KHUNG CHẤT LƯỢNG ĐỂ KẾT HỢP DỮ LIỆU ĐIỀU TRA THỐNG KÊ, DỮ LIỆU HÀNH CHÍNH VÀ DỮ LIỆU LỚN CHO THỐNG KÊ CHÍNH THỨC

Yvonne APM Gootzen, Piet JH Daas[*]; Arnout Van Delden[**]

Tóm tắt

Sản xuất số liệu thống kê bằng cách kết hợp các nguồn dữ liệu cho phép tạo ra số liệu thống kê mới, kịp thời và chi tiết hơn. Với kết quả thống kê từ các nguồn dữ liệu khác nhau, cần đánh giá tiềm năng của từng nguồn và khung chất lượng chính là công cụ cho những nhiệm vụ như vậy. Bài viết đề xuất một khung chất lượng bao gồm các chiều có thể áp dụng cho dữ liệu điều tra thống kê, dữ liệu hành chính và dữ liệu lớn để hỗ trợ đánh giá tiềm năng của từng nguồn dữ liệu trong việc đóng góp vào số liệu thống kê dự kiến. Khung chất lượng này đã được áp dụng cho nghiên cứu về dữ liệu di động và nghiên cứu về phát hiện hạt virus trong dữ liệu nước thải.

1. Giới thiệu

Số lượng dữ liệu thống kê đa nguồn dựa trên sự kết hợp giữa dữ liệu điều tra thống kê và dữ liệu hành chính đang ngày càng tăng. Bên cạnh dữ liệu điều tra thống kê và dữ liệu hành chính, dữ liệu lớn cũng đã được ứng dụng thành công trong thống kê chính thức. Dữ liệu lớn đi kèm với những thách thức riêng, khác với những thách thức của việc sử dụng dữ liệu điều tra thống kê hoặc dữ liệu hành chính. Để sản xuất một số liệu thống kê, trước tiên cần đánh giá nội dung và chất lượng của các nguồn dữ liệu sẵn có có đủ hay không. Điều này có thể được thực hiện bằng cách sử dụng khung chất lượng. Kinh nghiệm tại cơ quan Thống kê Hà Lan cho thấy, các khung chất lượng được tạo riêng cho dữ liệu điều tra thống kê và dữ liệu hành chính không thể áp dụng hoàn toàn cho các nguồn dữ liệu lớn. Trong các khung đó, bản chất của nguồn dữ liệu lớn khác nhau nên việc đánh giá trở nên thiếu thông tin. Tương tự, các khung chất lượng được xây dựng riêng cho dữ liệu lớn không được thiết kế để phù hợp với dữ liệu điều tra thống kê và dữ liệu hành chính. Do đó, nảy sinh nhu cầu về khung chất lượng đa nguồn phù hợp và có thể áp dụng cho dữ liệu điều tra thống kê, dữ liệu hành chính, dữ liệu lớn và sự kết hợp giữa chúng.

[*] *Cơ quan Thống kê Hà Lan; Đại học Công nghệ Eindhoven*

[**] *Cơ quan Thống kê Hà Lan*

Khung chất lượng cho hai trong số ba nguồn dữ liệu đã được phát triển cho dữ liệu điều tra thống kê và dữ liệu hành chính cũng như cho dữ liệu điều tra thống kê và dữ liệu lớn. Khung chất lượng không phụ thuộc vào loại dữ liệu đầu vào mà thường tập trung vào dữ liệu đầu ra. Mặc dù mục đích của các khung không phải là để đánh giá một nguồn dữ liệu duy nhất nhằm tạo ra số liệu thống kê đa nguồn, nhưng các chiều chất lượng cơ bản cần được xem xét đối với các nguồn cụ thể.

Bài viết đưa ra những điểm tương đồng giữa các khung chất lượng hiện có và đề xuất một khung có thể áp dụng khi điều tra thống kê, dữ liệu hành chính và dữ liệu lớn được kết hợp. Khung này sẽ được áp dụng trong giai đoạn thiết kế một thống kê (mới). Sau khi áp dụng khung và chọn nguồn sẽ được sử dụng để thống kê, nên phân tích các nguồn dữ liệu để xác thực các đánh giá về khung và chọn phương pháp phù hợp nhất. Về mặt siêu đa chiều do Karr giới thiệu, khung chủ yếu tập trung vào siêu đa chiều dữ liệu về chất lượng. Khung tính đến biến số mục tiêu và nhóm đối tượng mục tiêu của số liệu thống kê dự định. Ngoài ra, nó bao gồm mức tổng hợp dự kiến và dữ liệu đi kèm có sẵn.

Phần còn lại của bài viết được sắp xếp như sau: Phần 2: Mô tả các chiều và danh mục của khung chất lượng; Phần 3: Khung chất lượng được áp dụng cho hai trường hợp nghiên cứu; Phần 4: Kết luận.

