ĐÁNH GIÁ CHẤT LƯỢNG DỮ LIỆU HÀNH CHÍNH KHI SỬ DỤNG TRONG CUỘC TỔNG ĐIỀU TRA DÂN SỐ: SỬ DỤNG KẾT QUẢ TRONG CUỘC TỔNG ĐIỀU TRA THỬ Ở INDONESIA

Tóm tắt:

Cơ quan Thống kê Indonesia dự kiến thực hiện cuộc tổng điều tra dân số lần thứ 7 vào năm 2020. Đối lập với phương pháp truyền thống đã sử dụng trong các cuộc tổng điều tra trước, cuộc tổng điều tra lần này sẽ sử dụng dữ liệu hành chính đã thu thập được từ Bộ Nội vụ kết hợp với điều tra thực địa. Các thông tin cá nhân đã được lưu lại trong kho dữ liệu hành chính sẽ được sử dụng như nguồn dữ liệu gốc để điều tra thực địa đồng thời nhằm đảm bảo đủ độ bao phủ. Những thay đổi này có thể là bước tiến đầu tiên để thực hiện một cuộc tổng điều tra dựa trên dữ liệu đăng ký sau này. Với vai trò quan trọng của dữ liệu hành chính trong cuộc tổng điều tra sắp tới, chúng ta cần phải đánh giá chất lượng dữ liệu hành chính đã được Bộ Nội vụ lưu trữ. Do đó, nghiên cứu này nhằm mục đích phân tích liệu có những khác biệt lớn trong việc chọn các biến cá nhân giữa dữ liệu lấy từ nguồn dữ liệu hành chính và từ kết quả cuộc điều tra thử. Cuộc điều tra thử đã được thực hiện ở 3 ngôi làng được chọn, sử dụng hình thức thu thập dữ liệu theo nhiều cách khác nhau. Việc so sánh các biến cá nhân đã chọn và các kết quả của cuộc điều tra thử cho thấy rằng dữ liệu hành chính đảm bảo chất lượng để sử dụng như nguồn dữ liệu trong cuộc tổng điều tra áp dụng phương pháp kết hợp ở Indonesia.

1. Giới thiệu

Với vai trò là một phần trong hệ thống thống kê quốc gia, Cơ quan Liên hợp quốc đề xuất tất cả các quốc gia trên thế giới thực hiện cuộc tổng điều tra dân số và nhà ở ít nhất 1 lần trong chu kỳ 10 năm. [1] Ngoài ra, các cuộc tổng điều tra dân số và nhà ở ở Indonesia đều phải tuân thủ theo luật thống kê quốc gia. Indonesia, thông qua Cơ quan Thống kê Indonesia – đơn vị chịu trách nhiệm thống kê chính thức đã thực hiện 6 cuộc tổng điều tra dân số và nhà ở từ năm 1961. Tính đến tận cuộc tổng điều tra trước vào năm 2010, công tác thu thập dữ liệu vẫn được thực hiện bằng phương pháp truyền thống. Bằng cách liệt kê tất cả các trường dữ liệu, thông tin về các cá nhân và hộ gia đình đã được thu thập bằng cách sử dụng bảng hỏi, các thống kê viên đến phỏng vấn tận nhà. Phương pháp này được coi là phương pháp rất phức tạp và đòi hỏi có các nguồn lực lớn để thực hiện.

Sau khi cân nhắc những hạn chế của cuộc tổng điều tra sử dụng phương pháp truyền thống, Cơ quan thống kê Indonesia đã bắt đầu tìm kiếm các giải pháp thay thế để thực hiện trong các cuộc tổng điều tra trong tương lai. Hy vọng rằng, trong tương lai, Indonesia có thể thực hiện cuộc tổng điều tra sử dụng dữ liệu đăng ký. Các cuộc tổng điều tra sử dụng dữ liệu đăng ký được xem là rẻ hơn, nhanh hơn và có thể giảm gánh nặng cho đối tượng được điều tra. [2] Phương pháp tổng điều tra sử dụng dữ liệu đăng ký được thực hiện bởi nhiều quốc gia như: Hà Lan, Đan Mạch, Phần Lan, Na Uy, Thụy Điển, Áo, Slovenia, Triều Tiên và Singapore (2-4). Trước khi có thể thực hiện một cuộc tổng điều tra sử dụng dữ liệu đăng ký, cơ quan Thống kê Indonesia sẽ sử dụng phương

