DỮ LIỆU LỚN: NHỮNG XEM XÉT ĐƯỢC ĐƯA RA

Fride Eeg – Henriksen và Peter Hackl
(Trích phần 2 - Tạp chí Khoa học Thống kê của Hiệp hội Quốc tế về Thống kê Nhà nước[39] tháng 6/2015)

Tạp chí Khoa học Thống kê là tạp chí hàng đầu của Hiệp hội Quốc tế về Thống kê Nhà nước, chủ đề được đề cập bao gồm phương pháp luận, ứng dụng và những vấn đề về thống kê đang được thế giới quan tâm. Tập 31, số 2 xuất bản vào tháng 6/2015 đã dành một phần riêng về chủ đề "Dữ liệu lớn" để giới thiệu một số bài nghiên cứu: mô tả tổng quan về dữ liệu lớn; kinh nghiệm của các nước tiên phong trong ứng dụng dữ liệu lớn, đồng thời minh chứng cho sự thích hợp của dữ liệu lớn khi thay thế các dữ liệu truyền thống; những bình luận về cuộc cách mạng dữ liệu và thách thức đặt ra khi ứng dụng. Thông tin khoa học Thống kê trân trọng giới thiệu đến quý độc giả những nội dung trao đổi trên về "Dữ liệu lớn".

Dữ liệu lớn: Những xem xét đưa ra

Dữ liệu lớn là khái niệm mà tại thời điểm hiện tại có lẽ hầu như thường được tham khảo trong bối cảnh của khoa học thông tin và công nghệ thông tin; sự quan tâm khác thường hoặc những cường điệu có thể xảy ra cũng ảnh hưởng đến thống kê nhà nước. Đó là do hai yếu tố:

- Dữ liệu lớn là một từ đồng nghĩa về sự tồn tại của một số lượng khổng lồ và phát triển của thông tin số từ tất cả các lĩnh vực của đời sống con người.

- Thông tin tới tập là dấu hiệu hứa hẹn sẽ nhìn thấy và hiểu được và chi tiết hơn thực tế và các mối quan hệ đang thống trị thế giới chúng ta.

Khái niệm dữ liệu lớn

Dù có mối quan tâm lớn và sự phổ biến về Dữ liệu lớn, việc xác định khái niệm Dữ liệu lớn được chấp thuận rõ ràng và chung còn rất xa vời [2]. Sự phát triển công nghệ, xã hội và kinh tế hiện tại bao gồm sự tăng trưởng dịch vụ và cơ sở vật chất thông minh, việc tăng trưởng tính có lợi và hiệu quả của mạng Internet, sự hấp dẫn của các trang web mạng xã hội và sự phổ biến và có mặt khắp nơi của hệ thống công nghệ thông tin là kết quả của sự ra đời luồng rất lớn về dữ liệu số. Sự phức tạp về cấu trúc và năng động của bộ dữ liệu tương ứng, những thách thức về phát triển công cụ phần mềm phù hợp cho phân tích dữ liệu, nhìn chung tính đa dạng của các tiềm năng tận dụng khối lượng lớn dự liệu hiện có làm nó khó khăn trong việc tìm ra một định nghĩa phù hợp và có thể ứng dụng nói chung. Đặc điểm thường được đề cập của Dữ liệu lớn bởi 3 - hoặc nhiều hơn - Vs (số lượng lưu trữ, tốc độ xử lý, tính đa dạng - cũng như độ chính xác và giá trị thông tin), không nắm bắt được phạm vi lớn của các

[39] *Statistical Journal of The International Association for Official Statistics (IAOS)*

tập hợp dữ liệu tương ứng và các tiềm năng rộng lớn của việc sử dụng những dữ liệu này. Một khía cạnh có liên quan cao là Dữ liệu lớn quá rộng và phức tạp mà các công cụ quản lý cơ sở dữ liệu truyền thống và ứng dụng xử lý dữ liệu là những phương tiện không thực hiện được và không hiệu quả. Điều này được minh họa bởi cái nhìn của các loại nguồn dữ liệu được nhìn thấy điển hình trong bối cảnh của Dữ liệu lớn: Những nguồn dữ liệu có thể là:

- Hành chính, ví dụ hồ sơ bệnh án, hồ sơ bảo hiểm, hồ sơ ngân hàng.
- Các giao dịch thương mại, ví dụ giao dịch thẻ tín dụng, máy quét trong siêu thị.
- Các cảm biến, ví dụ ảnh vệ tinh, cảm biến môi trường, cảm biến đường.
- Thiết bị theo dõi, ví dụ dữ liệu theo dõi từ điện thoại di động, GPS.
- Dấu vết của hành vi con người, ví dụ tìm kiếm trực tuyến, xem trang trực tuyến.
- Tài liệu dẫn chứng về ý kiến, ví dụ các bình luận đăng lên ở các truyền thông xã hội.