2. Các chiều cho danh mục của tập dữ liệu

Khung chất lượng được trình bày trong bài viết được áp dụng cho nhiều tập dữ liệu bằng cách đánh giá riêng từng tập dữ liệu và sau đó kết hợp các kết quả. Chất lượng của từng tập dữ liệu riêng lẻ được đánh giá bằng một tập hợp các chiều. Mỗi chiều bao gồm các danh mục đã được chọn sao cho chúng tóm tắt thông tin liên quan đến quá trình kết hợp tập dữ liệu với các tập dữ liệu khác để tạo ra một thống kê dự kiến trong một bối cảnh nhất định. Các tập dữ liệu khác được xem xét để kết hợp với tập dữ liệu hiện tại được gọi là dữ liệu đi kèm. Nhiều hơn một danh mục có thể được áp dụng cho mỗi chiều.

Việc đầu tiên trước khi áp dụng khung là xác định bối cảnh khung sẽ được áp dụng. Bối cảnh có thể được hiểu là góc độ xem xét tập dữ liệu và cần được xác định trước khi tập dữ liệu được phân loại. Bối cảnh của số liệu thống kê dự kiến bao gồm biến số mục tiêu, đối tượng mục tiêu và mức độ tổng hợp, dữ liệu đi kèm. Chiều thời gian được coi là một phần của mức tổng hợp.

Mỗi nguồn sẽ được đánh giá riêng lẻ. Khi đó các nguồn khác không thuộc đối tượng của đánh giá sẽ được coi là dữ liệu đi kèm. Sau đó, các đánh giá trên mỗi nguồn được kết hợp để xem xét liệu sự kết hợp giữa các nguồn có thể được sử dụng để tạo ra số liệu thống kê dự kiến hay không.

Việc đánh giá một số chiều đòi hỏi thông tin về phương pháp thu thập dữ liệu, đặc biệt khi được thực hiện bởi các tổ chức khác nhau. Nếu thông tin không có sẵn thì bản thân dữ liệu có thể không thể hiện danh mục được áp dụng. Trong trường hợp này, danh mục không xác định được áp dụng nhiều nhất.

Mỗi chiều nhằm trả lời một câu hỏi chính liên quan đến tính hữu ích của nguồn dữ liệu cho mục đích trong ngữ cảnh. Các chiều sau được đề xuất để phân loại tập dữ liệu nhằm

mục đích kết hợp nó với các nguồn khác trong bối cảnh cụ thể:

***Mức độ liên quan**:* Dữ liệu có chứa thông tin liên quan đến thống kê mục tiêu không?

• *Liên quan trực tiếp*. Dữ liệu chứa biến mục tiêu hoặc một biến rất giống với biến mục tiêu, ngụ ý rằng không cần dữ liệu, biến hoặc mô hình đi kèm để trích xuất biến mục tiêu ở mức tổng hợp theo dự kiến.

• *Liên quan gián tiếp*. Dữ liệu chứa thông tin có thể liên quan đến thống kê dự định, nhưng chỉ khi kết hợp với tập dữ liệu, biến hoặc mô hình đi kèm. Hoặc việc tổng hợp biến mục tiêu khả dụng cho một loại đơn vị được liên kết với mức tổng hợp dự định.

• *Không liên quan.* Dữ liệu không chứa thông tin có liên quan đến người dùng. Nếu dữ liệu này không có sẵn, nó sẽ không ảnh hưởng đến kết quả cuối cùng. Nếu dữ liệu đi kèm có sẵn trong tương lai, việc phân loại trong danh mục này cần được xem xét lại.

***Độ bao phủ tổng thể**:* Tổng thể trong dữ liệu hoàn chỉnh đến mức nào so với tổng thể mục tiêu?

• *Bao phủ đầy đủ*. Mỗi đơn vị trong tổng thể mục tiêu xuất hiện chính xác một lần trong dữ liệu.

• *Lặp lại*. Các đơn vị của nhóm đối tượng mục tiêu được đưa vào dữ liệu nhiều lần.

• *Bao phủ quá mức.* Dữ liệu chứa các đơn vị không thuộc nhóm tổng thể mục tiêu.

• *Bao phủ ngầm*. Dữ liệu chứa các đơn vị của tổng thể mục tiêu. Một số đơn vị của nhóm đối tượng mục tiêu không có trong dữ liệu.

• *Chưa xác định*. Không có liên kết trực tiếp giữa các đơn vị trong dữ liệu và nhóm đối tượng mục tiêu. Không thể đưa ra tuyên bố về phạm vi bao phủ của tập dữ liệu.

• *Không có phạm vi bao phủ theo loại đơn vị*. Loại đơn vị của dữ liệu khác với loại đơn vị của nhóm đối tượng mục tiêu. Cần có một số dữ liệu hoặc mô hình đi kèm để chuyển đổi loại đơn vị thành loại đơn vị của nhóm tổng thể mục tiêu, trước khi đánh giá mức độ bao phủ của nhóm tổng thể.

***Tính đại diện của tổng thể**:* Có thể suy ra được liệu tập đơn vị trong dữ liệu có đại diện cho tổng thể mục tiêu hay không?

• *Xác suất được chọn đã biết*. Xác suất được chọn của các đơn vị trong tổng thể mục tiêu đã được biết. Điều này bao gồm các trường hợp có mẫu xác suất hoặc lựa chọn xác định.

• *Xác suất được chọn chưa biết*. Xác suất được chọn của các đơn vị trong tổng thể mục tiêu không được biết đến. Bao gồm các trường hợp có mẫu phi xác suất.