pháp kết hợp trong cuộc tổng điều tra sắp tới và tiến hành một số thay đổi trong quá trình thực hiện. Phương pháp này đã được sử dụng bởi Cơ quan Thống kê Hà Lan từ năm 1981 bằng cách sử dụng kết hợp dữ liệu đăng ký và dữ liệu điều tra. [5] Trước khi tiến tới thực hiện cuộc tổng điều tra sử dụng dữ liệu đăng ký vào năm 2021, Tây Ban Nha cũng đã thực hiện cuộc tổng điều tra sử dụng phương pháp kết hợp vào năm 2011. [6] Trong cuộc tổng điều tra năm 2020, Cơ quan thống kê Indonesia sẽ sử dụng dữ liệu hành chính kết hợp với điều tra thực địa. Công tác điều tra sẽ được thực hiện bằng máy tính bảng (CAPI) và sử dụng bảng hỏi trực tuyến (CAWI) bên cạnh bảng hỏi giấy.

Hiện nay, Indonesia đã có 1 hệ thống cơ sở dữ liệu hành chính dân số do Bộ Nội vụ quản lý. Mỗi cư dân Indonesia đã đăng ký trong hệ thống này đều có một số định danh nhất định. Mặc dù không phải tất cả cư dân Indonesia đều được đăng ký trong hệ thống này nhưng ước tính khoảng 97% tổng dân số nước này đã được lưu trữ thông tin trong hệ thống. Hệ thống dữ liệu hành chính Indonesia sử dụng công nghệ sinh trắc học. Hệ thống định danh số sử dụng công nghệ sinh trắc học nhằm đảm bảo rằng mỗi người chỉ có 1 số định danh duy nhất. [8] Bởi vì việc sử dụng số định danh đang trở thành yêu cầu bắt buộc để truy cập hầu hết tất cả dịch vụ và tài liệu hành chính nên người ta cho rằng sẽ có càng nhiều người dân chủ động yêu cầu để được đăng ký vào hệ thống. Tuy nhiên, mặc dù dữ liệu hành chính luôn sẵn có và phủ rộng hầu hết tất cả dân số nhưng những thay đổi về dữ liệu cá nhân không được cập nhật thường xuyên trong hệ thống. Do đó, có thể có một số dữ liệu cá nhân đã được lưu trữ trong hệ thống nhưng vẫn chưa được cập nhật kịp thời.

Vì những hạn chế như vậy trong dữ liệu hành chính mà cuộc tổng điều tra dân số sắp tới tại Indonesia không thể thực hiện hoàn toàn dựa trên các dữ liệu đăng ký. Tuy nhiên, mặc dù có những hạn chế nêu trên nhưng Cơ quan Thống kê Indonesia có thể vẫn sử dụng nguồn dữ liệu hành chính trong cuộc tổng điều tra năm 2020 như Khung cơ bản để đảm bảo đủ độ bao phủ. Có thể mượn một số thông tin cá nhân từ dữ liệu hành chính để có được các biến dân số xác định. Điều tra thực địa vẫn rất cần thiết để hoàn thiện thông tin cá nhân và hộ gia đình – những thông tin mà không liên quan hoặc không sẵn có trong dữ liệu hành chính. Do đó, cuộc tổng điều tra sử dụng phương pháp kết hợp này sẽ là điểm khởi đầu để thực hiện 1 cuộc tổng điều tra dựa trên dữ liệu đăng ký trong tương lai.