Dữ liệu lớn và thống kê nhà nước

Đối với thống kê nhà nước, một vài nguồn có thể, hoặc được hy vọng được sử dụng như nguồn dữ liệu thay thế hoặc bổ sung. Để hoàn thành bổn phận bắt buộc bởi chương trình thống kê, các Viện thống kê quốc gia (NSIs) thu thập dữ liệu trong các cuộc Tổng điều tra hoặc điều tra, hoặc họ sử dụng dữ liệu từ nguồn hành chính. Xu hướng giảm bớt gánh nặng trả lời cho các doanh nghiệp và hộ gia đình và tăng nhu cầu cho các sản phẩm thống kê mới cho phép NSIs tìm ra nguồn dữ liệu mới. Tính đa dạng và sẵn có của dữ liệu hành chính ngày càng tăng đang đạt tới sự thích hợp trong sản phẩm thống kê. Nhưng các nguồn dữ liệu khác được đề cập ở trên cũng rất thú vị có tiềm năng như một đầu ra của thống kê nhà nước. Sử dụng những dữ liệu này có thể giảm thời gian sản xuất và chi phí của thống kê, thực tế khác nữa là còn làm tăng sức hút của những nguồn dữ liệu này.

Mối quan tâm trong sử dụng các nguồn dữ liệu được đề cập cho sản xuất thống kê nhà nước bắt đầu từ khoảng nửa thế kỉ trước. Theo sau một yêu cầu của những người tham dự tại Hội nghị cấp cao về Hợp lý hóa các dịch vụ và sản phẩm thống kê năm 2012, báo cáo ""Dữ liệu lớn" có ý nghĩa gì đối với thống kê nhà nước?" [3] vạch ra những cơ hội và thách thức mà Dữ liệu lớn đề ra cho thống kê nhà nước. Hưởng ứng bản báo cáo này và theo sau đề xuất của nhóm nhiệm vụ bao gồm đại diện của 13 tổ chức thống kê quốc gia và quốc tế, Dự án Dữ liệu lớn [4] được thành lập. Báo cáo "Dữ liệu lớn lớn như thế nào?" [5] là một mô tả có giá trị và cập nhập vai trò tiềm năng của Dữ liệu lớn đối với thống kê nhà nước, đặc biệt là những thách thức và yêu cầu xét trên phương diện các phương pháp thống kê bao gồm ban hành chất lượng, của công nghệ thông tin, và của năng lực và kĩ năng của nhân viên. Trong năm 2014, Ủy ban Thống kê Liên hợp quốc thành lập Nhóm làm việc toàn cầu Liên hợp quốc với 8 đội làm nhiệm vụ về nhiều chủ đề bao gồm xây

dựng đào tạo và năng lực, dữ liệu điện thoại di động, hình ảnh vệ tinh và dữ liệu truyền thông xã hội [7]. Cục Thống kê của Cộng đồng châu Âu đã đang bao gồm tất cả những hoạt động này từ lúc bắt đầu. Có nhiều quốc gia đi tiên phong trong việc khảo sát các tiềm năng của Dữ liệu lớn. Ví dụ nổi bật nhất có lẽ là Văn phòng Thống kê Úc (ABS); xem Tam và Clarke (2015).

Bằng chứng cho mối quan tâm rất lớn của thống kê nhà nước về Dữ liệu lớn là Dữ liệu lớn được đề cao tại nhiều hội nghị, hội thảo, và các sự kiện khác trong suốt những năm gần đây. Ví dụ như Sự kiện Dữ liệu lớn của Thống kê châu Âu tại Rome (2014), Hội nghị quốc tế về Dữ liệu lớn trong Thống kê nhà nước tại Bắc Kinh (2014), và Hội thảo vệ tinh UNECE NTTS 2015 về Dữ liệu lớn ở Brussels (2015). Các bài luận liên quan đến và các bản báo cáo về vấn đề Dữ liệu lớn đang đóng vai trò tiên phong trong các sự kiện như Hội nghị của Giám đốc các Cơ quan thống kê quốc gia DGINS 2013 ở The Hague, Hội nghị Chất lượng Thống kê châu Âu 2014 (Q2014) ở Vienna, Hội nghị của Hiệp hội quốc tế về Thống kê nhà nước 2014 (IAOS 2014) ở Đà Nẵng, Ủy ban Thống kê Liên hợp quốc năm 2015, và nhiều sự kiện khác. Rất nhiều đóng góp đang giải quyết các vấn đề về khái niệm hoặc chiến lược. Tuy nhiên, các báo cáo về Dữ liệu lớn - chủ yếu dựa trên kinh nghiệm - giải thích Dữ liệu lớn có thể được sử dụng như thế nào trong thống kê nhà nước và các vấn đề phương pháp luận và công nghệ phải được giải quyết là gì.