• *Xác suất được chọn khác không*. Tất cả xác suất được chọn của các đơn vị trong tổng thể mục tiêu đều lớn hơn 0.

• *Xác suất được chọn bằng không*. Một số xác suất được chọn trong tổng thể mục tiêu là 0.

• *Chưa xác định*. Tính đại diện của tổng thể không thể được xác định nếu không có sẵn mã định danh duy nhất của các đơn vị hoặc loại đơn vị của tổng thể mục tiêu không có trong dữ liệu. Không có biến đi kèm để đo lường tính đại diện.

***Hiệu lực của biến**:* Tập dữ liệu đo lường biến mục tiêu tốt đến mức nào?

• *Hoàn hảo*. Định nghĩa của biến mục tiêu giống với định nghĩa được sử dụng trong dữ liệu và không xảy ra lỗi đo lường.

• *Định nghĩa không nhất quán*. Định nghĩa của biến mục tiêu hoặc loại đơn vị khác với định nghĩa trong dữ liệu. Sự không nhất quán trong định nghĩa còn được gọi là giá trị (trong) khái niệm.

• *Lỗi đo lường*. Lỗi đo lường làm cho các giá trị trong dữ liệu khác với định nghĩa dự định trong tập dữ liệu.

• *Lỗi mô hình hóa*. Biến trong tập dữ liệu trước đây được lấy từ một biến khác bằng một mô hình hoặc đạo hàm không hoàn hảo.

• *Lỗi xử lý*. Biến chứa lỗi từ bước xử lý trước đó, chẳng hạn như nhập dữ liệu hoặc chỉnh sửa thủ công.

• *Lỗi nhân quả*. Lỗi đã xảy ra do bỏ qua các kết nối nhân quả giữa các biến trong phiên bản trước của dữ liệu (hoặc nhiều tập dữ liệu nếu tập hiện tại là sự kết hợp của các tập dữ liệu).

• *Chưa xác định*. Định nghĩa, quy trình đo lường hoặc quy trình mô hình hóa của biến trong dữ liệu chưa được biết (một phần).

***Tính ổn định của khái niệm**:* Đánh giá tập dữ liệu theo chiều hiệu lực biến số có ổn định theo thời gian không?

• *Ổn định*. Mức độ nhất quán về định nghĩa, sai số đo lường và sai số mô hình hóa của tập dữ liệu đối với biến mục tiêu ổn định theo thời gian.

• *Phân tán*. Mức độ nhất quán về định nghĩa của tập dữ liệu so với biến mục tiêu thay đổi theo thời gian. Do sự thay đổi định nghĩa của biến mục tiêu theo thời gian, khi sự thay đổi đó không xuất hiện trong tập dữ liệu.

• *Không ổn định*. Lỗi đo lường hoặc lỗi mô hình hóa của tập dữ liệu thay đổi theo thời gian.

• *Không áp dụng được*. Đối với mục đích của nghiên cứu này, khái niệm ổn định là không liên quan. Đây có thể là trường hợp đối với các nguồn mà biến mục tiêu không được đưa vào nguồn dữ liệu.

***Khả năng sửa chữa**:* Liệu những điểm không chính xác (ví dụ sai lệch) trong dữ liệu có thể được sửa chữa hay không, chẳng hạn bằng cách lập mô hình hoặc bằng cách kết hợp với các bộ dữ liệu khác?

• *Không cần thiết*. Không cần chỉnh sửa vì dữ liệu đo lường chính xác biến mục tiêu.

• *Tự sửa lỗi*. Những điểm không chính xác trong tập dữ liệu có thể được sửa bằng cách sử dụng các biến đi kèm trong chính tập dữ liệu đó mà không cần sử dụng các tập dữ liệu khác.

• *Có thể bổ sung - có thể sửa được*. Độ lệch trong tập dữ liệu có thể được sửa bằng cách sử dụng các bộ đi kèm, có thể bằng cách liên kết chúng với các biến từ tập dữ liệu hiện tại.

• *Không thể sửa được*. Dữ liệu không thể được sửa trong bối cảnh nhất định.

• *Chưa xác định.* Không rõ liệu dữ liệu có thể được sửa trong bối cảnh nhất định hay không.

***Tính mới**:* Bản chất của độ trễ thời gian giữa sự xuất hiện của một hiện tượng và thời điểm nó được báo cáo lần đầu tiên là gì?

• *Dựa trên sự kiện*. Dữ liệu liên quan đến một sự kiện sẽ có sẵn sớm sau khi sự kiện xảy ra mà không cần nhóm nhiều sự kiện thành một "phân phối" dữ liệu duy nhất.

• *Định kỳ.* Có một hệ thống đảm bảo việc phát hành dữ liệu định kỳ.

• *Không thường xuyên*. Tính sẵn có của dữ liệu phụ thuộc vào các hành động khó dự đoán của từng cá nhân. Hoặc có thể không có bất kỳ đảm bảo nào rằng tập dữ liệu kế thừa sẽ có sẵn trong tương lai.

