SỰ CẦN THIẾT TRONG VIỆC KHAI THÁC DỮ LIỆU HÀNH CHÍNH ĐỂ SẢN XUẤT THÔNG TIN THỐNG KÊ NHÀ NƯỚC CỦA VIỆT NAM

ThS. Nguyễn Thanh Ngọc
Cục Thu thập DL và Ứng dụng CNTT Thống kê - TCTK

Trên thế giới

Trên thế giới có một số nước ứng dụng khai thác, sử dụng dữ liệu hành chính (DLHC) phục vụ cho mục đích thống kê đã được thực hiện từ lâu và rất thành công.

Có thể kể đến các nước trong khối Scandinavia như: Na-Uy, Thụy Điển, Đan Mạch và Phần Lan đã sử dụng dữ liệu hành chính trong công tác thống kê từ rất sớm và khá thành công. Cụ thể: Ở Đan Mạch, dữ liệu hành chính được xây dựng căn cứ vào sổ đăng ký chính của họ - Sổ Đăng ký Cá nhân Trung tâm (CPR), Sổ Đăng ký kinh doanh (CVR) và Sổ Đăng ký Nhà ở và Hộ Gia đình (BBR). Ở Phần Lan, việc khai thác sử dụng dữ liệu hành chính cho mục đích sản xuất số liệu thống kê kinh tế - xã hội được xây dựng ngay từ trong luật và thực hiện rất tốt. Thống kê Phần Lan (cụ thể là văn phòng thống kê trung ương) thu thập gần như tất cả (chiếm khoảng 93%) dữ liệu từ các nguồn hành chính. Từ năm 1990, Phần Lan cũng đã hoàn toàn dựa vào dữ liệu hành chính để thực hiện điều tra dân số và nhà ở. Ngay trong các văn bản quy phạm pháp luật của mình, Phần Lan cũng dựa trên nguyên tắc cốt lõi là khai thác và tận dụng tối đa nguồn dữ liệu hành chính để đưa ra các quyết sách.

Bên cạnh các nước trong khối Scandinavia, hiện nay trên thế giới việc nghiên cứu ứng dụng khai thác dữ liệu hành chính trong sản xuất số liệu thống kê đã và đang dần được chú trọng, có nhiều nghiên cứu và đã được áp dụng thành công ở một số nước Châu Á (Sinh-ga-po, Hàn Quốc) cũng như một số nước Châu Âu (Đan Mạch, Na-uy, Đức, Anh…) và Châu Úc (Australia) hay như ở các nước Bắc Mỹ (Ca-na-đa).

Tại Ca-na-đa, từ năm 1921 cơ quan Thống kê của nước này đã thu thập các dữ liệu thống kê quan trọng từ các tỉnh cũng như các vùng lãnh thổ của họ. Ngoài ra Ca-na-đa đã nhập và xuất dữ liệu về các doanh nghiệp trên lãnh thổ của mình từ những 1938. Có thể nói, tính đến nay Ca-na-đa đã sử dụng dữ liệu hành chính trong gần một thế kỷ. Hiện tại, cơ quan thống kê Ca-na-đa đang sử dụng hơn 800 tệp dữ liệu hành chính trong công tác thống kê và 40% các chương trình của thống kê Ca-na-đa dựa trên toàn bộ hoặc một phần dữ liệu hành chính sẵn có.

Tại Việt Nam

Hiện nay, Tổng cục Thống kê (TCTK) sử dụng 3 nguồn dữ liệu chính: dữ liệu từ điều tra, DLHC và các báo cáo thống kê. Theo nhận định, nguồn dữ liệu từ chế độ báo cáo và DLHC có thể cung cấp thông tin tính toán khoảng gần một nửa trong tổng số 350 chỉ tiêu trong Hệ thống chỉ tiêu Thống kê quốc gia. Đối với các báo cáo thống kê, nguồn dữ liệu đầu vào phục vụ cho báo cáo lại chính là các nguồn DLHC của khu vực công.

Từ tháng 8 năm 2016 đến tháng 4 năm 2017, Tổng cục Thống kê đã nhận được sự hỗ trợ từ UN-ESCAP giúp xây dựng phần mềm chiết xuất số liệu thống kê kinh tế từ dữ liệu thuế, dựa trên Biên bản ghi nhớ giữa Tổng cục Thống kê và Tổng cục Thuế (ban hành ngày 10 tháng 11 năm 2015).

