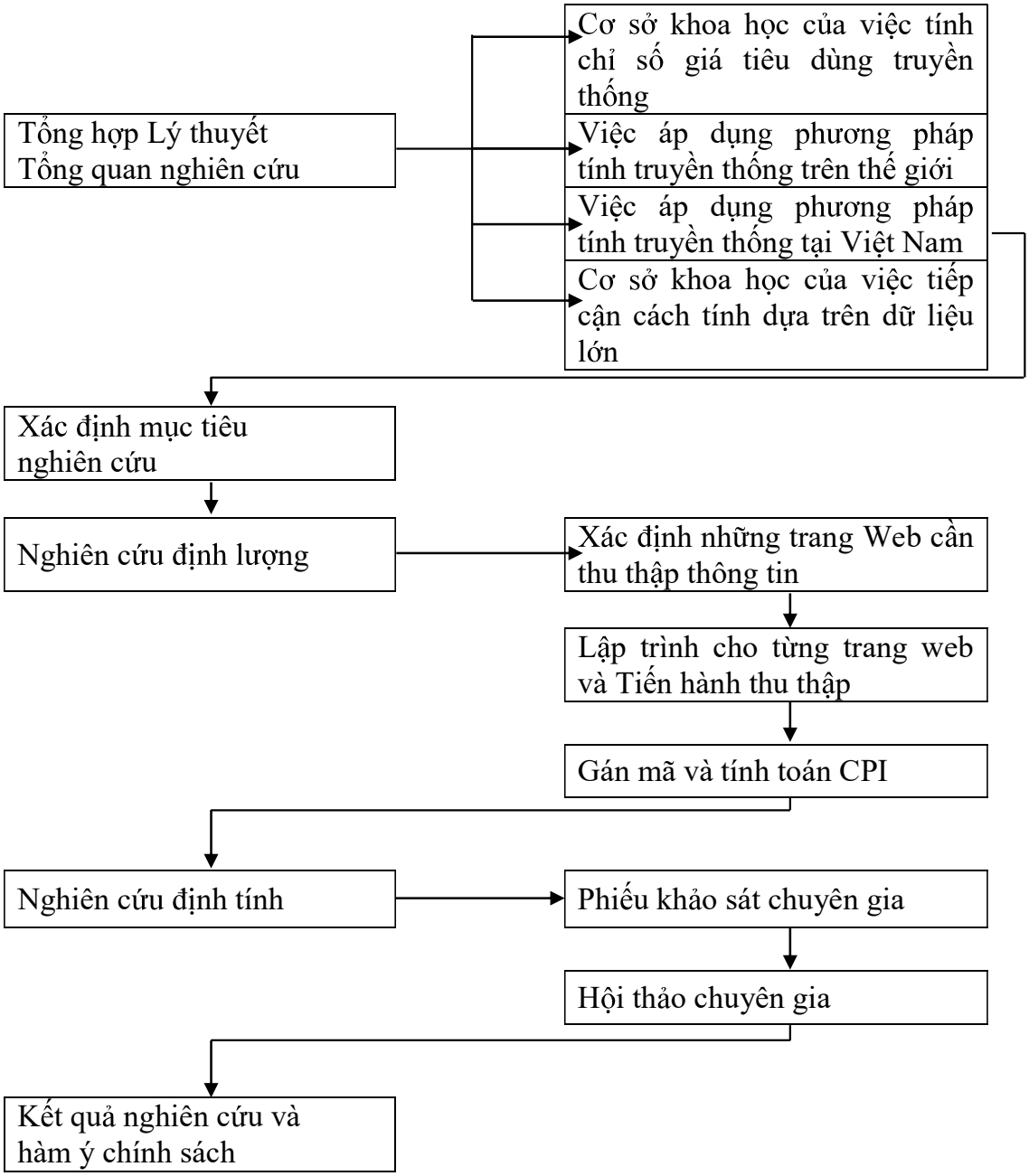
Chỉ số giá và các thông tin về giá cả thị trường có một vị trí, vai trò rất quan trọng trong công tác quản lý và điều hành các chính sách vĩ mô của nhà nước như các chính sách về quản lý tài chính tiền tệ, kiềm chế lạm phát, điều chỉnh lãi suất ngân hàng, điều chỉnh tỷ giá, ban hành các chính sách phát triển kinh tế xã hội theo vùng miền, các chính sách tiền lương..., qua đó góp phần phát triển hoạt động sản xuất kinh doanh và trao đổi thương mại quốc tế. Trong đó giá tiêu dùng và chỉ số giá tiêu dùng là một trong những chỉ số rất quan trọng. Ngành thống kê tiến hành điều tra, thu thập thông tin, tính chỉ số giá tiêu dùng và công bố hàng tháng. 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. Cuộc điều tra được thực hiện ở cả 63 tỉnh, thành phố và được công bố hàng tháng vào các ngày cuối tháng. Phương pháp thu thập thông tin giá tiêu dùng hiện nay vẫn thực hiện theo phương pháp truyền thống, điều tra viên sẽ trực tiếp đến các mạng lưới bán lẻ để thu thập thông tin về giá. Việc thu thập hiện nay gặp một số trở ngại như: Công tác thu thập tại địa bàn ngày một khó khăn hơn; Sai số phi chọn mẫu vẫn còn cao vì cũng còn hiện tượng điều tra viên chủ quan, không đến trực tiếp địa điểm kinh doanh để thu thập giá; chi phí cho cuộc điều tra lớn do phải huy động nhiều lực lượng điều tra viên.

Vì vậy, nghiên cứu giải pháp tận dụng nguồn dữ liệu lớn để tính chỉ số giá tiêu dùng ở Việt Nam là một việc rất cần thiết và phù hợp với xu hướng của thế giới. Chính vì những nguyên nhân này, nhóm tác giả tiến hành nghiên cứu đề tài: "Khai thác dữ liệu lớn trong việc tính chỉ số giá tiêu dùng ở Việt Nam (trường hợp Thành phố Hồ Chí Minh)" nhằm mong muốn đóng góp một phần vào kinh nghiệm khai thác dữ liệu lớn tại Việt Nam. Bài viết phân tích các vấn đề chính trong việc triển khai điều tra thu thập thông tin và tính toán chỉ số giá tiêu dùng ở Việt Nam hiện nay, giới thiệu phương pháp thu thập giá từ các trang web để tính toán chỉ số giá tiêu

dùng, các thách thức và đưa ra các khuyến nghị nhằm cải thiện chất lượng trong việc thu thập thông tin giá tiêu dùng.

2. ĐÁNH GIÁ THỰC TRẠNG CÔNG TÁC THỐNG KÊ GIÁ TIÊU DÙNG VÀ GIỚI THIỆU PHƯƠNG PHÁP KHAI THÁC DỮ LIỆU LỚN

Để thực hiện nghiên cứu này, nhóm tác giả thực hiện Quy trình nghiên cứu như sau:



Hình 1: Quy trình nghiên cứu

Nguồn: Đề xuất của nhóm tác giả

2.1. Đánh giá thực trạng công tác thống kê giá tiêu dùng

"Giá tiêu dùng là số tiền do người tiêu dùng phải chi trả khi mua một đơn vị hàng hóa hoặc dịch vụ phục vụ trực tiếp cho đời sống hàng ngày. Giá tiêu dùng được biểu hiện bằng giá bán lẻ hàng hóa trên thị trường hoặc giá dịch vụ phục vụ sinh hoạt đời sống dân cư. Trong trường hợp hàng hóa hoặc dịch vụ không có giá niêm yết, người mua có thể mặc cả thì giá tiêu dùng là giá người mua thực trả sau khi thỏa thuận với người bán. Chỉ số giá tiêu dùng là chỉ tiêu tương đối (tính bằng %) phản ánh xu hướng và mức độ biến động giá theo thời gian của các mặt hàng trong rổ hàng hóa và dịch vụ tiêu dùng đại diện" (Tổng cục Thống kê, 2018, trang 609).

Ở Việt Nam, để có thể tính và công bố được chỉ số giá tiêu dùng hàng tháng, quí, năm như hiện nay, ngành thống kê thực hiện cuộc điều tra giá tiêu dùng (Tổng cục Thống kê, 2015) với những nội dung chính trong phương án điều tra giá tiêu dùng giai đoạn 2014-2019 như sau:

Đơn vị điều tra: các sạp, quầy hàng tại các chợ, điểm bán hàng (chuyên bán lẻ), các cơ sở kinh doanh dịch vụ,… có địa điểm kinh doanh ổn định trong những khu vực điều tra đã được chọn mẫu.

Phạm vi điều tra: toàn bộ tất cả những điểm điều tra tại tất cả tỉnh, thành phố trực thuộc Trung ương được Tổng cục Thống kê chọn.

Thời điểm điều tra giá tiêu dùng: căn cứ vào thời điểm điều tra, 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. Tổng số mặt hàng lấy giá là: 654 mặt hàng. Số lượng cụ thể các mặt hàng theo từng kỳ và từng điểm điều tra như sau: có 126 mặt hàng lấy giá tại 1 đến 3 nơi điều tra trong mỗi khu vực điều tra và lấy 3 lần/tháng; có 50 mặt hàng lấy giá tại 1 đến 3 nơi điều tra trong mỗi khu vực điều tra và lấy 1 lần/tháng; có 18 mặt hàng lấy giá tại 1 nơi điều tra trong mỗi khu vực điều tra và lấy 3 lần/tháng; có 453 mặt hàng lấy giá tại 1 nơi điều tra trong mỗi khu vực điều tra và lấy 1 lần/tháng; có 5 mặt hàng lấy giá tại 1 nơi điều tra trong mỗi khu vực điều tra và lấy theo số lần phát sinh trong tháng.

Loại điều tra: Đây là cuộc điều tra chọn mẫu, được thực hiện theo các bước sau:

- Xây dựng dàn mẫu điều tra là danh mục mặt hàng đại diện: dựa vào danh mục điều tra giá tiêu dùng chung cả nước, các tỉnh, thành phố trực thuộc Trung ương tiến hành rà soát và xác định danh mục điều tra cụ thể cho địa phương của mình và danh mục này được dùng làm cơ sở để thu thập giá. Danh mục điều tra giá của địa phương phải đảm bảo hai tiêu chí: một là phải có trong danh mục chung của cả nước; Hai là phải đảm bảo là hàng hóa và dịch vụ phổ biến tiêu dùng tại địa phương. Một yêu cầu bắt buộc để thu thập giá được chính xác và đảm bảo đồng nhất khi so sánh là phải mô tả chi tiết nhãn mác, phẩm cấp, quy cách, cụ thể các loại hàng hóa và dịch vụ trong danh mục điều tra. Ngoại trừ trong danh mục chuẩn, các hàng hóa và dịch vụ phải thống nhất nhãn mác, phẩm cấp, quy cách trên phạm vi cả nước, các mặt hàng và dịch vụ còn lại có thể được chọn theo đặc điểm tiêu dùng của từng tỉnh, thành phố do mỗi địa phương có mức sống, đặc điểm vùng miền và tập quán tiêu dùng khác nhau.

- Thu thập giá kỳ gốc: Sau khi tiến hành rà soát và xác định được danh mục hàng hóa và dịch vụ đại diện của tỉnh, thành phố, các tỉnh, thành phố sẽ tiến hành lập bảng giá kỳ gốc.

Phương pháp xử lý thông tin

- Xây dựng quyền số giá tiêu dùng: Quyền số tính chỉ số giá tiêu dùng là cơ cấu chi tiêu các nhóm hàng hóa và dịch vụ trong tổng chi tiêu của hộ gia đình. Quyền số để tính chỉ số giá tiêu dùng của Việt Nam là cơ cấu chi tiêu của từng vùng so với tổng chi tiêu của cả quốc gia chia theo từng nhóm hàng.

- Cấu trúc của chỉ số giá tiêu dùng: Cấu trúc của chỉ số giá tiêu dùng được Tổng cục Thống kê xây dựng luôn đảm bảo 2 yêu cầu: Một là đảm bảo tính liên tục của chuỗi chỉ số giá tiêu dùng qua thời gian, hai là phải phù hợp với cơ cấu tiêu dùng của hộ gia đình trong giai đoạn hiện tại. Hiện nay chỉ số giá tiêu dùng có cấu trúc như sau: Nhóm cấp 1 bao gồm 11 nhóm, nhóm ngành cấp 2 có 32 nhóm, cấp 3 có 86 nhóm và cấp 4 có 266 nhóm.

- Công thức áp dụng tính chỉ số giá tiêu dùng: Áp dụng công thức Laspeyres bình quân nhân để tính chỉ số giá tiêu dùng.