***Thời gian xử lý**:* Bản chất của độ trễ thời gian giữa việc có được quyền truy cập vào dữ liệu và số liệu thống kê dự định đã sẵn sàng để xuất bản là gì?

• *Ngay lập tức*. Một hệ thống tự động được áp dụng để đảm bảo dữ liệu có thể được xử lý gần như ngay lập tức, ít nhất là trước khi có đợt dữ liệu tiếp theo.

• *Tự động hóa.* Bất cứ khi nào có một phần dữ liệu mới, nó có thể được xử lý mà không cần nhiều sự can thiệp của con người.

• *Theo yêu cầu*. Dữ liệu được xử lý thủ công và quá trình này được bắt đầu theo yêu cầu mỗi khi có phiên bản mới của dữ liệu.

***Khả năng truy cập**:* Có những hạn chế trong việc truy cập dữ liệu ở mức độ nào?

• *Toàn quyền truy cập*. Quyền truy cập hợp pháp được đảm bảo trong tương lai gần và không giới hạn các tùy chọn. Việc sử dụng dữ liệu được phép để xuất bản số liệu thống kê dự định.

• *Truy cập trả phí*. Dữ liệu có thể truy cập được sau khi trả phí.

• *Quyền truy cập hạn chế*. Các vấn đề pháp lý ngăn cản người dùng truy cập vào toàn bộ dữ liệu hoặc giới hạn phạm vi thống kê dự kiến.

• *Không có quyền truy cập*. Không có quyền truy cập vào dữ liệu hoặc không được phép sử dụng dữ liệu cho thống kê dự định.

***Siêu dữ liệu**:* Các định nghĩa và thông tin về dữ liệu được biết đến ở mức độ nào?

• *Kết hợp*. Siêu dữ liệu đầy đủ, được xác định rõ ràng và hoàn toàn phù hợp với siêu dữ liệu từ các tập dữ liệu đi kèm và số liệu thống kê dự định.

• *Tuân thủ tiêu chuẩn*. Siêu dữ liệu tuân thủ các tiêu chuẩn trong lĩnh vực ứng dụng. Các tiêu chuẩn có thể khác nhau giữa các lĩnh vực ứng dụng khác nhau.

• *Được xác định rõ ràng.* Siêu dữ liệu đầy đủ và được xác định rõ ràng nhưng không phù hợp lắm với siêu dữ liệu từ các tập dữ liệu đi kèm hoặc thống kê dự định.

• *Không xác định được*. Siêu dữ liệu phần lớn có sẵn nhưng chưa rõ ràng và cho phép diễn giải bằng nhiều cách.

• *Chưa đầy đủ.* Siêu dữ liệu phần lớn không có sẵn và cần phải phân tích hoặc giả định dữ liệu khám phá để diễn giải dữ liệu.

***Khả năng so sánh**:* Dữ liệu có thể được so sánh với dữ liệu từ nghiên cứu song song ở mức độ nào?

• *Hoàn toàn có thể so sánh*. Có sự đồng thuận chung về các định nghĩa được sử dụng trong dữ liệu và trong các dự án song song. Về lý thuyết, dữ liệu có thể được trao đổi với dữ liệu từ nghiên cứu song song mà

không có những phức tạp lớn về phương pháp luận.

• *Có thể so sánh một phần*. Có thể dự kiến sẽ có một số khác biệt giữa dữ liệu và dữ liệu trong các nghiên cứu song song, nhưng việc chuyển đổi cho phép so sánh kết quả của nghiên cứu với các nghiên cứu song song.

• *Không thể so sánh*. Dữ liệu này độc đáo đến mức khó có thể so sánh được kết quả với các nghiên cứu song song.

Các chiều: độ bao phủ tổng thể, tính đại diện của tổng thể, hiệu lực của biến số, tính ổn định của khái niệm và khả năng sửa chữa là các chiều của độ chính xác. Các chiều: tính gần đây, thời gian xử lý là các chiều của tính kịp thời và đúng giờ.

Lưu ý rằng một số danh mục không loại trừ lẫn nhau. Mặc dù quá trình phân loại sẽ đơn giản hơn nếu tất cả các danh mục trong một chiều loại trừ lẫn nhau, nhưng số lượng danh mục cần thiết sẽ cao quá mức. Một số lỗi và sự không nhất quán có thể xảy ra đồng thời. Trạng thái của ba loại mâu thuẫn và lỗi này có thể là "không có vấn đề"; "vấn đề hiện tại" và "không biết liệu vấn đề có hiện diện hay không". Điều này dẫn đến tổng cộng 7^3=343 danh mục, trong khi 07 danh mục hiện tại cung cấp đủ định nghĩa để mô tả chính xác một tập dữ liệu theo chiều giá trị. Việc lựa chọn các danh mục không loại trừ lẫn nhau được thực hiện để giữ cho khung có trật tự, đồng thời không làm mất đi sự phức tạp của một tập dữ liệu.