Tổng cục Thống kê cũng đang xúc tiến hợp tác với Tổng cục Hải quan trong việc xây dựng quy trình, biểu mẫu và thử nghiệm sử dụng dữ liệu hải quan để sản xuất số liệu thống kê xuất nhập khẩu. Liên quan tới việc sử dụng số liệu hành chính trong thống kê giáo dục, cơ quan Thống kê Đan Mạch đã thực hiện đợt khảo sát thí điểm một số trường tiểu học, trung học cơ sở và đại học tại tỉnh Bắc Ninh để xem xét, đánh giá việc sử dụng dữ liệu hành chính trong sản xuất số liệu thống kê cũng như luồng dữ liệu từ các trường đến cơ quan chủ quản, từ đó nghiên cứu xây dựng báo cáo nghiên cứu khả thi làm cơ sở để tìm các nhà tài trợ thực hiện dự án cho Tổng cục Thống kê.

Ngoài ra, hoạt động đăng ký hộ tịch hiện đang được thiết kế và thực hiện tại Việt Nam, điều này sẽ tạo điều kiện thuận lợi giúp chuyển dịch việc sản xuất một số chỉ tiêu thống kê nhất định từ điều tra thống kê sang sử dụng hồ sơ hành chính, đặc biệt việc sử dụng duy nhất một định danh cá nhân sẽ giúp Tổng cục Thống kê xây dựng kho dữ liệu chuỗi thời gian theo chiều dọc (longitudinal time series) trong lĩnh vực thống kê xã hội.

Dự án "Hiện đại hóa sản xuất thống kê của Việt Nam" do Ngân hàng Thế giới (WB) hỗ trợ cũng đang bước đầu giúp Việt Nam trong việc đánh giá và lồng ghép dữ liệu hành chính về giáo dục, thuế và hải quan phục vụ sản xuất số liệu thống kê nhà nước.

Ngày 22/5/2015, Thủ tướng Chính phủ ban hành Quyết định số 714/QĐ-TTg về Danh mục Cơ sở dữ liệu quốc gia (CSDLQG). Theo đó cần ưu tiên triển khai khai tạo nền tảng phát triển chính phủ điện tử, bao gồm 6 CSDLQG.

STT	Cơ sở dữ liệu quốc gia	Cơ quan chủ quản	Mô tả vắn tắt
1	Cơ sở dữ liệu quốc gia về Dân cư	Bộ Công an	Thông tin gốc về người dân phục vụ quản lý hành chính về cư trú, hộ tịch và sử dụng chung giữa các cơ quan nhà nước; đơn giản hóa thủ tục hành chính liên quan đến người dân
2	Cơ sở dữ liệu Đất đai quốc gia	Bộ Tài nguyên và Môi trường	Thông tin về sử dụng đất đai
3	Cơ sở dữ liệu quốc gia về Đăng ký doanh nghiệp	Bộ Kế hoạch và Đầu tư	Lưu trữ thông tin cơ bản về doanh nghiệp, phục vụ: Quản lý và chia sẻ, sử dụng chung giữa các cơ quan nhà nước; đơn giản hóa thủ tục hành chính liên quan đến doanh nghiệp
4	Cơ sở dữ liệu quốc gia về Thống kê tổng hợp về dân số	Bộ Kế hoạch và Đầu tư	Thông tin tổng hợp về người dân phục vụ chia sẻ, dùng chung đa ngành, đa lĩnh vực
5	Cơ sở dữ liệu quốc gia về Tài chính	Bộ Tài chính	Thông tin cơ bản về tài chính, ngân sách như: Thu/chi ngân sách nhà nước; nợ công; vốn nhà nước tại doanh nghiệp...
6	Cơ sở dữ liệu quốc gia về Bảo hiểm	Bảo hiểm Xã hội Việt Nam	Thông tin cơ bản về bảo hiểm y tế, bảo hiểm xã hội

Có thể thấy, việc khai thác và sử dụng dữ liệu hành chính trong sản xuất thông tin thống kê đang trở thành xu hướng mới trong công tác thống kê của nhiều nước trên thế giới nói chung và Việt Nam nói riêng.