Từ phương án triển khai điều tra của Tổng cục Thống kê và thực tiễn thu thập thông tin giá tại địa phương, dựa vào nghiên cứu của Berry và các cộng sự (2019) đưa ra các đánh giá về các thông lệ quốc tế tốt nhất trong

việc tính toán chỉ số giá tiêu dùng, có thể thấy được những ưu điểm trong việc tính chỉ số giá tiêu dùng ở Việt Nam như:

- Tần suất công bố và tính kịp thời của dữ liệu cung cấp một dấu hiệu về việc cải thiện chất lượng và mức độ phù hợp. Thông thường thì tần suất phổ biến số liệu gia tăng đồng thời với những cải tiến về phương pháp luận. IMF đã thiết lập các tiêu chuẩn phổ biến dữ liệu kinh tế và tài chính, trong đó có cả chỉ số giá tiêu dùng. Tiêu chuẩn Phổ biến dữ liệu đặc biệt (SDDS) quy định việc phổ biến CPI hàng tháng, còn về thời gian công bố số liệu CPI, tốt nhất theo thông lệ quốc tế là trong vòng 1 tháng sau tháng tham chiếu. Việt Nam đã đáp ứng được cả 2 tiêu chuẩn này với việc công bố CPI hàng tháng vào các ngày cuối tháng.

- Để tăng cường khả năng so sánh quốc tế của CPI, các tổ chức quốc tế khuyến nghị các quốc gia thống nhất sử dụng Bảng phân loại tiêu dùng cá nhân theo mục đích (COICOP), đã được các quốc gia thông qua dưới sự bảo trợ của Liên hợp quốc. Việt Nam cũng đang sử dụng bảng phân loại này.

- Thời gian cập nhật quyền số là một chỉ số về độ tin cậy và độ chính xác của dữ liệu, là cơ sở rất quan trọng để đánh giá chất lượng của số liệu thống kê. Tài liệu cẩm nang hướng dẫn CPI năm 2004 (ILO, 2014) khuyên việc cập nhật các quyền số CPI không quá 5 năm. Quyền số CPI ở Việt Nam hiện nay được cập nhật theo chu kỳ 5 năm.

- Ngoài ra, phạm địa lý của CPI là một thành phần quan trọng khác trong việc đánh giá tính hợp lý của phương pháp luận. Phạm vi địa lý đề cập đến phạm vi tính quyền số chi tiêu hoặc phạm vi thu thập giá thực tế. Tốt nhất là hai phạm vi này phải trùng nhau và chỉ số này nên bao gồm chi tiêu của tất cả các hộ gia đình, bao gồm cả thành thị, nông thôn và trong cả nước. CPI của Việt Nam hiện nay đã đáp ứng được cả 2 yêu cầu này.

Tuy nhiên cũng còn có một số bất cập sau:

- Công tác thu thập tại địa bàn ngày một khó khăn hơn. Tỷ lệ các điểm thu thập giá không hợp tác, gây khó khăn trong quá trình điều tra viên thu thập thông tin không giảm vì nhiều nguyên nhân như nhu cầu bảo mật thông tin kinh doanh, điều tra viên khai thác thông tin vào lúc cao điểm bán hàng nên sẽ gặp trở ngại trong việc cung cấp thông tin, thông tin thu thập nhiều, ý thức chấp hành chế độ báo cáo, điều tra thống kê còn hạn chế... Ngoài ra, nếu thời gian thu thập thông tin đúng vào những ngày địa phương thực hiện giãn cách xã hội vì một số lý do như dịch bệnh Covid-19 vừa

qua... hoặc thời gian thu thập thông tin đúng vào những ngày nghĩ Lễ dài hạn như Tết Âm lịch, Lễ 30/4 (Ngày Giải phóng hoàn toàn miền Nam, thống nhất đất nước), Lễ 1/5 (Ngày Quốc tế Lao động)..., phần lớn các cơ sở kinh doanh không mở cửa bán hàng cũng là một vấn đề lớn cần giải quyết trong quá trình thu thập giá tại địa bàn. Bên cạnh đó, giá cả hàng hóa và dịch vụ biến động lớn trong những ngày này cũng là một vấn đề khi xử lý, tính toán giá tiêu dùng.

- Sai số phi chọn mẫu vẫn còn phát sinh. Cũng còn nhiều điều tra viên chủ quan, không đến trực tiếp địa điểm kinh doanh để thu thập giá mà chỉ điện thoại nắm tình hình, thậm chí một số điều tra viên còn không liên lạc với đơn vị kinh doanh mà thu thập giá dựa trên nhận định chủ quan của mình. Các khu vực kinh doanh ngày càng mở rộng, việc gia tăng số lượng các cơ sở kinh doanh nhưng theo phương án điều tra thì lại không tăng số lượng mẫu khảo sát, không tính đến tỷ trọng doanh số bán hàng, nhiều cửa hàng trong mẫu điều tra ngày càng thu hẹp kinh doanh, thị phần giảm sút, ảnh hưởng đến chất lượng số liệu tổng hợp.

- 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. Ngoài ra việc chọn mẫu và tính toán số lượng các điểm bán hàng bình ổn ở địa phương, các siêu thị, trung tâm thương mại và các tập đoàn bán lẻ lớn cũng là một vấn đề cần phải giải quyết.

- Cuối cùng, một hạn chế lớn nhất trong việc thu thập giá truyền thống là chi phí cho cuộc điều tra lớn do phải huy động nhiều lực lượng điều tra viên. Hàng tháng ngành thống kê Thành phố Hồ Chí Minh phải triển khai thu thập giá 3 kỳ tại 8 quận, huyện, mỗi quận huyện thu thập từ 554 đến 651 mặt hàng, do đó phải huy động lực lượng điều tra viên nhiều và khá tốn kém.

Với nền tảng của cuộc cách mạng công nghiệp 4.0, cùng với sự phát triển của kinh tế số giúp tạo ra những nguồn dữ liệu mới là cơ hội tuyệt vời cho ngành thống kê cải thiện chất lượng, nâng cao hiệu quả công tác thu thập thông tin. Trong đó, nguồn dữ liệu giá thu thập trực tuyến với nhiều ưu điểm sẽ mang lại cơ hội tốt cho ngành thống kê xử lý các thách thức mà thống kê giá tiêu dùng truyền thống đang đối mặt. Với nguồn dữ liệu giá thu thập trực tuyến này sẽ giúp công tác thống kê giá đo lường chính xác hơn sự thay đổi giá, giúp mở rộng cỡ mẫu, việc người tiêu dùng sử dụng các mặt hàng thay thế sẽ được phản ánh chính xác hơn, sự thay đổi

chất lượng được xử lý tốt hơn, giảm hoặc loại bỏ áp lực từ người trả lời và giảm chi phí điều tra. Ngoài ra, nguồn dữ liệu thu thập từ các trang Web giúp cải thiện thời gian thu thập, chi tiết dữ liệu nhiều và đa dạng hơn, dữ liệu được thu thập với tần suất cao hơn. Chính vì vậy, nhóm tác giả đã tiến hành nghiên cứu và đề xuất giải pháp khai thác dữ liệu lớn trong việc tính toán chỉ số giá tiêu dùng tại Thành phố Hồ Chí Minh.

2.2. Khai thác dữ liệu lớn trong việc tính toán chỉ số giá tiêu dùng tại Thành phố Hồ Chí Minh và một số kết quả đạt được

2.2.1. Khái niệm dữ liệu lớn

Có rất nhiều định nghĩa về dữ liệu lớn nhưng không có một định nghĩa thật chính xác cho khái niệm dữ liệu lớn. Ý tưởng đầu tiên cho khái niệm dữ liệu lớn xuất phát từ việc dung lượng thông tin đã tăng quá lớn tới mức không còn vừa vào các bộ nhớ máy vi tính dùng để xử lý. Như vậy các chuyên gia công nghệ thông tin phải cải tạo các công cụ họ đang dùng để có thể phân tích được đầy đủ tất cả thông tin.Những công nghệ này cho phép ta quản lý những khối lượng dữ liệu lớn hơn nhiều so với trước đây và quan trọng là không cần đưa dữ liệu vào các bảng cơ sở dữ liệu cổ điển. Như vậy dữ liệu lớn là một thuật ngữ chỉ bộ dữ liệu lớn hoặc phức tạp mà các phương pháp truyền thống không đủ các ứng dụng để xử lý dữ liệu này. Vào năm 2001, nhà phân tích Doug Laney của hãng META đã nói rằng những thách thức và cơ hội nằm trong việc tăng trưởng dữ liệu có thể được mô tả bằng ba chiều: tăng về lượng (Volume), tăng về vận tốc (Velocity) và tăng về chủng loại (Variety). Giờ đây, Gartner cùng với nhiều công ty và tổ chức khác trong lĩnh vực công nghệ thông tin tiếp tục sử dụng mô hình "3V" này để định nghĩa nên Big Data. Đến năm 2012, Gartner bổ sung thêm hai tính chất quan trọng trong định nghĩa về dữ liệu lớn là Độ tin cậy/chính xác (Veracity) và Giá trị thông tin (Value).

2.2.2. Tình hình sử dụng dữ liệu lớn trong công tác thống kê của các quốc gia trên thế giới

Hiện nay, rất nhiều cơ quan thống kê quốc gia của các nước đang triển khai nghiên cứu sử dụng nguồn dữ liệu lớn để sản xuất số liệu thống kê chính thức. Một số nước đã xây dựng hẳn một chiến lược về dữ liệu lớn như: Trung Quốc, Đan mạch, Phần Lan, Ý, Nhật, Bồ Đào Nha, Romania, Serbia, Thụy Điển, Anh. Một số dự án về dữ liệu lớn đang được thống kê các nước triển khai thực hiện như sau:

Thống kê Cameroon thực hiện các dự án: Xây dựng năng lực trong sử dụng Big data như nguồn số liệu thống kê chính thức; xây dựng năng lực cho việc sử dụng Big data cho mục đích thống kê.

Thống kê Anh đang thực hiện các dự án: Khai thác cơ sở dữ liệu về thị trường thương mại để ước tính số liệu điều tra dân số; Kiểu dữ liệu Smartmeter cho cấu trúc hộ gia đình/quy mô và nghề nghiệp; tiềm năng dữ liệu Smartmeter để phát hiện ngôi nhà vắng chủ; Dữ liệu điện thoại di động tích hợp để xác định mẫu.

Thống kê Mexico thực hiện các dự án: Phân tích Tweet; Thống kê chính thức sử dụng dữ liệu định vị của điện thoại di động với một ứng dụng cụ thể để xây dựng lưới dân số.

Thống kê Trung Quốc thực hiện các dự án: Chỉ số thống kê doanh nghiệp dữ liệu lớn; Thay đổi giá trực tuyến của các phương tiện sản xuất trong khu vực Zhuochang ở Sơn Đông.

Thống kê Canada thực hiện các dự án: Không gian nhà ở: Nghiên cứu khả thi; Chỉ số thị trường nhà (dựa trên thông tin trang web).

Thống kê Romania thực hiện dự án: Sử dụng dữ liệu máy quét.

Thống kê Nam Phi thực hiện dự án: Sử dụng dữ liệu máy quét để tính chỉ số giá tiêu dùng.

Thống kê Ý thực hiện dự án: Ước tính di động dựa trên dữ liệu điện thoại di động; Sử dụng dữ liệu máy quét cho chỉ số giá tiêu dùng; Internet như một nguồn dữ liệu cho việc sử dụng công nghệ thông tin và truyền thông của các doanh nghiệp và các tổ chức công.