Lưu ý về tính liên kết chặt chẽ

Người đọc quen thuộc với các chiều chất lượng trong số liệu thống kê chính thức có thể thắc mắc tại sao thuật ngữ liên kết chặt chẽ không được đưa vào danh sách các chiều trong khung. Eurostat tuyên bố rằng tính liên kết chặt chẽ thường được sử dụng khi đánh giá mức độ của các đầu ra từ các quy trình thống kê khác nhau có tiềm năng được sử dụng kết hợp. Cùng một quy trình thống kê nhưng đối với các khoảng thời gian khác nhau, các quốc gia/khu vực khác nhau và/hoặc các lĩnh vực khác nhau có tiềm năng được sử dụng khác nhau. Sự khác biệt chính giữa khả năng so sánh và tính liên kết chặt chẽ là khả năng so sánh liên quan đến hình thức và tính liên kết chặt chẽ liên quan đến giá trị của đầu ra. Khung chất lượng được thiết kế cho giai đoạn lập kế hoạch của một thống kê mới. Trong thời điểm này, đã biết đầu ra dự định nhưng không biết giá trị của đầu ra.

Ngoài ra, việc đưa số liệu thống kê có sẵn vào bối cảnh có thể khiến chúng được hiểu là những hạn chế. Không nên hiểu sai mức độ của một thống kê đã được thiết lập và không nên hạn chế sự phát triển của một thống kê dựa trên các bộ dữ liệu hoặc phương pháp mới. Trong trường hợp một số liệu thống kê mới không nhất quán với một số liệu thống kê đã được thiết lập, cả hai số liệu đều nên được thể hiện cùng với lời giải thích về sự khác biệt đó. Số liệu thống kê mới không nên bị loại bỏ vì sự khác biệt với số liệu thống kê đã được thiết lập.

3. Nghiên cứu ứng dụng trong các trường hợp

Khung chất lượng được áp dụng cho hai nghiên cứu trong đó có nhiều loại nguồn dữ liệu được kết hợp.

3.1. Nghiên cứu ứng dụng trong trường hợp: tính di động

Trong nghiên cứu này, dữ liệu hành chính, dữ liệu điều tra thống kê và dữ liệu lớn được kết hợp với dữ liệu mạng lưới đường bộ của Hà Lan. Cường độ giao thông trên mạng lưới đường bộ ở Hà Lan được nghiên cứu bằng cách kết hợp bốn nguồn khác nhau. Trước khi áp dụng khung này cho các tập dữ liệu có sẵn, trước tiên phải chính thức hóa bối cảnh của nghiên cứu. Biến số mục tiêu của thống kê dự định là số lượng ô tô, xe máy chở khách và tổng thể mục tiêu là tập hợp các đoạn đường ở Hà Lan. Mức tổng hợp được xác định là giờ cao điểm buổi sáng trên mỗi đoạn đường. Khung này được áp dụng bốn lần: một lần cho mỗi nguồn, ba nguồn còn lại hoạt động như nguồn dữ liệu đi kèm. Bảng 1 cho thấy kết quả của việc áp dụng khung chất lượng cho các tập dữ liệu được sử dụng trong nghiên cứu điển hình, dựa trên bối cảnh này.

Bốn nguồn dữ liệu đều có loại đơn vị riêng và thoạt nhìn thì không thể kết hợp được. Nguồn đầu tiên là điều tra thống kê du lịch quốc gia của Hà Lan. Trong đó mọi người được yêu cầu báo cáo về việc di chuyển bằng phương tiện giao thông của họ (bao gồm cả phương thức và mục đích) trong một ngày cụ thể. Loại đơn vị của nguồn này là "người". Dữ liệu điều tra thống kê những người đi công tác được sử dụng. Nguồn thứ hai liên quan đến sự kết hợp của các bộ dữ liệu hành chính khác nhau, loại đơn vị "người" và chứa các đặc điểm cơ bản. Nguồn thứ ba dựa trên dữ liệu Open Street Map, cụ thể hơn là mạng lưới đường bộ của Hà Lan. Đơn vị của nguồn này là "đoạn đường". Bao gồm vị trí và hình học của từng phân đoạn cũng như các kết nối với các phân đoạn khác. Nguồn cuối cùng là dữ liệu cảm biến vòng lặp giao thông chứa các quan sát về cường độ giao thông trên các đoạn đường mỗi phút.

Một số cách phân loại được chỉ định đã được đưa ra trong Bảng 1 dưới đây. Mỗi phân loại được đánh dấu bằng một số tương ứng.

Bảng 1. Áp dụng khung chất lượng vào nghiên cứu về di cư

	Dữ liệu điều tra thống kê	**Dữ liệu hành chính**	**Dữ liệu cơ sở hạ tầng**	**Dữ liệu cảm ứng**
Mức độ liên quan	Liên quan gián tiếp (1)	Liên quan gián tiếp	Liên quan gián tiếp	Liên quan trực tiếp
Độ bao phủ tổng thể	Không có phạm vi	Không có phạm vi (2)	Bao phủ hoàn toàn	Bao phủ ngầm
Tính đại diện của tổng thể	Chưa xác định	Chưa xác định	Xác xuất được chọn đã biết	Xác xuất được chọn đã biết
Hiệu lực của biến	Lỗi đo lường	Chưa xác định	Định nghĩa không nhất quán, lỗi mô hình hóa	Lỗi đo lường, định nghĩa không nhất quán
Sự ổn định của khái niệm	Không áp dụng (3)	Không áp dụng (3)	Phân tán (4)	Ổn định