Sự cấp thiết khai thác dữ liệu hành chính phục vụ công tác thống kê

Thực tế việc khai thác dữ liệu hành chính phục vụ sản xuất thông tin thống kê hiện nay là vô cùng cần thiết bởi những lý do sau:

Thứ nhất, xuất phát từ chính bản thân những ưu thế mà nguồn dữ liệu hành chính mang đến.

Một là, giảm chi phí thu thập số liệu thống kê

Nguồn dữ liệu hành chính là nguồn dữ liệu lớn, đa dạng và sẵn có, vì vậy nếu các cơ quan thống kê khi khai thác sử dụng nguồn dữ liệu này để sản xuất số liệu thống kê nhà nước thì sẽ tiết kiệm

được chi phí so với việc thu thập dữ liệu thống kê thông qua các cuộc điều tra bởi sẽ không mất thêm các khoản chi phí khác, ngoại trừ các khoản phụ phí hoặc chi phí liên quan đến làm sạch dữ liệu.

Trong một vài trường hợp hoặc một vài khâu chi phí để thu thập, khai thác DLHC gần bằng chi phí thu thập dữ liệu ở các cuộc điều tra tại địa bàn, tuy nhiên nếu xét toàn bộ quá trình từ lúc bắt đầu thu thập cho đến khi kết thúc để có thể ra được một bộ dữ liệu hoàn chỉnh thì khai thác DLHC có giá rẻ hơn nhiều.

Từ bảng chi phí cho cuộc tổng điều tra dân số và nhà ở của các quốc gia thuộc Liên minh châu Âu năm 2000-2001 (Theo Bảng 22 của ấn phẩm Eurostat) cho thấy sự khác biệt lớn về chi phí trên đầu người giữa Phần Lan, quốc gia điều tra dân số dựa trên các nguồn hành chính so với các quốc gia khác như Anh và Úc là hai quốc gia sử dụng bảng câu hỏi truyền thống.

Quốc gia	Tổng chi phí (Triệu EURO)	Chi phí bình quân một người (EURO)
Bỉ	24	2.3
Hy Lạp	50	4.5
Tây Ba Nha	167	4.1
Pháp	248	4.1
Ai len	44	11.2
Ý	298	5.3
Tiệp khắc	5	10.6
Úc	56	6.9
Bồ Đào Nha	46	4.5
Phần Lan	0.8	0.2
Anh	367	6.2
Na Uy	15	3.3
Thụy sỹ	99	13.6
Séc	80	7.8
Es-tôn-nia	10	7.4
Hung ga ry	40	3.9

Hai là, giảm tải gánh nặng của thu thập thông tin thống kê

Khi khai thác nguồn dữ liệu hành chính sẵn có sẽ giúp giảm tải gánh nặng đáng kể so với quy trình khai thác sản xuất số liệu thống kê truyền thống. Gồm: (i) Giảm gánh nặng về nguồn nhân lực và thủ tục hành chính. Việc giảm khối lượng thông tin cần thu thập trong bảng hỏi của các cuộc điều tra sẽ giúp giảm tải các gánh nặng về công tác chuẩn bị, tập huấn điều tra, các thủ tục hành chính cũng như giảm thời gian thu thập thông tin tại địa bàn. (ii) Giảm gánh nặng đối với người được phỏng vấn trong các cuộc điều tra. Việc tận dụng khai thác nguồn DLHC để giảm tải các chỉ tiêu cần thu thập qua điều tra, điều này sẽ giúp giảm gánh nặng đối với người trả lời.

Ba là, tính kịp thời và mức độ thường xuyên, liên tục của số liệu

(i) Thông tin thu thập từ các cuộc điều tra chuyên môn luôn có độ trễ nhất định do để triển khai một cuộc điều tra thống kê cần phải được tiến hành đúng theo trình tự, đảm bảo đúng và đầy đủ tất cả các khâu. Đặc biệt, đối với công tác thu thập thông tin tại địa bàn luôn mất một khoảng thời gian khá dài và sẽ luôn phát sinh các vấn đề ngoài dự tính trong quá trình điều tra làm tăng độ trễ của số liệu. Số liệu thu thập tại địa bàn sau đó sẽ cần khoảng thời gian để nhập tin, rà soát, làm sạch trước khi tiến hành tổng hợp, phân tích và tính toán.