Bảng 1. Các dự án thống kê chính thức liên quan đến giá có sử dụng dữ liệu lớn

STT	Tên dự án	Quốc gia	Đơn vị thực hiện	Nguồn dữ liệu
1	CPI với dữ liệu máy quét	Áo	Thống kê Áo	Dữ liệu từ máy quét
2	Thu thập giá tự động trên Internet: Sử dụng web scraping và web crawlers để	Áo	Thống kê Áo	Dữ liệu web scraping

STT	Tên dự án	Quốc gia	Đơn vị thực hiện	Nguồn dữ liệu
	tính chỉ số giá			
3	Sử dụng dữ liệu máy quét cho chỉ số giá	Bỉ	Thống kê Bỉ	Dữ liệu từ máy quét
4	Sử dụng dữ liệu giá web scraping để lập chỉ số giá thương mại điện tử	Trung Quốc	Cục Thống kê quốc gia	Dữ liệu web scraping
5	Sử dụng dữ liệu máy quét để tính CPI	Đan Mạch	Thống kê Đan Mạch	Dữ liệu từ máy quét
6	Nghiên cứu khả thi về việc tạo ra các chỉ số bằng cách sử dụng web scraping	Ecuador	Viện Thống kê về điều tra quốc gia	Dữ liệu web scraping
7	Số liệu thống kê giá tiêu dùng đa mục đích: dữ liệu máy quét	Europe	Ủy ban châu Âu - Eurostat	Dữ liệu từ máy quét
8	Hiện đại hóa việc thu thập và biên soạn giá với dữ liệu web và máy quét	Phần Lan	Thống kê Phần Lan	Dữ liệu Web scraping và máy quét
9	Web scraping cho thống kê giá	Đức	Cục Thống kê Liên bang	Dữ liệu web scraping
10	Dữ liệu web scraping từ các trang web bán lẻ để tính toán CPI	Hungari	Cục Thống kê Trung ương	Dữ liệu web scraping
11	Dữ liệu từ máy quét để tính CPI	Israel	Cục Thống kê Trung ương	Dữ liệu web scraping và máy quét

STT	Tên dự án	Quốc gia	Đơn vị thực hiện	Nguồn dữ liệu
12	Sử dụng dữ liệu máy quét cho CPI	Ý	Viện Thống kê quốc gia	Dữ liệu từ máy quét
13	Dữ liệu web scraping và dữ liệu máy quét cho thống kê giá	Nhật	Bộ Nội vụ và Truyền thông	Dữ liệu web scraping và máy quét
14	Sử dụng dữ liệu máy quét cho thống kê giá	Luxembourg	Viện Thống kê quốc gia	Dữ liệu từ máy quét
15	Thông tin giá dựa trên thông tin máy quét và các trang web	Hà Lan	Thống kê Hà Lan	Dữ liệu web scraping
16	Sử dụng dữ liệu trực tuyến trong HICP	Na Uy	Thống kê Na Uy	Dữ liệu web scraping
17	Thay đổi giá trực tuyến	Trung Quốc	Viện Thống kê quốc gia	Dữ liệu web scraping
18	Chỉ số giá trực tuyến	Hàn Quốc	Thống kê Hàn Quốc	Dữ liệu web scraping
19	Sử dụng dữ liệu máy quét cho thống kê giá và thống kê kinh tế	Romania	Viện Thống kê quốc gia	Dữ liệu từ máy quét
20	Thử nghiệm các công cụ web scraping và dữ liệu máy quét để thống kê giá	Slovenia	Cục Thống kê	Dữ liệu Web scraping và máy quét
21	Đánh giá việc sử dụng dữ liệu máy	Nam Phi	Thống kê	Dữ liệu từ

STT	Tên dự án	Quốc gia	Đơn vị thực hiện	Nguồn dữ liệu
	quét để tổng hợp chỉ số giá tiêu dùng		Nam Phi	máy quét
22	Web scraping về giá cho CPI	Tây Ban Nha	Viện Thống kê quốc gia	Dữ liệu web scraping
23	Sử dụng dữ liệu máy quét cho thống kê giá	Thụy Sĩ	Cục Thống kê Liên bang	Dữ liệu từ máy quét
24	Thu thập giá với dữ liệu máy quét	Thụy Sĩ	Cục Thống kê Liên bang	Dữ liệu từ máy quét
25	Web Scraper và giao diện trình ứng dụng (API): khám phá web scraping về dữ liệu giá	Mỹ	Văn phòng Quản lý và Ngân sách	Dữ liệu web scraping
26	Khám phá giá theo thời gian thực trên thị trường hàng hóa tương lai	Mỹ	Văn phòng Quản lý và Ngân sách	Khác

Nguồn: Tổng hợp của tác giả từ trang web www.unstats.un.org

2.2.3. Khai thác dữ liệu lớn tính chỉ số giá tiêu dùng

Nhóm nghiên cứu sử dụng phương pháp nghiên cứu định lượng kết hợp với nghiên cứu định tính (thông qua bảng hỏi chuyên gia và hội thảo chuyên gia) để thực hiện đề tài và đã xây dựng được các Quy trình: Quy trình nghiên cứu; Quy trình tính toán chỉ số giá tiêu dùng khai thác từ dữ liệu lớn.

Thời điểm thu thập gồm 3 kỳ:

- Kỳ 1 thu thập vào ngày 01 tháng báo cáo;

- Kỳ 2 thu thập vào ngày 10 tháng báo cáo;

- Kỳ 3 thu thập vào ngày 20 tháng báo cáo.

Thời gian thu thập: Bắt đầu từ năm 2017.

Quy trình sản xuất chỉ số giá tiêu dùng truyền thống tuân thủ theo "quy trình sản xuất thông tin thống kê" bao gồm các bước sau: Xác định nhu cầu thông tin; Chuẩn bị thu thập thông tin; Thu thập thông tin; Xử lý thông tin; Phân tích thông tin; Phổ biến thông tin; Lưu trữ thông tin.

Quy trình sản xuất chỉ số giá tiêu dùng dựa trên việc thu thập thông tin từ dữ liệu lớn cũng phải tuân thủ theo quy trình sản xuất thông tin thống kê của Tổng Cục Thống kê, tuy nhiên sẽ tập trung xử lý phức tạp hơn ở bước 3: Thu thập thông tin và bước 4: Xử lý thông tin.

Để đảm bảo tính thống nhất khi so sánh với chỉ số giá tiêu dùng theo phương pháp truyền thống, nghiên cứu thực hiện đúng theo phương án của Tổng cục Thống kê là tiến hành thu thập giá 3 lần 1 tháng trên các trang web bán hàng trực tuyến lớn, có đầy đủ thông tin. Ứng dụng thu thập thông tin giá của nghiên cứu vận hành chạy trên môi trường internet. Thông tin lưu trữ bằng công nghệ điện toán đám mây, ngoài ra còn kết hợp thêm việc lưu trữ bằng mạng nội bộ.

Bảng 2. Quy trình thu thập thông tin từ nguồn dữ liệu lớn

STT	**Nội dung**
1	Xác định những trang nguồn cần thu thập thông tin
2	Phân tích cấu trúc các trang web được chọn, nhận diện các thông tin cần lấy
3	Lập trình cho từng trang web để thu thập dữ liệu
4	Lập trình xây dựng bảng cấu trúc để nhóm các thông tin giống nhau lại
5	Tiến hành thu thập thí điểm
6	Trao đổi, rút kinh nghiệm về những điều chỉnh cấu trúc dữ liệu
7	Tổ chức, hoàn chỉnh lại cấu trúc dữ liệu
8	Tiến hành thu thập chính thức
9	Kiểm tra các trường hợp như: trang web nguồn đổi cấu trúc, chương trình thu thập bị chặn …
10	Nếu có vấn đề trong quá trình kiểm tra: Lập trình điều chỉnh lại

STT	Nội dung
11	Gán mã CPI. Nhận diện các hạng mục mới xuất hiện hay có thay đổi
12	Lưu cơ sở dữ liệu vào máy chủ nội bộ

Nguồn: tác giả tổng hợp

2.2.4.Quy trình tính toán chỉ số giá tiêu dùng khai thác từ dữ liệu lớn

Các bước thực hiện trong quy trình tổng hợp chỉ số giá tiêu dùng được thu thập từ dữ liệu lớn cụ thể như sau:

Bước 1: Thực hiện tính chỉ số giá từng mặt hàng cụ thể.

Bước 2: Thực hiện tính chỉ số giá của nhóm hàng hóa và dịch vụ cấp 4 kỳ báo cáo so với kỳ trước.

Phương pháp bình quân nhân giản đơn được sử dụng để tính chỉ số giá tiêu dùng của nhóm hàng hóa và dịch vụ cấp 4, công thức tính toán cụ thể sau đây:

$$I_p^{IV} = \sqrt[n]{\prod_{i=1}^{n}(i_{pi})} \quad (1)$$

Trong đó:

I_p^{IV} : chỉ số giá nhóm hàng hóa và dịch vụ cấp 4;

i_{pi} : chỉ số giá cá thể của các loại hàng hóa hoặc dịch vụ i trong nhóm hàng hóa và dịch vụ cấp 4 cần tính;

n: số mặt hàng tham gia tính chỉ số nhóm cấp 4.

Bước 3: Thực hiện tính chỉ số giá của nhóm hàng hóa cấp thấp nhất (cấp 4) so kỳ gốc.

Bước 4: Thực hiện tính chỉ số giá của nhóm hàng hóa, dịch vụ các cấp còn lại (từ cấp 3 trở lên) đến cấp 1 và chỉ số chung so kỳ gốc, ở bước này bắt đầu sử dụng quyền số.

Với thông tin giá thu được từ dữ liệu lớn, nghiên cứu thực hiện tính toán CPI theo các bước nêu ở trên (phương pháp dữ liệu lớn). Ngoài ra nghiên cứu cũng tiến hành tính toán chỉ số giá tiêu dùng bằng cách kết hợp với thông tin giá thu được bằng phương pháp truyền thống (kết hợp 1 và kết hợp 2). Cụ thể 2 cách tính kết hợp như sau:

- Cách tính kết hợp 1: Xem kết quả tính toán CPI được thu thập từ dữ liệu lớn đại diện cho khu vực thành thị. CPI chung được tổng hợp từ CPI dữ liệu lớn với quyền số là tỷ trọng chi tiêu của khu vực thành thị và CPI truyền thống (khu vực nông thôn) với quyền số là tỷ trọng chi tiêu của khu vực nông thôn (quyền số này là cố định, 5 năm mới thay đổi 1 lần). Cách tính chỉ số giá chung của toàn thành phố: Chỉ số giá của thành phố được tính từ chỉ số của các nhóm hàng tương ứng giữa hai khối: thu thập trực tuyến và thu thập theo phương pháp truyền thống. Quyền số ngang được sử dụng để tính chỉ số giá cả thành phố theo các nhóm hàng từ cấp 4 đến cấp 1 và chỉ số chung. Quyền số ngang là cơ cấu chi tiêu của 2 khu vực thành thị và nông thôn chia theo từng nhóm hàng.

- Cách tính kết hợp 2: Xem kết quả tính toán CPI được thu thập từ dữ liệu lớn đại diện cho khối doanh nghiệp. CPI truyền thống đại diện cho khối cá thể vì phần lớn mạng lưới được thu thập từ các chợ truyền thống. CPI chung được tổng hợp từ CPI dữ liệu lớn với quyền số là tỷ trọng doanh thu bán lẻ của khối doanh nghiệp và CPI truyền thống với quyền số là tỷ trọng doanh thu bán lẻ của khối cá thể. Quyền số này có thể tính toán được hàng tháng từ doanh thu tổng mức bán lẻ chia theo từng nhóm mặt hàng. Cách tính chỉ số giá thành phố: Chỉ số giá của thành phố được tính từ chỉ số của các nhóm hàng tương ứng giữa hai khối: thu thập trực tuyến và thu thập theo phương pháp truyền thống. Quyền số ngang được sử dụng để tính chỉ số giá cả thành phố theo các nhóm hàng từ cấp 4 đến cấp 1 và chỉ số chung. Quyền số ngang là tỷ trọng tổng mức bán lẻ trên địa bàn Thành phố Hồ Chí Minh của hai khối doanh nghiệp và cá thể phân theo các nhóm hàng.