Kinh nghiệm và Thách thức

Một cái nhìn gần hơn về những dự án này chỉ ra rằng các phương pháp thống kê và các công cụ công nghệ thông tin được sử dụng trong việc giải quyết dữ liệu từ các nguồn Dữ liệu lớn điển hình là đặc trưng cho sản phẩm thống kê. Trong các lĩnh vực tiếp theo, kinh nghiệm sử dụng Dữ liệu lớn trong thống kê nhà nước có trong:

- Thống kê sử dụng Công nghệ thông tin và truyền thông (ICT);
- Thống kê Giá;
- Thống kê Thị trường lao động;
- Thống kê Du lịch;
- Thống kê Giao thông và vận tải;
- Tổng điều tra Nông nghiệp và các cuộc điều tra;

Trong Dự án Dữ liệu lớn [4], nhiều Dự án Dữ liệu lớn đang được quản lý bởi NSIs từ các quốc gia tham gia như Hà Lan, Ý, Mỹ, Ireland, Úc và Slovenia.

Các nguồn dữ liệu là liên mạng trong hoàn cảnh của sử dụng thống kê ICT, giá và thị trường lao động [1]. Số lượng của dữ liệu liên quan thường khổng lồ và có khả năng là được phân bổ vượt quá số lượng lớn của các trang web. Điều này có nghĩa là các công cụ khôi phục lại các trang web liên quan là cần thiết cũng như các công cụ cho việc thu thập dữ liệu liên quan; trình duyệt thu thập dữ liệu các trang web và

trích xuất nội dung trang web là lần lượt là tên của những công cụ này. Để giải quyết kích thước lớn của các nguồn Dữ liệu lớn, môi trường lập trình đặc biệt đang được phát triển: ví dụ, Map - Reduce là một công cụ lập trình và môi trường liên kết cho hình thành và xử lý bộ dữ liệu rộng. Khung lập trình như hệ thống nguồn mở Hadoop cho phép soạn thảo chương trình để xử lý các vấn đề Map - Reduce qua bộ dữ liệu rộng sử dụng số lượng lớn máy tính và đưa ra tập tin đầu ra trong hệ thống tập tin tên là HDFS (Hệ lưu tập tin phân tán được dùng bởi Hadoop). Hadoop phù hợp cho các quy trình một đợt vận hành dài, như khai phá dữ liệu; các công cụ như Big Query cho phép lệnh hỏi đặc biệt đòi hỏi các kết quả nhanh chóng. Những thách thức to lớn của Dữ liệu lớn đến công nghệ thông tin có hệ quả rằng các vấn đề IT và - các chuyên gia IT - chiếm ưu thế hơn trong các cuộc thảo luận về Dữ liệu lớn. Việc sử dụng Dữ liệu lớn trong thống kê nhà nước cũng cần thích nghi trong phương pháp luận thống kê. Mối quan tâm về phương pháp luận trong bối cảnh liên mạng và điện thoại di động đặt vào vị trí dữ liệu mang tính biểu tượng của kết quả thống kê: Cơ chế thu thập dữ liệu có cho phép biểu diễn các tập hợp đối với sản phẩm thống kê làm đại diện không, và tập hợp này có trùng hợp với tập hợp đích mà sản phẩm thống kê được xây dựng không? Nếu không, sản phẩm thống kê có thể được giải thích như thế nào? Các vấn đề về phương pháp luận khác quan tâm đến thẩm định chất lượng dữ liệu và các sản phẩm thống kê, sự kết hợp dữ liệu từ những nguồn khác nhau, tính dễ biến động của nguồn dữ liệu qua thời gian, các mối quan tâm riêng, tính bảo mật, và những vấn đề khác. Các vấn đề về phương pháp luận là riêng biệt với các bộ dữ liệu và phải được giải quyết riêng lẻ cho từng bộ dữ liệu. Các báo cáo về những dự án được đề cập bao hàm các vấn đề về phương pháp luận trong phương thúc cũng khá là chung. Tính biểu tượng và cũng như những khía cạnh chất lượng khác của Dữ liệu lớn - trên cơ sở các sản phẩm thống kê là những khía cạnh then chốt cho tính đáng tin cậy của chúng. Sử dụng Dữ liệu lớn trong thống kê nhà nước cần các kĩ năng và năng lực mới. Một cuộc điều tra trong các tổ chức thống kê [6] đã chỉ ra rằng khoảng 37% làm việc với Dữ liệu lớn, và còn lại 43% dự định sẽ làm điều này trong tương lai gần. Trong khi hầu hết người được hỏi nói rằng nhân viên của họ thường ở trình độ trung bình hoặc cao cấp với các công cụ IT như Java, SAS, cơ sở dữ liệu SQL, và R, không hoặc chỉ các kĩ năng cơ bản được nói đến tồn tại trong các công cụ như Map Reduce và Hadoop. Kết quả này bộc lộ cả sự chú ý mạnh mẽ của các tổ chức thống kê về Dữ liệu lớn và nhu cầu để nâng cao năng lực và kĩ năng để tích hợp những tiềm lực mới trong đời sống hàng ngày của các tổ chức thống kê. Các khóa đào tạo trong NSIs hoặc các cơ quan như Cục Thống kê của Cộng đồng châu Âu cũng như các dự án thiết thực sẽ giúp xây dựng khả năng về công nghệ IT mà còn trong phương pháp luận thống kê.