	Dữ liệu điều tra thống kê	Dữ liệu hành chính	Dữ liệu cơ sở hạ tầng	Dữ liệu cảm ứng
Khả năng sửa chữa	Bổ sung - có thể sửa được	Bổ sung - có thể sửa được	Không cần thiết	Tự sửa lỗi (5)
Tính mới	Định kỳ	Định kỳ	Định kỳ	Dựa trên sự kiện
Thời gian xử lý	Tự động	Tự động	Tự động	Tự động
Khả năng tiếp cận	Toàn quyền truy cập	Toàn quyền truy cập	Toàn quyền truy cập	Toàn quyền truy cập
Siêu dữ liệu	Kết hợp, tuân thủ tiêu chuẩn	Kết hợp, tuân thủ tiêu chuẩn	Được xác định rõ	Được xác định rõ
So sánh	Có thể so sánh một phần	Có thể so sánh một phần	Hoàn toàn có thể so sánh được	Có thể so sánh hoàn toàn (6)

1. Mức độ liên quan: dữ liệu điều tra thống kê. Một trong những lý do để chỉ định chiều liên quan gián tiếp là do dữ liệu điều tra thống kê có loại đơn vị là "người", trong khi đối tượng mục tiêu của thống kê dự định là ở cấp độ các đoạn đường. Việc áp dụng công cụ lập kế hoạch tuyến đường trên dữ liệu cơ sở hạ tầng cho phép thu hẹp khoảng cách và sử dụng thông tin từ cuộc điều tra thống kê trên mạng lưới đường bộ.

2. *Phạm vi tổng thể: dữ liệu hành chính.* Đối tượng mục tiêu được xác định là các đoạn đường ở Hà Lan. Dữ liệu hành chính dựa trên con người, một loại đơn vị hoàn toàn khác.

3. *Tính ổn định của khái niệm: dữ liệu điều tra thống kê và dữ liệu hành chính*. Biến mục tiêu của số liệu thống kê dự định không có trong dữ liệu điều tra thống kê và hành chính. Do đó, khái niệm ổn định chiều không thể áp dụng được cho các nguồn này.

4. *Tính ổn định của khái niệm: dữ liệu cơ sở hạ tầng*. Cơ sở hạ tầng sẵn có cho việc đi lại có thể thay đổi theo thời gian, điều này có thể ảnh hưởng đến các tuyến đường được tính toán từ dữ liệu cơ sở hạ tầng. Việc sử dụng bộ dữ liệu tương ứng với thời gian của dữ liệu hành chính và dữ liệu cảm biến sẽ loại bỏ ảnh hưởng của sự phân tán.

5. *Khả năng sửa lỗi: tự sửa lỗi*. Dữ liệu theo phút được tổng hợp thành một giá trị trung bình duy nhất cho từng đoạn đường nhằm giống với giờ cao điểm buổi sáng (5 giờ sáng đến 9 giờ sáng). Để sửa lỗi đo lường, tổng số được tính trung bình cho tất cả các ngày làm việc thông thường trong cả tháng. Vì những hiệu chỉnh này được áp dụng mà không sử dụng dữ liệu đi kèm nên dữ liệu cảm biến được phân loại có thể tự sửa.

6. *Khả năng so sánh: dữ liệu cảm biến*. Dữ liệu cảm biến hoàn toàn có thể so sánh được vì các quốc gia khác cũng có thể có các dữ liệu được sử dụng với vai trò tương tự nếu dự án được thực hiện tại quốc gia đó.

Số liệu thống kê dự định là kết quả của cách tiếp cận như sau: (i) Đầu tiên, dữ liệu

điều tra thống kê được sử dụng để thử mô hình phương thức vận tải nhằm xác định xác suất của một phương thức nhất định, dựa trên các đặc điểm cơ bản của một người. (ii) Thứ hai, mô hình phương thức vận tải được áp dụng cho dữ liệu hành chính, sau đó được tổng hợp thành ma trận điểm đi – điểm đến (OD). Ma trận OD bao gồm các cặp khu dân cư và số người dự kiến sẽ đi làm bằng ô tô. Thứ ba, dự án phần mềm nguồn mở (Open Trip Planner) được sử dụng để chuyển đổi các cặp OD thành các tuyến đường, gồm các đoạn đường dẫn đến mật độ giao thông dự kiến cho từng đoạn đường. Về cơ bản, công cụ lập kế hoạch tuyến đường đóng vai trò là công cụ chuyển đổi giữa hai loại đơn vị (khu dân cư và đoạn đường). Cuối cùng, dữ liệu vòng lặp giao thông theo phút được lọc dựa trên chiều dài phương tiện để chỉ bao gồm các phương tiện ngắn như ô tô con và xe máy, đồng thời loại trừ các phương tiện dài hơn như xe tải, xe khách. Sau đó, dữ liệu được tổng hợp thành một giá trị cường độ cho mỗi cảm biến bằng cách lấy tổng tất cả các quan sát trong giờ cao điểm buổi sáng. Bằng cách này, hầu hết việc đi lại từ nhà đến nơi làm việc đều được tính đến. Đồng thời giảm thiểu việc đưa hoạt động đi lại để giải trí hoặc đi từ nơi làm việc về nhà vào, vì họ chủ yếu có xu hướng đi lại ngoài giờ cao điểm buổi sáng. Kết quả trung gian là một tập dữ liệu có hai biến cho một tập hợp các đoạn đường: cường độ dự kiến và cường độ quan sát được. Giả sử cường độ quan sát được là thực tế cơ bản, một mô hình đã được thử để hiệu chỉnh cường độ dự kiến gần với giá trị quan sát được hơn. Sau đó, mô hình này có thể được áp dụng cho tất cả các đoạn đường (ngay cả những đoạn đường không có cường độ quan sát được) để tạo ra giá trị hiệu chỉnh của cường độ dự kiến.