Bảng 3. Số lượng trang web và số lượng mặt hàng thu thập

STT	Tên trang web	Đơn vị quản lý	Số lượng
01	adayroi.com	Cty CP DV TM TH Vincommerce	17.403
02	Aeoneshop.com	Cty TNHH Aeon Việt Nam	8.913
03	bachhoaxanh.com	Cty CP Bách Hóa Xanh	4.732
04	baza.vn	Cty CP Baza Việt Nam	4.937
05	concung.com	Cty CP Con Cưng	2.482

STT	Tên trang web	Đơn vị quản lý	Số lượng
06	csfood.vn	Cty TNHH TM XD Toàn Lực	2.572
07	dienmaycholon.vn	Cty TNHH Cao Phong	4.424
08	foodandy.vn	Cty TNHH AN&D	36
09	fptshop.com.vn	Cty CP Bán Lẻ Kỹ Thuật Số FPT	661
10	hc.com.vn	Cty TNHH Thương mại VHC	2.689
11	juno.vn	Cty CP SX thương mại dịch vụ JUNO	272
12	muachung.vn	Cty CP VCCorp	352
13	pico.vn	Cty CP Pi Co	10.275
14	shop2banh.vn	Cty TNHH Truyền thông số	659
15	thucphamnhanh.com	Cty TNHH Thực phẩm nhanh	71
16	tiki.vn	Cty CP Ti Ki	21.442
17	vascara.com	Cty CP Thương Mại Global Fashion	133
18	carmudi.vn	Cty TNHH MTV Xe Classified	4.463
19	dienmaythienhoa.vn	TT Điện Máy & Nội Thất Thiên Hòa	2.546
20	fahasa.com	Cty CP Phát Hành Sách TPHCM	273
21	tuticare.com	Công ty CP VEETEX	6.154
22	vatgia.com	Công ty Cổ phần Vật Giá Việt Nam	8.261
23	vinabook.com	Cty CP TM DV Mê Kông Com	90
24	yes24.vn	Cty TNHH Hansaeyes24 Vina	56.334
25	Hieusach.vn	Cty CP truyền thông Văn Hóa Việt	52.557
26	adayroi.com	Cty CP DV TM TH Vincommerce	17.403

STT	Tên trang web	Đơn vị quản lý	Số lượng
27	Aeoneshop.com	Cty TNHH Aeon Việt Nam	8.913
28	laptopxachtayshop.com	Cửa hàng laptop xách tay	974
	TỔNG CỘNG		**246.069**

Nguồn: Tác giả tổng hợp

Tổng số trang Web được thu thập dữ liệu là 28 trang Web với 246.069 mặt hàng. Trang web có số lượng mặt hàng lớn nhất là yes24.vn với 56.334 mặt hàng. Đây là trang web bán rất nhiều loại nhóm hàng như kim khí điện máy, máy vi tính, điện thoại di động, mỹ phẩm, thời trang nam nữ, giày dép, dây nịt, đồ chơi trẻ em, đồ dùng cho mẹ và bé, gia dụng, đồ dùng nhà bếp, đồ dùng, dụng cụ thể thao, đồ trang sức, trang trí nội thất… Trang web có số lượng mặt hàng ít nhất là foodandy.vn với 36 mặt hàng. Đây là trang web bán thịt bò, cừu, cá hồi… Sau khi xử lý dữ liệu và mã hóa, tất cả 246.069 mặt hàng này được đưa vào tính toán CPI.

Các nhóm hàng thu thập được nhiều mặt hàng là: sách các loại (52.568 mặt hàng), quần áo may sẵn khác (21.286 mặt hàng), hàng chăm sóc cơ thể (8.450 mặt hàng), túi xách, va ly, ví (7.280 mặt hàng), đồ trang sức (5.436 mặt hàng), giầy dép (sandan) da, nữ người lớn (4.665 mặt hàng), máy vi tính và phụ kiện (5.636 mặt hàng)…

Các nhóm hàng có số lượng mặt hàng thu thập được tương đối ít là các loại rau, quả như: su hào, rau muống, rau chế biến các loại, măng tươi, chuối, rượu mạnh, rượu nhẹ, gạo, hoa tươi… Trung bình chỉ thu được từ 4-20 mặt hàng.

Một số nhóm hàng thống nhất trong toàn quốc: bao gồm 4 nhóm hàng là vé tàu hỏa, vé máy bay, dịch vụ bưu điện, dịch vụ viễn thông được cập nhật thống nhất từ thông báo chính thức hàng tháng của Tổng cục Thống kê.

Các nhóm hàng chưa thu thập được từ các trang web là:

Nhóm hàng đặc thù (8 mặt hàng), không thể thu thập được trên web, do đó phải sử dụng giá thu thập truyền thống là: Nhà chủ sở hữu tính quy

đổi, nước sinh hoạt, dịch vụ nước sinh hoạt, điện sinh hoạt, dịch vụ điện sinh hoạt, gas, dầu hỏa….

Một số nhóm hàng trong danh mục hàng hóa và dịnh vụ tính CPI, không có trong 28 trang web kể trên (117 mặt hàng) nên khi tính toán phải sử dụng giá thu thập truyền thống. Các mặt hàng này chủ yếu là các loại dịch vụ như: tiền công may quần áo, giặt là quần áo, thuê quần áo, nhà ở thuê, sửa chữa thiết bị, khám chữa bệnh nội trú, thuê đồ dùng, học phí… và sẽ được thu thập vào giai đoạn 2 của dự án.

2.2.5. Một số kết quả tính toán CPI

Kết quả tính toán thực nghiệm từ dữ liệu lớn đã thu thập và tổng hợp được cụ thể như sau:

Bảng 4. Chỉ số giá tiêu dùng so tháng trước chia theo tháng

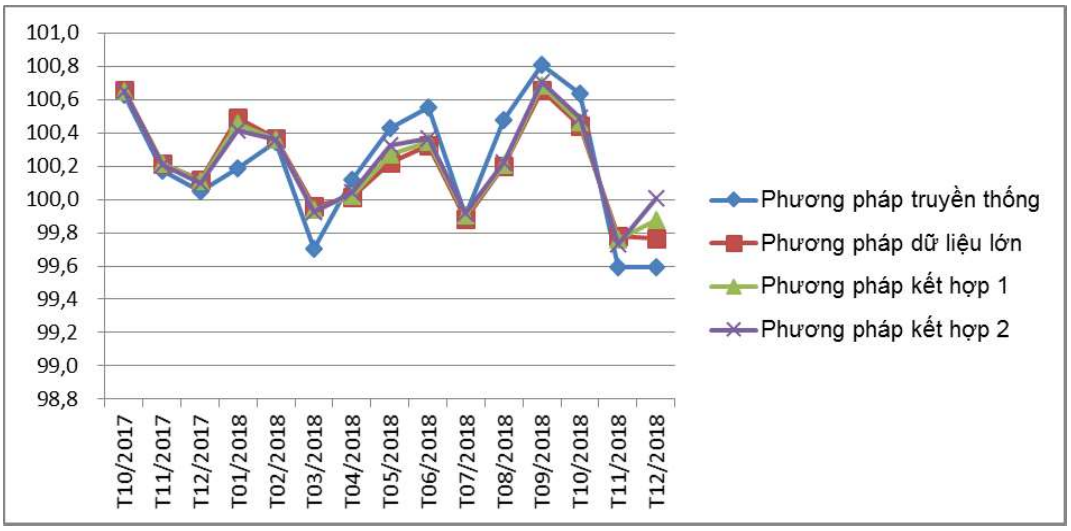
ĐVT: %

	Phương pháp truyền thống	Phương pháp dữ liệu lớn	Phương pháp kết hợp 1	Phương pháp kết hợp 2	Chênh lệnh giữa (1) và (2)	Chênh lệnh giữa (1) và (3)	Chênh lệnh giữa (1) và (4)
A	1	2	3	4	5	6	7
Tháng 10/2017	100,63	100,66	100,65	100,65	-0,03	-0,02	-0,02
Tháng 11/2017	100,17	100,22	100,21	100,21	-0,05	-0,04	-0,04
Tháng 12/2017	100,05	100,12	100,11	100,10	-0,07	-0,06	-0,05
Tháng 01/2018	100,19	100,49	100,46	100,42	-0,30	-0,27	-0,23
Tháng 02/2018	100,34	100,37	100,36	100,36	-0,03	-0,02	-0,02
Tháng 03/2018	99,70	99,96	99,94	99,92	-0,26	-0,24	-0,22
Tháng 04/2018	100,12	100,01	100,02	100,04	0,11	0,10	0,08
Tháng 05/2018	100,43	100,22	100,27	100,32	0,21	0,16	0,11
Tháng 06/2018	100,55	100,33	100,34	100,36	0,22	0,21	0,19
Tháng 07/2018	99,91	99,89	99,90	99,92	0,02	0,01	-0,01
Tháng 08/2018	100,48	100,20	100,21	100,22	0,28	0,27	0,26
Tháng 09/2018	100,81	100,66	100,68	100,71	0,15	0,13	0,10

	Phương pháp truyền thống	Phương pháp dữ liệu lớn	Phương pháp kết hợp 1	Phương pháp kết hợp 2	Chênh lệnh giữa (1) và (2)	Chênh lệnh giữa (1) và (3)	Chênh lệnh giữa (1) và (4)
Tháng 10/2018	100,64	100,44	100,46	100,49	0,20	0,18	0,15
Tháng 11/2018	99,59	99,78	99,76	99,73	-0,19	-0,17	-0,14
Tháng 12/2018	100,75	99,76	99,88	100,01	0,99	0,87	0,74

Nguồn: tác giả tổng hợp

Kết quả tính toán CPI từ dữ liệu lớn ở tất cả các phương pháp đều thể hiện đúng xu hướng và không có chênh lệnh nhiều so với CPI truyền thống. Trong giai đoạn 15 tháng tính toán CPI so tháng trước thì có 7 tháng CPI tính theo phương pháp truyền thống cao hơn CPI tính từ dữ liệu lớn và có 8 tháng thấp hơn. Đặc biệt có 3 tháng, mức chênh lệch này chỉ có 0,02% là các tháng: tháng 10/2017, tháng 2/2018 và tháng 7/2018. Tháng có mức chênh lệch cao nhất là tháng 12/2018, CPI tính theo phương pháp truyền thống cao hơn CPI tính từ dữ liệu lớn 0,75%.

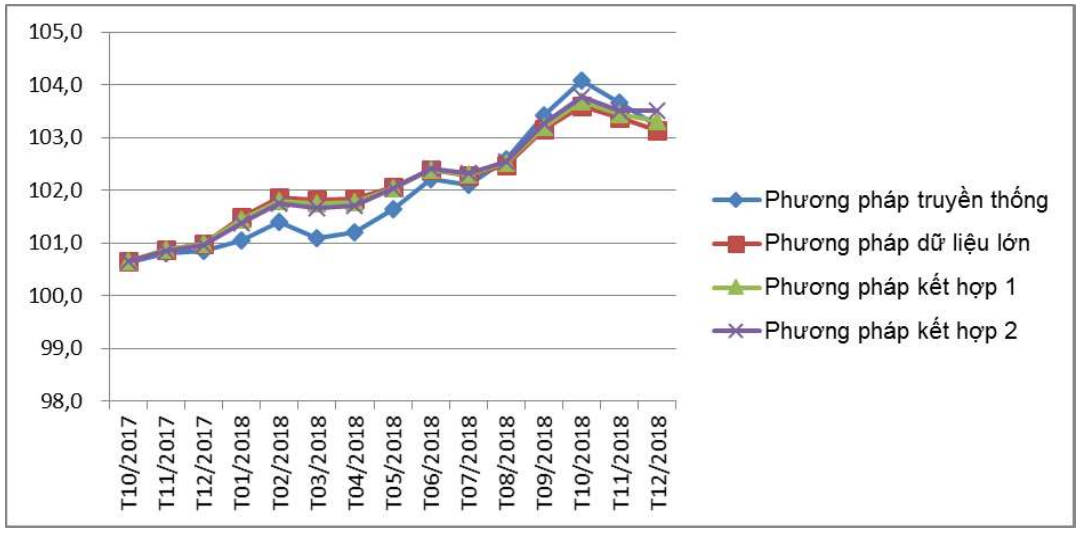


Hình 1. Chỉ số giá tiêu dùng so tháng trước

Nguồn: tác giả tổng hợp

So với CPI truyền thống, CPI được tính theo phương pháp kết hợp 2 giữa truyền thống và dữ liệu lớn có mức độ chênh lệch thấp nhất trong các phương pháp tính và là phương pháp kế thừa, kết hợp giữa phương pháp thu

thập theo dữ liệu lớn và CPI tính theo phương pháp truyền thống. Phân tích số liệu trong giai đoạn từ tháng 10/2017 đến tháng 12/2018 cho thấy, hai phương pháp thu thập dữ liệu (truyền thống và dữ liệu lớn) cho kết quả tương đối giống nhau: Chỉ số giá tiêu dùng của 2 phương pháp này tăng lần lượt là 3,23% và 3,51% (chênh lệch không đáng kể 0,28% trong giai đoạn 15 tháng).



Hình 2. Chỉ số giá tiêu dùng truyền thống và chỉ số giá được tính từ dữ liệu lớn so tháng 9 năm 2017

Nguồn: tác giả tổng hợp

Do số lượng các mặt hàng thu thập giá lớn hơn rất nhiều so với CPI tính theo phương pháp truyền thống nên CPI khai thác từ dữ liệu lớn thể hiện **biến động giá của thị trường nhạy hơn** so với phương pháp truyền thống. Kết quả tính toán cũng thể hiện rất rõ xu hướng này. Nếu chỉ tính đến 11 ngành cấp 1, từ tháng 10/2017 đến tháng 12/2018, CPI so tháng trước tính theo phương pháp truyền thống tháng nào cũng có nhóm ngành cấp 1 bằng tháng trước, đặc biệt là tháng 5/2018 có 8/11 nhóm hàng bằng giá tháng trước. Số liệu cụ thể như sau: tháng 10/2017: 2 nhóm, tháng 11/2017: 4 nhóm, tháng 12/2017: 3 nhóm, tháng 1/2018: 1 nhóm, tháng 3/2018: 5 nhóm, tháng 4/2018: 1 nhóm, tháng 5/2018: 8 nhóm, tháng 6/2018: 7 nhóm, tháng 7/2018: 3 nhóm, tháng 8/2018: 5 nhóm, tháng 9/2018: 6 nhóm, tháng 10/2018: 1 nhóm, tháng 11/2018: 6 nhóm. Trong khi đó CPI so tháng trước tính theo phương pháp thu thập từ dữ liệu lớn luôn có sự biến động về giá qua từng tháng.