Dữ liệu lớn được cho rằng tạo ra nhiều cơ hội trong phạm vi phổ biến thống kê, một phạm vi mà còn chưa được chú trọng nhiều không chỉ trong các cuộc thảo luận

mà còn trong các bài báo về các phần đặc biệt. Sự quan tâm tăng lên về hiển thị thống kê là một mặt quan trọng của việc này. Những triển vọng mới cho phân tích và hiển thị chắc chắn cũng tạo ra nhiều thách thức cho khả năng xây dựng các cơ quan thống kế quốc gia và quốc tế.

Mục đích của phần đặc biệt này

Phần đặc biệt này trình bày một bài viết tổng quan cũng như kinh nghiệm từ các khu vực đi tiên phong chỉ ra nơi mà các nguồn Dữ liệu lớn có thể chứng minh thích hợp để thay thế các nguồn dữ liệu truyền thống hoặc có thể cho phép sản xuất số liệu thống kê mới.

"Hợp tác quốc tế để hiểu sự phù hợp của Dữ liệu lớn trong thống kê nhà nước" của Steve Vale đưa ra bản kê khai các dự án Dữ liệu lớn được tổ chức để đáp ứng cho các Nhóm cấp cao về hiện đại hoá sản xuất và dịch vụ thống kê [3]. Bài báo mô tả các mục tiêu và ưu tiên của dự án, việc thành lập môi trường điện toán, được gọi là "sandbox", cho quản trị và phân tích các bộ dữ liệu quy mô lớn và đưa ra một cái nhìn tổng quan về kết quả. Sự liên quan cao của dự án và kết quả của nó là do thực tế 7 đội đã tham gia vào công việc thử nghiệm thực tế trong khu vực như chỉ số giá tiêu dùng, dữ liệu điện thoại di động, dụng cụ đo thông minh, vòng lặp giao thông, cổng thông tin việc làm, trích rút nội dung trang web và phương tiện truyền thông xã hội.

Trong các chương trình của Hội nghị Chất lượng của Cục Thống kê của Cộng đồng châu Âu 2014 (Q2014) và Hội nghị IAOS 2014 (IAOS2014) tại Đà Nẵng, một số báo cáo về các dự án tiên phong cụ thể được đưa ra. Các bài viết khác đều dựa trên các bài luận đã được trình bày tại một trong hai hội nghị này.

Một bài báo báo cáo về những kinh nghiệm trong việc sản xuất các chỉ số giá tiêu dùng. "Kỹ thuật Trích rút nội dung trang web để thu thập dữ liệu trên thiết bị điện tử tiêu dùng và giá vé máy bay do HICP của Italia biên soạn" của Riccardo Giannini và các đồng tác giả từ Viện Nghiên cứu thống kê quốc gia ý (ISTAT) cho thấy chi tiết làm thế nào kỹ thuật trích rút nội dung trang web có thể được sử dụng để thu thập dữ liệu về giá cho thiết bị điện tử tiêu dùng và giá vé máy bay. "Việc sản xuất các hồ sơ/mẫu lương của các chuyên gia ICT: Chuyển từ cơ sở dữ liệu có cấu trúc sang phân tích dữ liệu lớn" của Ramachandran Ramasamy từ Hiệp hội công nghệ thông tin quốc gia của Malaysia báo cáo về sản xuất hồ sơ/mẫu lương trên cơ sở dữ liệu từ một hệ thống đăng ký việc làm trực tuyến khu vực tư nhân. Thống kê được cung cấp tại một mức độ cao về phân tách. Bài viết thảo luận chi tiết các vấn đề về chất lượng bao gồm sự nhất quán và ổn định trong xu hướng.