3.2. Nghiên cứu ứng dụng trong trường hợp: Phát hiện hạt virus trong dữ liệu nước thải

Trường hợp nghiên cứu thứ hai là một dự án của Viện Y tế công cộng và môi trường quốc gia Hà Lan (RIVM), nơi virus Corona được theo dõi trong dữ liệu nước thải và kết hợp với các nguồn dữ liệu hành chính để tạo ra số liệu thống kê về số lượng virus trên 100.000 người ở địa phương, khu vực. Trong nghiên cứu, biến số mục tiêu là số lượng virus trên 100.000 người mỗi ngày, tổng thể mục tiêu là tất cả cư dân của Hà Lan và cấp độ tổng hợp là khu vực hành chính.

Nguồn dữ liệu đầu tiên được sử dụng trong nghiên cứu là số liệu đo lường nước thải cho tất cả các hệ thống xử lý nước thải ở Hà Lan. Đơn vị của nguồn này là số lượng hệ thống lắp đặt nước thải. Mỗi hệ thống lắp đặt được lấy mẫu khoảng bốn lần một tuần, từ đó có được số liệu đo lường cho khu vực nước thải tương ứng. Số lượng hạt virus là biểu hiện của bệnh COVID-19 trong khu vực được lắp đặt. Đơn vị dữ liệu địa lý do cơ quan Thống kê Hà Lan công bố là các khu vực địa lý được sử dụng phổ biến như các tỉnh và thành phố. Tuy nhiên, ranh giới của các khu vực lắp đặt hệ thống xử lý nước thải không phải lúc nào cũng phù hợp với các khu vực của đô thị. Nguồn dữ liệu thứ hai là các khu vực xử lý nước thải, chứa thông tin địa lý về các khu vực có hệ thống xử lý nước thải. Những dữ liệu này được lấy từ cơ quan cấp nước địa phương và có thể có ngày tham chiếu khác nhau. Những thay đổi trong mạng lưới nước thải có thể không được đưa ngay vào dữ liệu. Nguồn dữ liệu thứ ba và cũng là

nguồn cuối cùng trong nghiên cứu do cơ quan Thống kê Hà Lan cung cấp là số lượng cư dân ở các khu vực dịch vụ kết hợp với các khu vực có trụ sở tại đô thị. Nguồn dữ liệu này chứa cả các đơn vị lắp đặt nước thải cũng như các đô thị. Có thể coi dữ liệu này là nguồn dựa trên sổ đăng ký. Bản thân nó là kết quả của việc kết hợp nhiều nguồn: dữ liệu cấp địa chỉ địa lý chi tiết (BAG), sổ đăng ký cá nhân (BRP) và các khu vực nước thải.

Một số cách phân loại được chỉ định đã được đưa ra trong Bảng 2 dưới đây. Mỗi phân loại được đánh dấu bằng một số tương ứng.

Bảng 2. Áp dụng khung chất lượng vào nghiên cứu về dữ liệu nước thải

	Dữ liệu điều tra thống kê	**Dữ liệu khu vực**	**Dữ liệu đăng ký**
Mức độ liên quan	Liên quan gián tiếp	Liên quan gián tiếp	Liên quan gián tiếp
Độ bao phủ tổng thể	Bao phủ hoàn toàn	Bao phủ hoàn toàn	Không có phạm vi bao phủ theo đơn vị
Tính đại diện của tổng thể	Xác xuất được chọn đã biết	Xác xuất được chọn đã biết	Chưa xác định
Hiệu lực của biến số	Định nghĩa không nhất quán, sai số đo lường (1)	Định nghĩa không nhất quán (2)	Chưa xác định
Sự ổn định của khái niệm	Phân tán (3)	Phân tán	Phân tán
Khả năng sửa chữa	Bổ sung-có thể sửa được	Không cần thiết	Không cần thiết
Tính mới	Dựa trên sự kiện	Định kỳ	Không thường xuyên
Thời gian xử lý	Tự động (1)	Tự động	Theo yêu cầu
Khả năng tiếp cận	Toàn quyền truy cập (4)	Truy cập hạn chế (4)	Truy cập hạn chế (4)
Siêu dữ liệu	Kết hợp	Kết hợp	Kết hợp, tuân thủ tiêu chuẩn
So sánh	Có thể so sánh một phần	Hoàn toàn có thể so sánh	Hoàn toàn có thể so sánh

1. *Hiệu lực biến số & Thời gian xử lý: dữ liệu đo lường*. Cần mất nhiều công sức để xử lý mẫu và tạo ra các phép đo. Việc làm thủ công có thể dẫn đến sai số đo lường. Quá trình này đã được chứng minh là nhanh chóng và đáng tin cậy và do đó được coi là tự động.