Phương pháp thu thập từ dữ liệu lớn **ổn định hơn** so với phương pháp truyền thống. Phương pháp truyền thống thu thập mẫu nhỏ: Theo phương án điều tra của Tổng cục Thống kê, trong tháng báo cáo chỉ có 503 mặt hàng lấy giá 1 lần, 144 mặt hàng lấy giá 3 lần, có 5 mặt hàng lấy giá khi có phát sinh và có 2 mặt hàng lấy giá hàng ngày tại 1 điểm điều tra trong mỗi khu vực điều tra. Trong khi đó phương pháp thu thập giá từ dữ liệu lớn thu thập khoảng 250 ngàn mặt hàng vào 3 kỳ / tháng. Như vậy nếu như một mặt hàng trong rổ hàng hóa của CPI truyền thống (đặc biệt là mặt hàng chỉ thu thập 1 kỳ / 1 tháng) biến động mạnh (tăng cao hoặc giảm sâu) thì sẽ ảnh hưởng rất lớn đến CPI chung. Trong khi đó, đối với dữ liệu lớn sẽ không ảnh hưởng nhiều vì có rất nhiều hàng hóa khác cùng nhóm được thu thập nhưng không có sự biến động lớn.

Phương pháp thu thập từ dữ liệu lớn có tính **đại diện mẫu cao hơn** nhiều so với phương pháp truyền thống vì phương pháp thu thập từ dữ liệu lớn thu thập giá của tất cả các mặt hàng, do đó có thể xem mẫu gần như là đại diện cho toàn bộ các hàng hóa có trên thị trường. Ngược lại, đối với CPI truyền thống: do rổ hàng hóa được cập nhật theo chu kỳ 5 năm nên có rất nhiều mặt hàng tiêu dùng phổ biến xuất hiện trên thị trường nhưng chưa được cập nhật, hơn nữa rổ hàng hóa cũng chỉ có 654 mặt hàng đại diện tiêu dùng phổ biến nên cũng thiếu rất nhiều so với thực tế biến động liên tục của thị trường.

Tuy nhiên, trong tình hình hiện nay, việc thu thập dữ liệu lớn để phục vụ công tác tính toán chỉ số giá tiêu dùng cũng gặp một số hạn chế như:

- Không phải tất cả các sản phẩm trong rổ hàng hóa và dịch vụ đều được thu thập bằng nguồn dữ liệu lớn. Trong thực tế tính toán, để phù hợp với quyền số và phạm vi so sánh với CPI truyền thống, đề tài phải lấy một số hàng hóa và dịch vụ từ rổ hàng hóa truyền thống để xử lý.

- Phần lớn các trang web trực tuyến hoạt động và cung cấp hàng hóa cho khắp các địa phương trên cả nước, do đó rất khó phân tách để tính doanh thu chia theo địa phương, làm cơ sở tính trọng số phục vụ việc tính CPI cho từng địa phương. Đây cũng là một vấn đề khó cần phải xử lý khi tính CPI bằng phương pháp truyền thống bởi vì hiện nay có nhiều tập đoàn, nhà bán lẻ có tầm cỡ quốc gia có nhiều cửa hàng bán lẻ khắp cả nước và bán với cùng một mức giá.

- Hiện nay, hoạt động thương mại điện tử chỉ phát triển mạnh và tập trung ở khu đô thị, các thành phố lớn, do vậy việc tính toán chỉ số giá tiêu dùng sẽ bị giới hạn bởi phạm vi địa lý, đặc biệt là ở các vùng nông thôn.

3. THẢO LUẬN VÀ GỢI Ý CHÍNH SÁCH

Dữ liệu lớn sẽ là một xu hướng tất yếu trong tương lai và có tiềm năng tăng trưởng rất lớn trong nền kinh tế toàn cầu. Ở Việt Nam, đặc biệt là ngành thống kê, nếu triển khai được công tác thu thập thông tin dựa trên dữ liệu lớn sẽ đem lại nhiều lợi ích như: cắt giảm chi phí, giảm thời gian thu thập thông tin, tăng chất lượng và tối ưu hóa số liệu thống kê. Tuy nhiên để triển khai được hiệu quả, chúng ta phải đối diện với nhiều thách thức:

3.1. Về tính pháp lý

- Thách thức lớn nhất là vấn đề đảm bảo tính pháp lý đối với thông tin thu thập được từ dữ liệu lớn (nguồn dữ liệu trích xuất từ các trang Web). Trong Luật Thống kê (Chương III: Thu thập thông tin thống kê Nhà nước) chỉ mới quy định các nguồn dữ liệu: Điều tra thống kê, dữ liệu hành chính và chế độ báo cáo thống kê là ba nguồn dữ liệu chính thức phục vụ sản xuất thông tin thống kê trong hệ thống thống kê nhà nước. Vì vậy, để các nguồn dữ liệu khác, đặc biệt là nguồn dữ liệu lớn trở thành nguồn dữ liệu chính thống, được áp dụng chính thức vào quá trình sản xuất thông tin thống kê trong hệ thống thống kê nhà nước cần phải có sự nỗ lực rất lớn từ các cơ quan chức năng có liên quan mà đơn vị giữ vị trí trọng yếu là Tổng cục Thống kê và đặc biệt là vai trò trung tâm của Bộ Kế hoạch và Đầu tư, cơ quan quản lý nhà nước về thống kê để chuẩn bị điều kiện cần thiết như cải thiện cơ sở hạ tầng, hoàn thiện khuôn khổ pháp lý liên quan đến dữ liệu lớn. Công việc cụ thể, trước mắt là cố gắng thực hiện tốt Đề án ứng dụng công nghệ thông tin - truyền thông trong Hệ thống thống kê nhà nước giai đoạn 2017- 2025, tầm nhìn đến năm 2030, đây cũng là nhiệm vụ trọng tâm để góp phần thúc đẩy sự phát triển kinh tế số ở Việt Nam.

- Một nội dung nửa về vấn đề tính pháp lý, đó là việc nếu ngành thống kê tiến hành triển khai chính thức việc thu thập thông tin giá trên các trang Web thì các quy định, các thỏa thuận về những điều khoản dịch vụ tại các trang web sẽ có các nội dung gây khó khăn cho các cơ quan thống kê trong quá trình thu thập và công bố số liệu. Việc thu thập thông tin trên các trang Web do các cơ quan thống kê tiến hành phải có được sự cho phép của đơn vị chủ quản trang Web. Để đảm bảo tất cả dữ liệu thu thập được sử dụng cho mục đích quản lý nhà nước hoặc phục vụ trong nghiên cứu, khi

thu thập dữ liệu trực tuyến cho dù là thủ công hay tự động, cơ quan thống kê phải cung cấp cho các đơn vị chủ quản trang Web các cam kết bảo mật và sẽ chỉ sử dụng thông tin cho mục đích thống kê.

Hòa nhập vào xu hướng chung của thống kê thế giới, việc chuyển đổi sang các phương pháp thống kê tiên tiến là rất cần thiết nhưng cũng cần có đồng thời những quy định pháp lý cụ thể về việc tiếp cận các nguồn dữ liệu lớn để tạo điều kiện cho việc phát triển kênh thông tin đầu vào này dần thay thế kênh điều tra, nâng cao hiệu quả trong việc sử dụng các nguồn lực xã hội. Do đó, cần thiết phải tiến hành rà soát, cập nhật và bổ sung các văn bản quy phạm pháp luật, đặc biệt là Luật Thống kê và các Nghị định hướng dẫn để tạo cơ sở pháp lý cho việc khai thác, sử dụng nguồn dữ liệu lớn trong công tác thống kê nhà nước. Ngoài ra, nhà nước cũng sớm triển khai xây dựng cơ sở pháp lý đối với các doanh nghiệp quản lý, vận hành các trang web thương mại điện tử trong việc cung cấp thông tin thống kê. Cụ thể:

(1) Nghị định 97/2016/NĐ-CP ngày 1 tháng 7 năm 2016 của Chính phủ quy định nội dung chỉ tiêu thống kê thuộc hệ thống chỉ tiêu thống kê quốc gia chỉ rõ nguồn số liệu để tính chỉ số giá tiêu dùng bao gồm 2 nguồn: Điều tra giá tiêu dùng và Khảo sát mức sống dân cư Việt Nam. Đề nghị bổ sung thêm nguồn thông tin "Khai thác từ dữ liệu lớn và các nguồn thông tin khác" phục vụ biên soạn CPI ở Việt Nam trong quá trình sửa đổi Luật Thống kê.

(2) Bổ sung Luật Thống kê tại Điều 33, mở rộng thêm đối tượng cung cấp thông tin thống kê (ngoài các "tổ chức, cá nhân được điều tra thống kê" nên bổ sung "các tổ chức, cá nhân chia sẻ dữ liệu thống kê").

3.2. Về nguồn nhân lực

Khai thác dữ liệu lớn phục vụ công tác thống kê là một nội dung mới, vì vậy nguồn nhân lực trong lĩnh vực này còn thiếu về số lượng và cũng có những hạn chế nhất định về chất lượng. Công việc vận hành hệ thống, phát triển phần mềm, khai phá dữ liệu… là các kỹ năng cần thiết để thực hiện tốt công việc này. Việc khai thác dữ liệu lớn khác hoàn toàn với việc khai thác dữ liệu truyền thống có dung lượng dữ liệu nhỏ trước đây. Ở Việt Nam, phần lớn mọi người vẫn còn thói quen thao tác trên những dạng dữ liệu có cấu trúc và sử dụng SQL để khai thác các cơ sở dữ liệu quan hệ. Chính vì vậy việc phát triển nguồn nhân lực đáp ứng cho công tác thống kê trong thời đại kinh tế số, cụ thể là nguồn nhân lực vận hành và khai thác nguồn dữ liệu là một vấn đề lớn đối với ngành thống kê.

Một trong những nguồn lực quan trọng tác động lớn đến quá trình phát triển kinh tế của quốc gia là nguồn nhân lực, đây là cơ sở quyết định để tạo ra lợi thế quốc gia. Lịch sử thế giới đã chứng minh, trong quá trình phát triển kinh tế - xã hội của quốc gia, nếu các nước tận dụng và phát triển tốt nguồn nhân lực sẽ chiếm được nhiều lợi thế. Ngày nay, cùng với cuộc cách mạng công nghiệp 4.0 và sự phát triển mạnh mẽ của nền kinh tế số, thế giới nói chung và Việt Nam nói riêng đang trong quá trình chuyển đổi nền kinh tế phát triển dựa vào tài nguyên sang quá trình tăng trưởng dựa vào kinh tế tri thức. Do đó vấn đề phát triển nguồn nhân lực lại càng phải được đặt lên hàng đầu. Một trong những phương pháp tối ưu để nâng cao kiến thức, phát triển chất lượng nguồn nhân lực là tăng cường quá trình học tập.

Để phát triển nguồn nhân lực chất lượng cao, đặc biệt nguồn nhân lực trong lĩnh vực dữ liệu lớn, cần phải tăng cường vai trò của Nhà nước. Nhà nước xây dựng cơ chế phối hợp, tạo điều kiện thuận lợi để tạo mối quan hệ mật thiết giữa các đơn vị sử dụng lao động với nhà trường để có thể đào tạo nguồn nhân lực chất lượng, phù hợp với nhu cầu của doanh nghiệp và đơn vị sử dụng lao động. Ngoài ra, các cơ sở giáo dục cần nâng cao chất lượng công tác đào tạo, chương trình đào tạo phải bám sát thực tiễn, đáp ứng được nhu cầu công việc. Trên cơ sở đó giúp các học viên có thể xử lý được các bài toán thực tiễn về dữ liệu lớn khi được tuyển dụng. Ngoài hai đơn vị đào tạo trong ngành thống kê là Trường Cao đẳng Thống kê Bắc Ninh và Trường Cao đẳng Thống kê II Đồng Nai, ngành thống kê nên có kế hoạch hợp tác toàn diện với các cơ sở đào tạo chuyên ngành thống kê lớn trong nước như Trường Đại học Kinh tế Thành phố Hồ Chí Minh, trường Đại học Kinh tế quốc dân, Đại học Kinh tế Đà Nẵng… trong các lĩnh vực như đào tạo sau đại học, tổ chức các khóa đào tạo các chuyên ngành sâu về thống kê, khoa học dữ liệu, đặc biệt là dữ liệu lớn, nghiên cứu khoa học, hướng dẫn sinh viên thực tập… Qua đó chất lượng đào tạo sẽ dần được cải thiện và sinh viên ra trường sẽ đáp ứng tốt hơn nhu cầu của đơn vị sử dụng lao động.