Hai đóng góp khác đang giải quyết các khía cạnh phương pháp luận. "Remake-Remodel - dữ liệu lớn nên thay đổi các mô hình mẫu trong thống kê nhà nước không?" của Barteld Braaksma và Kees Zeelenberg từ Cơ quan thống kê của Hà Lan thảo luận về việc sử dụng các mô hình để đánh giá và nâng cao tính biểu tượng của

các nguồn Big data. Bài luận phác thảo những ứng dụng có thể xảy ra. Bài luận "Chỉ số chất lượng cho thống kê dựa trên nhiều nguồn" của Mihaela Agafitei và các đồng tác giả từ Cục Thống kê của Cộng đồng châu Âu phân tích sự phù hợp của các biện pháp tiêu chuẩn chất lượng cho nhiều thống kê nguồn và đề xuất những cải tiến chất lượng mà từ đó phải được điều tra trong công việc sau này hơn nữa.

Kết luận

Kết luận chung từ tập hợp các bài viết trong phần đặc biệt này có thể được rút ra như sau: Tính khả thi và tiềm năng của việc sử dụng Dữ liệu lớn trong thống kê nhà nước phải được đánh giá theo từng trường hợp. Trong một số lĩnh vực sử dụng các nguồn Dữ liệu lớn đã được chứng minh là có tính khả thi. Việc lựa chọn các công nghệ IT và phương pháp thống kê thích hợp phải được cụ thể trong từng tình huống. Ngoài ra các vấn đề như tính hình tượng và chất lượng của kết quả thống kê, hoặc sự bảo mật và nguy cơ tiết lộ dữ liệu cá nhân cần phải được đánh giá riêng trong từng trường hợp. Không còn nghi ngờ gì nữa Dữ liệu lớn sẽ có một chỗ đứng trong tương lai trong thống kê nhà nước, giúp giảm chi phí và gánh nặng của người trả lời. Tuy nhiên, những nỗ lực lớn sẽ là cần thiết để thiết lập việc sử dụng thành thạo thường xuyên Dữ liệu lớn, và những cách tiếp cận mới sẽ cần thiết để đánh giá tất cả các khía cạnh của chất lượng.

Tài liệu tham khảo:

[1] G. Barcaroli et al, đối phó với Big tata trong thống kê nhà nước. Hội nghị về quản lý Hệ thống thông tin thống kê (MSIs 2014);

[2] C. Reimsbach-Kounatze, (2015), Sự phát triển của "Big data" và mối liên quan tới các cơ quan thống kê và số liệu thống kê nhà nước: Phân tích sơ bộ, Tạp chí Kinh tế kỹ thuật số OECD số 245, tại http://dx.doi.org/10.1787/5js7t9wqzvg8-en;

[3] UNECE (2013), Những gì là "dữ liệu lớn" cho thống kê nhà nước? Báo cáo của Nhóm cấp cao về hiện đại hoá Sản xuất và Dịch vụ thống kê (HLG), tại http://www1.unece.org/stat/platform/display/hlgbas;

[4] UNECE (2014a), Dự án đề xuất: Vai trò của Big aata trong hiện đại hoá sản xuất thống kê, tại http://www1.unece.org/stat/platform/display/bigdata/2014+Project;

[5] UNECE (2014b), Thế nào là Big data? Vai trò của của Big data trong thống kê nhà nước. Báo cáo Hội thảo Virtual Sprint, tại http://www1.unece.org/stat/platform/display/ bigdata/How+big+is+Big+Data;

[6] UNECE (2014c), Câu hỏi về các kỹ năng cần thiết cho những người làm việc với Big data trong tổ chức thống kê. Báo cáo từ tháng 10/2014, tại http://www1.unece.org/stat/platform;

[7] UNECE (2014d), Báo cáo của Nhóm công tác toàn cầu về Big data cho thống kê nhà nước. Ghi chú của Tổng thư ký Hội đồng Bảo an LHQ lần thứ 46, tại http://unstats.un.org/unsd/statcom/doc15.