Trái lại, với nguồn DLHC thì các thông tin đã sẵn có không cần phải tiến hành các khâu: chuẩn bị, tập huấn và triển khai thu thập dưới địa bàn.

(ii) Mức độ thường xuyên, liên tục của nguồn số liệu:

Dữ liệu hành chính là dữ liệu được thu thập có tính liên tục, trực tiếp. Các DLHC luôn được cập nhật thường xuyên hàng năm, hàng quý, hàng tháng và thậm chí hàng ngày, hàng giờ tùy thuộc vào nhu cầu về nguồn dữ liệu theo quy định của pháp luật. Đối với quá trình phân tích xu hướng của các hiện tượng kinh tế - xã hội và dự báo thống kê thì việc có được nguồn số liệu thường xuyên liên tục cập nhật theo một chuỗi thời gian như vậy là điều thực sự cần thiết, có thể nói đây được xem là một trong những thế mạnh thực sự của nguồn dữ liệu hành chính.

Trong khi thông tin thống kê cập nhật theo hàng tháng, hàng quý thu thập, tổng hợp từ các cuộc tổng điều tra, điều tra chọn mẫu hay điều tra chuyên đề... là không khả thi. Các cuộc điều tra từ nguồn ngân sách hay được tài trợ để khi tiến hành thường sẽ thực hiện theo định kỳ hàng năm hoặc 3 đến 5 năm hoặc lâu hơn (ví dụ: tổng điều tra dân số và nhà ở diễn ra 5 năm hoặc 10 tùy từng quốc gia, ở Việt Nam được tiến hành 10 năm một lần).

Bốn là, cung cấp các thông tin mang tính "lịch sử"

Nguồn DLHC có thể giúp cho các nhà làm thống kê cũng như các cơ quan thống kê có thể khai thác và phân tích dữ liệu theo một chuỗi thời gian giúp nghiên cứu được sự biến đổi của các hiện tượng kinh tế - xã hội số lớn theo thời gian, từ đó có thể chỉ ra được bước ngoặt biến đổi của các hiện tượng kinh tế - xã hội và gắn nó với sự biến đổi về mặt lịch sử, chính trị, văn hóa và xã hội của mỗi quốc gia cũng như của toàn thế giới.

Năm là, có độ bao phủ rộng, thông tin đa dạng và phân tổ được theo nhiều tiêu thức

Dữ liệu hành chính được thu thập dựa trên quy định của pháp luật phục vụ công tác quản lý của các cơ quan hành chính của tất cả từ các cơ quan, tổ chức đến các cá nhân, chính vì thế, dữ liệu hành chính sẽ có độ bao phủ rộng. Ở nhiều quốc gia trên thế giới hay ở một số lĩnh vực cụ thể thì DLHC có tính bao phủ gần như 100% dân số giúp có thể phân tổ cũng như đảm bảo độ tin cậy ở cấp nhỏ.

Sáu là, giảm sai số trong điều tra thống kê

Trong nghiên cứu thống kê có hai loại sai số là "sai số phi chọn mẫu" và "sai số chọn mẫu". Sai số phi chọn mẫu là sai số do đăng ký, ghi chép và nó xảy ra với tất cả các cuộc điều tra thống kê (điều tra mẫu, điều tra trọng điểm, điều tra chuyên đề và tổng điều tra) cũng như xảy ra đối với cả công tác tổng hợp báo cáo thống kê định kỳ. Sai số chọn mẫu hay còn gọi là sai số do tính đại diện, sai số này chỉ xảy ra trong điều tra chọn mẫu. Cả hai loại sai số này sẽ được khắc phục thay thế bằng việc khai thác nguồn dữ liệu hành chính sẵn có.