Với sự phát triển của nền kinh tế số và sự phát triển của quá trình chuyển đổi số, cần phải đổi mới về chương trình đào tạo, điều chỉnh về cơ cấu đào tạo ngành nghề, trong đó tập trung vào công tác đào tạo nguồn nhân lực về công nghệ thông tin. Đối với chương trình đào tạo cho chuyên ngành thống kê, cần phải bổ sung kiến thức các môn học về khoa học dữ liệu, khai phá dữ liệu, đào tạo các kỹ năng như: khai thác dữ liệu lớn, xử lý dữ liệu lớn, phân tích dữ liệu lớn, kỹ thuật học máy... Đối với chương trình

đào tạo về công nghệ thông tin, do xu hướng công nghệ phát triển ngày càng nhanh, cần phải thường xuyên cập nhật tài liệu kỹ thuật, cập nhật giáo trình công nghệ thông tin liên quan đến xu hướng cách mạng công nghiệp 4.0, xu hướng chuyển đổi số, kinh tế số, các công nghệ mới như Dữ liệu lớn, Trí tuệ nhân tạo (Artificial Intelligence - AI), Kỹ thuật học máy (Machine Learning)…

Các trường đại học trong nước cố gắng phát huy tối đa nội lực của mình trong các hoạt động như tăng cường công tác nghiên cứu khoa học; hợp tác nghiên cứu với các doanh nghiệp trong lĩnh vực kinh tế số, chuyển đổi số, dữ liệu lớn…; cải tiến chương trình đào tạo; cập nhật giáo trình, các công nghệ mới, tiên tiến…. Bên cạnh đó tích cực hợp tác với các viện nghiên cứu lớn trên thế giới về dữ liệu lớn, kinh tế số, nền tảng công nghệ nhằm hướng đến phát triển đại học thông minh và các trung tâm nghiên cứu về khoa học dữ liệu, dữ liệu lớn hàng đầu.

Về công tác tổ chức đào tạo đối với ngành thống kê: Tập trung đào tạo, nâng cao năng lực, kỹ năng, nâng cao trình độ ứng dụng công nghệ thông tin, các kiến thức, kỹ năng có liên quan đến dữ liệu lớn cho cán bộ làm công tác nghiệp vụ cũng như cho người làm công nghệ thông tin trong toàn ngành. Nghiên cứu ứng dụng khoa học dữ liệu và các công nghệ tiên tiến hiện đại, áp dụng phù hợp cho công tác thống kê tại Việt Nam.

Đổi mới tư duy từ việc cho rằng "chỉ cần học một lần để làm việc suốt đời" sang khuynh hướng "học tập suốt đời mới đủ khả năng làm việc suốt đời". Theo các chuyên gia, trong kỷ nguyên số 4.0, công nghệ thay đổi rất nhanh, trường học khó có thể cập nhật kiến thức mới cho học viên. Vì vậy, học viên phải tự học để nâng cao chất lượng nguồn nhân lực. Một trong những nguồn để nâng cao kiến thức là Internet, Internet sẽ giúp sinh viên nắm được kiến thức nhanh trong khoảng thời gian ngắn.

Với những thành tựu của Đề án “Xây dựng xã hội học tập giai đoạn 2012-2020”, tiếp tục xây dựng và ban hành đề án mới để tiếp tục hình thành và phát triển xã hội học tập giai đoạn 2021-2030, cố gắng hình thành mô hình công dân học tập có những năng lực cần thiết để xây dựng nền kinh tế tri thức, đáp ứng yêu cầu nhân lực chất lượng cao của nền sản xuất kỹ thuật số và hội nhập quốc tế. Xã hội học tập sắp tới phải hướng tới mục tiêu tạo ra các công dân số với đầy đủ năng lực, kỹ năng số, phục vụ việc phát triển bản thân, cộng đồng và đất nước.

3.3. Về kỹ thuật học máy trong xử lý dữ liệu lớn

Khi đề cập đến dữ liệu lớn, một khái niệm cũng thường hay được nhắc đến là kỹ thuật học máy (Machine learning - ML). Kỹ thuật học máy là một trong những nội dung rất quan trọng của trí tuệ nhân tạo được nghiên cứu để giải quyết các vấn đề trong việc xử lý dữ liệu lớn. Học máy là việc dạy cho máy tính làm được những gì mà một cách tự nhiên con người có thể làm được, chủ yếu đó là việc học hỏi từ kinh nghiệm. Dữ liệu lớn và học máy đặc biệt phát triển và gặt hái được nhiều thành công trong thời gian gần đây và điều này là cơ sở và động lực quan trọng để phát triển các hoạt động thống kê, giúp cho vị thế của ngành thống kê ngày càng được nâng cao trong xã hội hiện đại, nhất là trong kỹ nguyên số hiện nay. Đề tài cũng đã sử dụng kỹ thuật học máy trong việc phân tích, sàng lọc thông tin để đưa các dữ liệu từ phi cấu trúc thành có cấu trúc như tách được đặc điểm hàng hóa, đơn vị tính, giá hàng hóa,… Tuy nhiên, một trong những công đoạn khó khăn và tốn thời gian nhất của đề tài là mã hóa các mặt hàng theo danh mục hàng hóa và dịch vụ đại diện (mã hàng hóa cấp 5) của chỉ số giá tiêu dùng, hiện nay công đoạn này phải làm thủ công. Do đó nếu nghiên cứu và ứng dụng được thuật toán học máy trong công tác mã hóa dữ liệu thì khối lượng công việc sẽ giảm đi nhiều. Phân ngành kinh tế và mã hóa dữ liệu là công việc thường xuyên, cần thiết và rất quan trọng trong công tác thống kê để đảm bảo tất cả dữ liệu thu thập có thể so sánh được với nhau, thông thường quy trình này đều được thực hiện thủ công và huy động rất nhiều nguồn nhân lực, các chuyên viên xử lý sẽ đọc thông tin của người trả lời phiếu điều tra hoặc đọc thêm thông tin có được từ các nguồn dữ liệu khác và gán mã thích hợp. Công việc này đòi hỏi nhiều thời gian và kinh nghiệm xử lý. Tuy nhiên, với kỹ thuật học máy phát triển như hiện nay, chúng ta có thể tự động hóa toàn bộ quá trình này. Đầu tiên, chúng ta có thể chọn một mẫu nhỏ trong tập dữ liệu để các chuyên gia mã hóa, sau đó tiến hành các thuật toán giúp cho máy học hỏi công việc, kinh nghiệm những từ mô hình đã được mã hóa của các chuyên gia, cuối cùng sử dụng kỹ thuật học máy này để phân loại hay mã hóa phần dữ liệu còn lại bằng những công việc đã học hỏi từ việc mã hóa của chuyên gia. Trong quá trình triển khai, để đảm bảo độ chính xác, độ tin cậy của dữ liệu đã được mã hóa, có thể tiến hành lặp lại nhiều lần thao tác: chọn một mẫu nhỏ, cho máy học, mở rộng mẫu cho máy mã hóa, kiểm tra mẫu, nếu tỷ lệ đạt yêu cầu thì triển khai làm toàn bộ, nếu tỷ lệ chưa đạt yêu cầu thì tiến hành mở rộng mẫu để máy học thêm… Việc mã hóa tự động như quy trình ở trên sẽ giúp cho việc tổng hợp và công bố số liệu sớm hơn và như vậy số liệu công

bố sẽ có giá trị hơn đối với người dùng tin. Hiện nay, trên thế giới cũng đã và đang nghiên cứu, triển khai nhiều dự án về học máy để mở đường cho sản xuất thống kê hiệu quả và hiện đại. Ủy ban Kinh tế và Xã hội của Liên hợp quốc ở châu Âu với chương trình hợp tác quốc tế nhằm tạo điều kiện, giúp đỡ các tổ chức thống kê trên toàn thế giới tiến tới việc sản xuất các số liệu thống kê quan trọng theo những cách sáng tạo dựa trên học máy và trí tuệ nhân tạo. Dựa trên các câu trả lời cho các câu hỏi điều tra mở, các nước như Canada, Mexico, Serbia và Iceland đang thử nghiệm học máy để phân loại công việc mà mọi người nắm giữ và ngành họ làm việc. Anh là quốc gia hiện đang đứng đầu về việc sử dụng học máy có đạo đức trong thống kê; IMF, Mexico, Thụy Điển và những nước khác thì tập trung vào việc tích hợp học máy vào sản xuất số liệu thống kê; Phần Lan sẽ khám phá các vấn đề về chất lượng trong bộ dữ liệu sử dụng các thuật toán học máy; Mexico sẽ tiếp tục nỗ lực để thiết lập một khuôn khổ chất lượng được quốc tế thống nhất cho học máy trong số liệu thống kê chính thức (UNECE, 2021). Nguyên tắc cơ bản của thống kê nhà nước là các số liệu thống kê phải được tạo ra trên cơ sở khoa học, minh bạch và đáng tin cậy. Do đó, điều quan trọng là phải chứng minh các công nghệ và kỹ thuật mới được khai thác một cách khoa học, chính xác để có thể duy trì được lòng tin của người sử dụng thông tin. Để có thể ứng dụng được kỹ thuật học máy vào công tác thống kê, ngành thống kê phải triển khai và phát triển khung chất lượng cho các thuật toán thống kê, nhằm hướng dẫn việc triển khai và để đảm bảo chất lượng khi sử dụng kỹ thuật học máy.

3.4. Về cơ sở hạ tầng công nghệ thông tin

Việc khai thác, phân tích dữ liệu lớn đòi hỏi phải có cơ sở hạ tầng công nghệ thông tin phát triển mạnh và các công nghệ nổi trội. Muốn phát triển lĩnh vực này cần phải tập trung đầu tư phát triển về hạ tầng tính toán. Mặc dù trong những năm gần đây hạ tầng công nghệ thông tin của ngành thống kê cũng đã được cải thiện rất nhiều, Tổng cục Thống kê đã quan tâm nâng cấp hạ tầng công nghệ thông tin và các giải pháp bảo đảm an toàn, an ninh hệ thống như tiến hành thuê hạ tầng công nghệ thông tin phục vụ hệ thống thu thập thông tin bằng phiếu điện tử, nâng cao năng lực Trung tâm máy chủ... nhưng nếu triển khai hệ thống thu thập thông tin trực tuyến trên các trang web và xử lý thông tin theo thời gian thực cũng còn một số hạn chế nhất định. Hiện nay dữ liệu trong lĩnh vực thống kê nhà nước đã được lưu trữ là vô cùng lớn nhưng chúng chưa được khai thác một cách hiệu

quả, đúng nghĩa để mang lại giá trị phục vụ tốt cho công tác quản lý điều hành, xây dựng các chính sách phát triển kinh tế xã hội.

Ngoài ra, một vấn đề quan trọng cần phải cân nhắc trong việc đầu tư phát triển hạ tầng công nghệ thông tin là tốc độ phát triển công nghệ trong thời đại ngày nay rất nhanh dẫn đến nguy cơ rất dễ bị lạc hậu về công nghệ nếu chúng ta không xây dựng kiến trúc tổng thể để hệ thống có độ mở, sẵn sàng tích hợp và nâng cấp mở rộng hệ thống khi cần thiết. Vì vậy việc đẩy mạnh nghiên cứu xây dựng cơ sở hạ tầng và triển khai việc nghiên cứu, sử dụng khai thác dữ liệu lớn trong thống kê nhà nước cần phải sớm được triển khai.

3.5. Về kinh phí:

Các phương pháp thu thập dữ liệu truyền thống hiện nay đều phải triển khai thu thập thông tin tại địa bàn, nếu chúng ta khai thác được dữ liệu lớn phục vụ cho công tác thống kê sẽ tiết kiệm được nhiều kinh phí. Tuy nhiên, có một khó khăn rất lớn trong quá trình triển khai là việc phải đầu tư một nguồn kinh phí ban đầu và kinh phí cho quá trình vận hành và quản lý dữ liệu. Hiện nay, hầu hết các hệ thống công nghệ thông tin hiện đại thường có chi phí cao, bên cạnh đó còn có các chi phí phát sinh kèm theo như: chi phí vận hành, duy trì hệ thống đặc biệt là chi phí về bản quyền phần mềm, thiết bị an toàn, an ninh mạng... Do đó khi triển khai việc khai thác thông tin từ dữ liệu lớn đòi hỏi cần phải xem xét tính hiệu quả trong việc đầu tư với bối cảnh nguồn kinh phí từ ngân sách ngày càng bị cắt giảm.