2. *Giá trị của biến: dữ liệu khu vực.* Dữ liệu khu vực được cung cấp bởi cơ quan quản lý nước địa phương và có thể có một số điểm không nhất quán nhỏ trong định nghĩa giữa các cơ quan chức năng khác nhau.

3. *Khái niệm ổn định: dữ liệu đo lường*. Phương pháp đo có thể cần cập nhật khi phát sinh các biến thể virus mới.

4. *Khả năng truy cập: dữ liệu đo lường, dữ liệu khu vực và dữ liệu đăng ký*. RIVM được chọn làm quan điểm để đánh giá khả năng tiếp cận.

Để đạt được số lượng virus trên 100.000 người mỗi ngày ở các khu vực địa phương, các nguồn được kết hợp theo cách sau. Số lượng dân cư trong dữ liệu đăng ký được sử dụng để tính quyền số cho từng tổ hợp khu

vực nước thải và đô thị. Tập dữ liệu này đã được tạo trong nhiều năm, với quyền số thay đổi mỗi năm và ngày đầu tiên của tháng một là ngày tham chiếu. Các quyền số được sử dụng để chuyển đổi dữ liệu đo lường trên mỗi lần cài đặt thành các phép đo trên mỗi khu vực dựa trên đô thị. Kết quả thống kê có thể được cập nhật cho mỗi phép đo mới, việc này diễn ra khoảng bốn lần một tuần trong thực tế.

4. Kết luận và thảo luận

Bài viết đề xuất một khung chất lượng có thể sử dụng khi kết hợp dữ liệu điều tra thống kê, dữ liệu hành chính và dữ liệu lớn cho thống kê chính thức. Khung bao gồm các chiều và danh mục tập trung vào chất lượng đầu vào của một nguồn riêng lẻ. Các danh mục được chọn tùy thuộc vào bối cảnh: số liệu thống kê dự định và dữ liệu đi kèm. Khung cho phép người dùng tìm ra điểm mạnh và điểm yếu của từng nguồn dữ liệu cũng như sự kết hợp của chúng. Nó giúp người sử dụng hiểu rõ hơn về cách kết hợp các nguồn dữ liệu. Khung chất lượng đã được áp dụng cho hai nghiên cứu trong đó cả ba loại nguồn dữ liệu được kết hợp. Việc áp dụng khung này cho từng nguồn riêng biệt và so sánh kết quả tổng thể cho phép có cái nhìn tổng quan về những thách thức gặp phải trong mỗi nghiên cứu.

Khung chất lượng được thiết kế cho các trường hợp có một hoặc nhiều biến mục tiêu trong đó quá trình đạt được kết quả dự định là (gần) giống nhau đối với tất cả các biến mục tiêu. Trong trường hợp có nhiều biến mục tiêu được dự định và mỗi biến yêu cầu một cách tiếp cận riêng biệt về kết hợp các nguồn, thì khung này phải được áp dụng riêng cho từng biến mục tiêu.

Chất lượng của kết quả thống kê cuối cùng không chỉ phụ thuộc vào chất lượng của các nguồn dữ liệu đầu vào mà còn phụ thuộc vào sự lựa chọn về quy trình và mô hình được sử dụng để kết hợp các nguồn này. Sau khi đánh giá các nguồn dữ liệu với khung, công việc thiết kế là cần thiết để lựa chọn ra phương pháp phù hợp nhất. Ngoài ra, những đánh giá tích cực về một nguồn theo khung không nên được coi là sự đảm bảo rằng các lần lặp lại nguồn trong tương lai cũng phù hợp với mục đích tạo ra số liệu thống kê dự kiến. Mọi thay đổi về nguồn dữ liệu hoặc bối cảnh phải được đánh giá lại bằng khung. Khung xem xét một cách có hệ thống các vấn đề chất lượng tiềm ẩn khi kết hợp các nguồn. Tuy nhiên, nó không đề xuất giải pháp cho những vấn đề chất lượng này.

Các nghiên cứu đã minh họa một số điểm chính, đặc biệt là trong các bước lập mô hình để có được số liệu thống kê dự kiến. Trong nghiên cứu về tính di động, nguồn dữ liệu mạng đóng vai trò là cầu nối giữa hai nhóm. Mặc dù các tổng thể khác nhau ở một số nguồn, chúng vẫn có thể được kết hợp. Trường hợp tương tự cũng được quan sát trong nghiên cứu về nước thải, trong đó quyền số chuyển đổi giữa hai định nghĩa khu vực không giống nhau là chìa khóa để kết hợp dữ liệu đo lường và dữ liệu hành chính. Việc kết hợp phân loại cho từng nguồn giúp xác định trước các điểm để tích hợp dữ liệu và giúp xác định các trường hợp tương tự, từ đó cho phép sử dụng lại giải pháp.

Phạm Hạnh (dịch)

Nguồn: Quality framework for combining survey, administrative and big data for official statistics