Bảy là, khắc phục hiện tượng từ chối trả lời phỏng vấn

Việc từ chối phỏng vấn là một vấn đề cần được quan tâm trong điều tra thống kê, khi tỷ lệ từ chối phỏng vấn nhiều sẽ làm tăng tỷ lệ dữ liệu bị mất (missing data) dẫn đến làm giảm chất lượng của nguồn số liệu khi chúng ta tiến hành tổng hợp và suy rộng cho tổng thể. DLHC đa phần là các dữ liệu được thực hiện theo quy định của pháp luật nên các thông tin đăng ký, kê khai luôn luôn được thực hiện, khai thác nguồn DLHC là khai thác nguồn dữ liệu sẵn có, vì vậy sẽ khắc phục được hiện tượng từ chối trả lời ở điều tra thống kê.

Thứ hai, xuất phát từ thực tế nền thống kê Việt Nam

Đối với nền thống kê nước ta hiện nay sẽ khó khăn nếu chỉ dựa trên hệ thống sản xuất số liệu thống kê hiện có. Công tác thống kê ngoài việc cần phải thích ứng liên tục với các yêu cầu mới, để giảm gánh nặng thống kê nhà nước cần phải thay thế các quy trình sản xuất số liệu thống kê tốn kém và cồng kềnh bằng các quy trình sản xuất tích hợp giúp tiết kiệm chi phí, thời gian, kết hợp với việc sử dụng mới và mở rộng các nguồn dữ liệu hiện có, dựa nhiều hơn vào các dữ liệu hành chính sẵn có từ các cơ quan chính phủ.

Luật Thống kê số 89/2015/QH13 Mục 1 Chương III (từ Điều 36 đến Điều 39) đã có những quy định về việc sử dụng dữ liệu hành chính cho hoạt động thống kê nhà nước cho thấy Chính phủ và Tổng cục Thống kê luôn ý thức rằng việc sử dụng dữ liệu hành chính cho công tác sản xuất số liệu thống kê là hết sức cần thiết. Tuy nhiên thực tế hiện nay, Tổng cục Thống kê vẫn chỉ thu thập các thông tin thống kê dựa trên hai kênh chủ yếu là điều tra thống kê và chế độ báo cáo thống kê (thông qua các các thông tin từ các báo cáo thống kê của các bộ, ngành và địa phương), trong khi việc thu thập và khai thác thông tin từ các nguồn dữ liệu hành chính vẫn đang bị bỏ ngỏ.

Cuối cùng, hiện nay nguồn dữ liệu hành chính khá đầy đủ và sẵn có ở rất nhiều lĩnh vực liên quan đến thống kê kinh tế, tài chính hay xã hội và môi trường như: Thuế, Hải quan, Y tế, Văn hóa, Giáo dục, Thông tin Truyền thông, Tội phạm, An toàn Giao thông,… nên có thể thấy việc sử dụng dữ liệu hành chính có tiềm năng rất lớn phục vụ trong công tác thống kê nếu được khai thác và tận dụng.

Có thể thấy, việc khai thác sử dụng nguồn dữ liệu hành chính là cần thiết, tuy nhiên hiện nay trong công tác Thống kê vẫn chưa được khai thác toàn diện vì các CSDL chuyên ngành vẫn đang trong quá trình xây dựng và hoàn thiện để kết nối lên cổng CSDL quốc gia. Hy vọng trong tương lai không xa, khi 6 CSDLQG hoàn thiện và có một khung pháp lý đầy đủ, kết hợp với cơ sở hạ tầng công nghệ thông tin đủ mạnh thì dữ liệu hành chính sẽ là một trong những kênh được khai thác và sử dụng chính trong công tác Thống kê nhà nước./.

Tài liệu tham khảo:

1. Asian Development Bank – *Administrative Data Sources for compiling millennium development goalds and related indicators;*

2. Anders Wallgren and Britt Wallgren - Register - Based Statistics: Statistical Methods for Administrative Data

3. European Commission (2017), Hướng tới hệ thống thống kê dựa trên đăng ký hành chính;

4. CODED Eurostat's Concepts and Definitions Database, Hyperlink:

http://ec.europa.eu/eur stat/ramon/nomenclatures/index.cfm?TargetUrl =DSP_GLOSSARY_NOM_DTL_VIEW&StrNom=CODED2&StrLanguageCode=EN&IntKey=20159524&RdoSearch=BEGIN&TxtSearch=admins&CboTheme=&IsTer=&IntCurrentPage=1&ter_valid=0