Để hệ thống vận hành tốt và ổn định, không phải chúng ta chỉ tập trung đầu tư về hạ tầng công nghệ thông tin để đáp ứng yêu cầu khai thác và sử dụng dữ liệu lớn, mà còn phải giải quyết bài toán đầu tư các chi phí liên quan đến cơ sở hạ tầng thống kê nhằm hình thành, phát triển và kiểm soát được hệ thống nhằm đáp ứng được yêu cầu, đảm bảo các kết quả đầu ra kịp thời hơn, chất lượng cao hơn và đặc biệt là dữ liệu an toàn hơn. Theo Nguyễn Văn Thụy (2017, trang 34), "dự kiến, trong giai đoạn 2016-2020, Chính phủ Úc sẽ đầu tư khoảng 250 triệu USD để chuyển đổi cơ sở hạ tầng, hệ thống và quy trình sản xuất số liệu thống kê". Đây là một khoản kinh đầu tư rất lớn, đặc biệt đối với điều kiện của Việt Nam hiện nay.

Ngoài ra việc truyền dữ liệu lớn thường phải gánh chịu chi phí cao, đây là thách thức cơ bản cần phải được tính đến khi nghiên cứu ứng dụng dữ liệu lớn bởi vì việc truyền dữ liệu là không thể tránh khỏi trong các ứng

dụng dữ liệu lớn. Do đó việc phân tích và tìm ra giải pháp góp phần nâng cao tính hiệu quả của việc truyền dữ liệu lớn là một yếu tố quan trọng trong việc khai thác và xử lý dữ liệu lớn.

Một nội dung cuối cùng về kinh phí là cần phải đảm bảo rằng việc nghiên cứu và triển khai sử dụng các nguồn dữ liệu khác thay thế như dữ liệu lớn sẽ góp phần làm giảm ngân sách chung, có nghĩa là khi triển khai phương án thu thập thông tin từ dữ liệu lớn chỉ được thực hiện khi chứng minh được phương án này thực sự tiết kiệm chi phí tổng thể khi xem xét trong khoảng thời gian nhất định.

3.6. Về bổ sung giá của các điểm bán hàng bình ổn trong Chương trình bình ổn của thành phố Hồ Chí Minh:

Trong những năm qua, với vai trò là đô thị đặc biệt, một trung tâm lớn về kinh tế, văn hóa, giáo dục đào tạo, khoa học công nghệ, đầu mối giao lưu và hội nhập quốc tế, Thành phố Hồ Chí Minh đã không ngừng sáng tạo, triển khai nhiều chính sách, giải pháp để phát triển đồng bộ các lĩnh vực, góp phần tích cực cùng cả nước kiểm soát lạm phát, ổn định kinh tế vĩ mô, đảm bảo an sinh xã hội. Một trong những chương trình rất thành công đó là Chương trình Bình ổn thị trường, được Thành phố triển khai liên tục từ năm 2002 tới nay đã khẳng định được hiệu quả lan tỏa, được nhân dân hưởng ứng tích cực, được Chính phủ đánh giá cao và chỉ đạo nhân rộng ra cả nước. Thành phố hiện đang triển khai chương trình bình ổn ở 4 nhóm hàng: Các mặt hàng lương thực, thực phẩm thiết yếu; các mặt hàng phục vụ Mùa khai trường; các mặt hàng Sữa, các mặt hàng Dược phẩm thiết yếu với rất nhiều doanh nghiệp lớn tham gia như Vissan, Ba Huân, Foodcosa, Thành Thành Công, CJ Cần Tre, Vinamilk, Fahasa, Saigon Co.op, Satra, BigC… Về hệ thống phân phối truyền thống, Thành phố có 239 chợ, gồm 3 chợ đầu mối, 14 chợ loại 1, 52 chợ loại 2, 170 chợ loại 3 với tỷ trọng chiếm khoảng 65-70% thị phần. Về hệ thống phân phối hiện đại, Thành phố có 206 siêu thị, 49 trung tâm thương mại và 2.566 cửa hàng tiện lợi (Sở Công Thương Thành phố Hồ Chí Minh, 2019). Hiện nay, đa số người dân có xu hướng tập trung mua sắm nhiều hơn ở các siêu thị, trung tâm thương mại, đặc biệt là những nơi có sự tham gia của các điểm bán hàng bình ổn giá của thành phố. Tùy theo ngành hàng, hiện nay các mặt hàng bình ổn chiếm từ 20% đến 50% nhu cầu thị trường. Do đó để phản ánh một cách đầy đủ hơn, chính xác hơn về tình hình giá cả của Thành phố Hồ Chí Minh, cần phải có sự kết hợp giữa dữ liệu thu thập từ dữ

liệu lớn và dữ liệu giá bán của các mặt hàng bình ổn. Nhóm tác giả có các đề xuất:

- Hàng năm, dựa vào kết quả điều tra doanh nghiệp và danh sách các doanh nghiệp đăng ký tham gia chương trình bình ổn sẽ tính toán được thị phần của từng mặt hàng bình ổn và tỷ trọng này sẽ làm quyền số để tính toán giá bình quân hàng tháng.

- Hàng tháng, căn cứ vào giá của các mặt hàng được bình ổn do Sở Tài chính và Sở Công Thương cung cấp và giá bình quân thu thập được từ dữ liệu lớn sẽ tính toán được giá bình quân chung với quyền số là tỷ trọng thị phần của hàng bình ổn trong tổng mức bán lẻ mặt hàng đó.

3.7. Về bổ sung thêm các nguồn dữ liệu thay thế khác

Trong xu thế phát triển của thời đại, Thống kê Việt Nam phải đối diện với một thách thức, khó khăn lớn là nhu cầu về thông tin, số liệu phục vụ quản lý điều hành ngày càng tăng, trong khi nguồn lực cho công tác thống kê có hạn. Với yêu cầu ngày càng nhiều thông tin và đòi hỏi rất chi tiết về số liệu của người dùng tin đòi hỏi nguồn thông tin đầu vào lớn, quá trình xử lý, tổng hợp, công bố cần có nguồn lực lớn, chỉ số giá tiêu dùng cũng không nằm ngoài xu hướng này. Do đó, việc tìm thêm các nguồn dữ liệu thay thế, bổ sung cho các nguồn dữ liệu truyền thống có ý nghĩa rất quan trọng. Các nguồn dữ liệu khác để thay thế có thể sẽ đem lại cơ hội tốt giúp ngành thống kê giải quyết nhiều thách thức mà cuộc điều tra giá tiêu dùng đang gặp phải. Các nguồn dữ liệu khác để thay thế có thể giúp cho việc đo lường chính xác hơn sự thay đổi về giá, tạo điều kiện mở rộng mẫu thu thập, giá thu thập là giá thực tế giao dịch thay vì giá do đơn vị cung cấp thông tin kê khai, khi người tiêu dùng sử dụng các mặt hàng thay thế, chỉ số giá sẽ được phản ánh chính xác hơn, hạn chế được sự thay đổi chất lượng và có thể giảm chi phí trong việc thu thập thông tin. Ngoài ra, tính kịp thời cũng là một ưu điểm lớn của nguồn dữ liệu khác để thay thế, nguồn dữ liệu này sẽ giúp cho việc tổng hợp và công bố chỉ số giá tiêu dùng kịp thời gian hơn. Số lượng hàng hóa nhiều hơn so với số lượng mẫu thu thập hiện tại, chi tiết dữ liệu cũng nhiều và đa dạng hơn, dữ liệu được thu thập với tần suất nhiều hơn.

Trong điều kiện hiện nay của Việt Nam, việc thu thập dữ liệu từ các trang web bán hàng trực tuyến cũng còn thiếu nhiều loại hàng hóa và dịch vụ. Với những phân tích ở trên, ngoài nguồn dữ liệu khai thác từ các trang web, nhóm tác giả đề nghị bổ sung thêm một nguồn dữ liệu thay thế khác

để khắc phục những hạn chế của việc thu thập dữ liệu từ các trang web bán hàng trực tuyến, đó là nguồn dữ liệu từ các doanh nghiệp cung cấp. Theo phương án điều tra giá tiêu dùng, đơn vị điều tra là các sạp, quầy hàng tại các chợ, điểm bán hàng (chuyên bán lẻ), các cơ sở kinh doanh dịch vụ,… có địa điểm kinh doanh ổn định. Tuy nhiên, chúng ta có thể khai thác thông tin về giá thông qua dữ liệu do các doanh nghiệp bán lẻ hay các doanh nghiệp hoạt động trong lĩnh vực cung cấp dịch vụ, thay vì người trả lời là chủ các sạp, quầy hàng. Một trong những nguồn dữ liệu từ các doanh nghiệp cung cấp mang lại nhiều cơ hội trong việc cải thiện tính chính xác cho số liệu chỉ số giá tiêu dùng là dữ liệu máy quét. Ngày nay, phần lớn các siêu thị, trung tâm thương mại, của hàng tiện lợi đều tích hợp công nghệ quét sản phẩm khi khách hàng thanh toán tiền và như vậy đã tạo nguồn dữ liệu vô cùng lớn. Với những đặc điểm của dữ liệu máy quét, các chuyên gia cũng xem đây là một dạng của dữ liệu lớn. Các cơ quan thống kê quốc gia ở một số nước như Úc, Hà Lan, New Zealand, Thụy Điển và Thụy Sĩ đã sử dụng dữ liệu máy quét để tính toán chỉ số giá tiêu dùng (IMF, 2018). Nếu chúng ta khai thác được nguồn dữ liệu này sẽ đem lại một số lợi ích như: đảm bảo tính kịp thời, thời gian cung ứng dữ liệu đáp ứng tính kịp thời cho việc tính toán CPI, chi phí thu thập thấp, giá thu thập là giá thực tế giao dịch, chất lượng số liệu đảm bảo, do mỗi mặt hàng đều có mã nhận diện khác nhau nên đảm bảo mức độ tích hợp của các mặt hàng, đảm bảo tính đồng nhất của thông tin. Ngoài ra, các doanh nghiệp là người cung cấp thông tin cũng là người chủ sở hữu thông tin nên cũng rất thuận tiện trong việc trao đổi, nắm tình hình khi có những vấn đề phát sinh về mặt số liệu. Tuy nhiên, khi triển khai thực tế, ngành thống kê cũng sẽ phải gặp nhiều thách thức, trong đó thách thức lớn nhất là cần phải có một hệ thống máy tính có thể đáp ứng việc lưu trữ, xử lý nguồn dữ liệu lớn này nếu muốn sử dụng các thông tin để tính chỉ số giá tiêu dùng. Do mục đích xây dựng hệ thống của các doanh nghiệp là nhằm phục vụ cho báo cáo trong nội bộ của doanh nghiệp chứ không phải được xây dựng cho việc tính toán chỉ số giá tiêu dùng nên mỗi doanh nghiệp sẽ tổ chức hệ thống cơ sở dữ liệu khác nhau hoàn toàn. Vì vậy, hệ thống công nghệ thông tin ngành thống kê cần phải đáp ứng và xử lý được các tập dữ liệu có cấu trúc, định dạng, nội dung khác nhau từ các doanh nghiệp bán lẻ khác nhau.

3.8. Về việc phát triển thị trường thương mại điện tử

Một trong những yếu tố tiên quyết quyết định đến sự thành công của việc khai thác dữ liệu của các trang web bán hàng trực tuyến trong tính

toán chỉ số giá tiêu dùng là vấn đề thúc đẩy thị trường thương mại điện tử phát triển. Thị trường thương mại điện tử Việt Nam đã và đang có những bước phát triển vượt bậc, đã góp phần quan trọng trong việc phát triển lưu chuyển hàng hóa và dịch vụ, giảm giá thành sản phẩm, nâng cao hiệu quả kinh doanh của doanh nghiệp. Hiện nay, thương mại điện tử Việt Nam được đánh giá là một trong những thị trường phát triển nhanh nhất ở khu vực Đông Nam Á. Về việc mua sắm trực tuyến trong khoảng thời gian 2015-2019, số người tham gia đã tăng bình quân 10,3%/năm, năm 2019 đạt 44,8 triệu người. Trung bình một người mua sắm trực tuyến bình quân tăng 8,8%/năm (tăng từ 160 USD lên đến 225 USD). Doanh số thương mại điện tử bán lẻ đạt 10,1 tỷ USD, chiếm 4,9% tổng mức bán lẻ hàng hóa và dịch vụ tiêu dùng cả nước (Nguyễn Thị Phương Thảo, 2021). Quyết định số 645/QĐ-TTg của Thủ tướng Chính phủ ban hành ngày 15/5/2020 về Phê duyệt Kế hoạch tổng thể phát triển thương mại điện tử quốc gia giai đoạn 2021-2025, trong đó có một mục tiêu quan trọng là đưa thương mại điện tử trở thành một trong các lĩnh vực tiên phong của nền kinh tế số đến năm 2025. Để thực hiện mục tiêu trên và góp phần thúc đẩy phát triển thị trường thương mại điện tử, nhóm tác giả đề xuất một số giải pháp:

- Tiếp tục hoàn thiện cơ chế, chính sách, tạo điều kiện, hỗ trợ các hoạt động ứng dụng thương mại điện tử và các mô hình kinh doanh mới trên nền tảng công nghệ số. Cụ thể tiếp tục công tác rà soát, bổ sung, hoàn thiện các chính sách, các văn bản quy phạm pháp luật nhằm góp phần khuyến khích, hỗ trợ các hoạt động ứng dụng thương mại điện tử, nghiên cứu ban hành các văn bản hướng dẫn phù hợp với tình hình phát triển ở Việt Nam và tình hình phát triển ở các nước trên thế giới, chính sách về thương mại điện tử trong nước so với các cam kết trong Hiệp định thương mại tự do (FTA). Nâng cao năng lực quản lý và tổ chức hoạt động thương mại điện tử.

- Tăng cường công tác tuyên truyền các lợi ích của thương mại điện tử cho người dân, xây dựng lòng tin từ người tiêu dùng trong thương mại điện tử, đấu tranh chống các hành vi gian lận thương mại, xâm phạm quyền sở hữu trí tuệ và cạnh tranh không lành mạnh.

- Đào tạo kỹ năng thương mại điện tử cho doanh nghiệp và các kiến thức cơ bản về thương mại điện tử cho người dân, xây dựng thói quen, kỹ năng về thương mại điện tử.

- Cải thiện năng lực các hệ thống hạ tầng và dịch vụ hỗ trợ cho thương mại điện tử, hoàn thiện các dịch vụ chuyển phát và logistics. Xây

dựng hệ thống quản lý các dịch vụ vận chuyển, giao nhận cho thương mại điện tử bao phủ tất cả các khu vực từ thành thị đến nông thôn và các tỉnh, thành phố trên cả nước.

Với những bất cập đã trình bày ở trên, để thực hiện tốt và có hiệu quả việc khai thác dữ liệu lớn phục vụ cho công tác thống kê chính thức, nhóm tác giả đề xuất một số khuyến nghị sau:

Ngành Thống kê sớm triển khai xây dựng cơ chế, chính sách tạo điều kiện cần thiết cho việc chuyển đổi từ việc điều tra trực tiếp theo phương pháp truyền thống sang việc khai thác các nguồn dữ liệu gián tiếp, khai thác tối đa việc sử dụng dữ liệu hành chính để tiết kiệm nguồn kinh phí, tiết kiệm nhân lực, nâng cao hiệu quả hoạt động thống kê, giảm bớt gánh nặng trong việc điều tra thống kê, giảm áp lực cho các đối tượng cung cấp thông tin và cũng là góp phần tích cực trong công cuộc chuyển đổi số ngành thống kê.

Chính phủ, cụ thể là Tổng cục Thống kê cần có quyết tâm cao trong việc xem và sử dụng nguồn dữ liệu lớn đầy tiềm năng như là một kênh quan trọng để thu thập thông tin, nhất là thông tin về giá để phục vụ cho công tác thống kê nhà nước. Trước mắt, cần tập trung phát triển việc xây dựng thể chế, tạo môi trường pháp lý cho việc khai thác dữ liệu lớn phục vụ cho công tác thống kê nhà nước, đặc biệt là công tác thu thập dữ liệu trên các trang web trực tuyến.

Tăng cường và chú trọng hình thức khai thác dữ liệu trực tuyến để thu thập thông tin, xử lý và cung cấp thông tin một cách tự động cho các đối tượng sử dụng; Tập trung nghiên cứu, xây dựng các Robot dữ liệu để có thể thu thập thông tin một cách tự động theo tần suất cố định (tuần, ngày, giờ...) để đáp ứng nhu cầu ngày càng tăng cho các đối tượng người sử dụng.

Trong việc quản lý các vấn đề của đời sống kinh tế và xã hội, Chính phủ cần tích cực triển khai việc đổi mới phương thức và biện pháp áp dụng công nghệ số, từng bước chuyển đổi số, xây dựng, dần dần hoàn thiện và chuyển đổi các hoạt động quản lý nhà nước sang Chính phủ điện tử dựa vào nền tảng công nghệ số. Khi mà tất cả hệ thống dịch vụ công được nhà nước cung cấp trực tuyến, mọi công dân đương nhiên sẽ trở thành công dân điện tử, tương tự vậy mọi doanh nghiệp cũng sẽ trở thành doanh nghiệp điện tử. Điều này sẽ tạo nhiều thuận lợi cho việc triển khai công tác thu thập giá trực tuyến.

Chính phủ sớm triển khai xây dựng và hoàn thiện hạ tầng cứng và mạng lưới viễn thông, qua đó làm cơ sở để phát triển hạ tầng mềm là dịch vụ số tạo điều kiện để nâng cao hiệu quả các hoạt động kinh tế và xã hội của đất nước.

Ngành thống kê tăng cường việc hợp tác nghiên cứu với các Viện nghiên cứu, các doanh nghiệp hoạt động trong các lĩnh vực có liên quan đến việc phân tích dữ liệu lớn, các nhà khoa học dữ liệu để xây dựng giải pháp khai thác và sử dụng các thông tin từ nguồn dữ liệu lớn có hiệu quả, xây dựng được khung phân tích, đánh giá chất lượng dữ liệu thu thập, phương pháp tính và các thách thức phát sinh khi tiếp cận nghiên cứu về dữ liệu lớn.

Bố trí kinh phí hàng năm từ nguồn ngân sách nhà nước và hỗ trợ kỹ thuật của các dự án hợp tác song phương... để thực hiện công tác này. Việc đầu tư kinh phí ban đầu là rất lớn, ngoài ra còn có các chí phí khác có liên quan đến quá trình vận hành, quản trị và khai thác dữ liệu là một bài toán khó trong tình hình hiện nay. Do đó, ở Việt Nam, để triển khai được nội dung này, cần phải có sự nỗ lực rất lớn từ nhiều phía. Chi phí tiết kiệm được do khai thác dữ liệu lớn sẽ bù đắp được các khoản chi phí đầu tư ban đầu và sẽ thật sự phát huy hiệu quả trong dài hạn.

4. KẾT LUẬN VÀ KHUYẾN NGHỊ GIẢI PHÁP THỰC HIỆN

Dữ liệu lớn sẽ là một xu hướng tất yếu trong tương lai và có tiềm năng tăng trưởng rất lớn trong nền kinh tế toàn cầu. Ở Việt Nam, đặc biệt là ngành thống kê, nếu triển khai được công tác thu thập thông tin dựa trên dữ liệu lớn sẽ đem lại nhiều lợi ích như: cắt giảm chi phí, giảm thời gian thu thập thông tin, tăng chất lượng và tối ưu hóa số liệu thống kê. Đề tài nghiên cứu đã trình bày được các nội dung: Cơ sở khoa học của việc tính chỉ số giá tiêu dùng truyền thống; Việc áp dụng phương pháp tính truyền thống tại các quốc gia trên thế giới; Việc áp dụng phương pháp tính truyền thống tại Việt Nam; Cơ sở khoa học của việc tiếp cận cách tính dựa trên dữ liệu lớn; Qua đó đề xuất giải pháp khai thác dữ liệu lớn phục vụ cho việc tính toán chỉ số giá tiêu dùng. Nhóm tác giả cũng đã tiến hành thu thập 246.069 mặt hàng trên 28 trang web bán hàng trực tuyến có uy tín và đã tính toán thử nghiệm chỉ số giá tiêu dùng với kết quả thể hiện đúng xu hướng và không có chênh lệnh nhiều so với CPI truyền thống. Mặc dù còn một số hạn chế như:

Không phải tất cả các sản phẩm trong rổ hàng hóa và dịch vụ đều được thu thập bằng nguồn dữ liệu lớn;

Phần lớn các trang web trực tuyến hoạt động và cung cấp hàng hóa cho khắp các địa phương trên cả nước, do đó rất khó phân tách để tính doanh thu chia theo địa phương, làm cơ sở tính trọng số phục vụ việc tính CPI cho từng địa phương;

Hoạt động thương mại điện tử chỉ phát triển mạnh và tập trung ở khu đô thị, các thành phố lớn, do vậy việc tính toán chỉ số giá tiêu dùng sẽ bị giới hạn bởi phạm vi địa lý, đặc biệt là ở các vùng nông thôn,

Tuy nhiên nghiên cứu này cũng đã chứng minh được vai trò quan trọng của dữ liệu lớn trong công tác thống kê, đặc biệt là thống kê giá. Cụ thể, nghiên cứu có những đóng góp chính sau:

- Xây dựng phương pháp tiếp cận mới trong việc thu thập thông tin thống kê, một trong những bước quan trọng nhất trong quá trình 7 bước sản xuất thông tin thống kê;

- Xây dựng được quy trình khai thác thông tin giá từ dữ liệu lớn, cụ thể là các trang web bán hàng trực tuyến;

- Xây dựng được quy trình tính chỉ số giá tiêu dùng từ dữ liệu lớn;

- Các doanh nghiệp có thể khai thác dữ liệu từ đề tài để phân tích dữ liệu khách hàng, phân tích thị trường ở các nội dung cụ thể như:

+ Phát hiện và nghiên cứu sản phẩm mới trên thị trường;

+ Phân tích dòng đời sản phẩm;

+ Xây dựng chiến lược kinh doanh, chiến lược về giá bán sản phẩm, các chương trình khuyến mại...

+ Hỗ trợ doanh nghiệp ra các quyết định về lượng hàng tồn kho, thời điểm nhập hàng, chính sách giá bán sản phẩm...

TÀI LIỆU THAM KHẢO

1. Berry, Francien, Brian Graf, Michael Stanger, and Mari Ylä-Jarkko, 2019. Price Statistics Compilation in 196 Economies: The Relevance for Policy Analysis. *IMF Working Papers*, International Monetary Fund, 2019.
2. Cục Thống kê thành phố Hồ Chí Minh, 2015. *Kế hoạch điều tra và báo cáo thống kê giá tiêu dùng (thời kỳ 2015-2019)*.
3. Cục Thống kê thành phố Hồ Chí Minh, 2018. *Báo cáo tình hình kinh tế xã hội năm 2018*.
4. Cục Thống kê thành phố Hồ Chí Minh, 2020. *Niên giám thống kê 2019*. Thành phố Hồ Chí Minh: Nhà xuất bản Thống kê.
5. Hiệp hội Thương mại điện tử Việt Nam, 2016. *Chỉ số Thương mại Điện tử Việt Nam 2015*.
6. IMF, 2018. CPI Manual update chapter 10.
7. Luật Thống kê, 2015.
8. Nguyễn Thị Phương Thảo, 2021. Thúc đẩy phát triển thị trường thương mại điện tử. *Con số & Sự kiện*, Năm thứ 60, kỳ II - 02-2021, trang 31-33.
9. Nguyễn Văn Thụy, 2017. Khai thác dữ liệu giao dịch để biên soạn chỉ số giá tiêu dùng kinh nghiệm của cơ quan thống kê quốc gia Úc. *Thông tin khoa học thống kê*, Số 3-2017, trang 31-39.
10. Sở Công Thương, 2019. *Báo cáo tình hình triển khai Chương trình bình ổn thị trường năm 2019* - Tết Canh tý 2020 trên địa bàn Thành phố Hồ Chí Minh.
11. Tổng cục Thống kê, 2015. *Phương án Điều tra và báo cáo thống kê giá tiêu dùng (thời kỳ 2015-2019)*.
12. UNECE, 2021. Machine learning paves the way for modern, efficient statistical production.
13. Vũ Thị Thu Thủy, 2015. Hệ thống thông tin giá và các cuộc điều tra thống kê giá. *Thông tin khoa học thống kê*, số 4-2015, trang 16-19.