Mặc dù một số cuộc điều tra thử đã được thực hiện vào năm 2018 và 2019 nhưng mục đích của cuộc điều tra thử năm 2018 bị giới hạn bởi các đặc điểm khác nhau của các phương pháp truyền thống hơn là sử dụng dữ liệu hành chính. Do đó, để thử nghiệm cơ chế hoạt động của cuộc tổng điều tra và việc sử dụng dữ liệu hành chính trong cuộc tổng điều tra, Cơ quan thống kê Indonesia đã thực hiện 1 cuộc điều tra thử sử dụng phương pháp kết hợp vào tháng 2 năm 2019. Cuộc điều tra thử sử dụng phương pháp kết hợp này rất quan trọng bởi vì nó tạo nên nền tảng đầu tiên trong việc thay đổi phương pháp thực hiện tổng điều tra từ phương pháp truyền thống sang phương pháp kết hợp. Cuộc thử nghiệm đã được thực hiện tại 3 ngôi làng ở 2 tỉnh; 1 ngôi làng nằm ở phía Tây tỉnh Java (Sukamiskin), và 2 làng còn lại ở phía Nam Kalimantan (Mekarsari và Indah Sari). Các địa điểm này đã được chọn từ các địa điểm trong cuộc điều tra thử lần trước vì vậy các kết quả của cuộc điều tra thử sử dụng phương pháp kết hợp có thể được so sánh với các kết

quả từ các cuộc điều tra thử trước đó cũng sử dụng các phương pháp truyền thống.

Đánh giá cơ sở dữ liệu hành chính là cần thiết để kiểm tra các biến tồn tại ở cả nguồn dữ liệu hành chính và dữ liệu thu thập được từ cuộc điều tra thử. Kinh nghiệm của Ba Lan và bản hướng dẫn của UNECE cho thấy rằng để khai thác sử dụng cơ sở dữ liệu hành chính trong cuộc tổng điều tra, chất lượng dữ liệu đăng ký là yếu tố quan trọng nhất cần được xem xét. [9] Một cách để đo lường chất lượng của dữ liệu hành chính là so sánh dữ liệu hành chính với dữ liệu cá nhân thu thập được từ 1 cuộc điều tra hay cuộc tổng điều tra. Để chuẩn bị cho cuộc Tổng điều tra năm 2021, Cơ quan Thống kê Bồ Đào Nha cũng đã đo lường chất lượng của dữ liệu hành chính sẵn có bằng cách so sánh dữ liệu này với dữ liệu vi mô của cuộc tổng điều tra. [10] Một phương pháp tương tự cũng đã được sử dụng để đánh giá chất lượng của dữ liệu hành chính sẵn có ở Indonesia với một số khía cạnh đã được điều chỉnh tùy theo bối cảnh của Indonesia.

Nghiên cứu này nhằm phân tích liệu có sự khác biệt đáng kể trong các biến cá nhân được chọn giữa dữ liệu từ nguồn dữ liệu hành chính và từ cuộc điều tra dân số thử nghiệm hay không. Các kết quả này cũng có thể gợi ý một số khía cạnh cần được cải tiến cả trong cơ chế điều tra hoặc trong dữ diệu hành chính để hướng tới các cuộc tổng điều tra trong tương lai.

2. Phương pháp

2.1 Nguồn dữ liệu hành chính

Dữ liệu hành chính được sử dụng làm cơ sở cho cuộc tổng điều tra dân số sử dụng phương pháp kết hợp của Indonesia có thể lấy được từ một nguồn dữ liệu duy nhất. Dữ liệu do Tổng cục Dân số và Đăng ký Hộ tịch, Bộ Nội vụ quản lý. Căn cứ vào Luật Quản lý Dân số số 23/2006, dữ liệu dân số và hồ sơ dân sự ở Indonesia được quản lý bằng cách áp dụng hệ thống thông tin hành chính dân số có tên là SIAK. Việc sử dụng SIAK cho phép thu thập nhiều dữ liệu dân cư hơn và cho phép cung cấp 1 cách tự động và lâu dài số định danh duy nhất cho mỗi cư dân Indonesia, bao gồm cả công dân nước ngoài có giấy phép thường trú. Do đó, hệ thống có thể loại bỏ quyền sở hữu nhiều số định danh. Việc lưu trữ dữ liệu cá nhân được thực hiện thông qua SIAK bởi đơn vị chính quyền cấp thấp nhất tại các địa phương ở Indonesia. Dữ liệu sau đó được lưu trữ trong cơ sở dữ liệu tổng hợp quốc gia thuộc Bộ Nội vụ. Để đảm bảo rằng dữ liệu được tổng hợp từ tất cả các vùng ở Indonesia được làm sạch và không có dữ liệu thừa, công tác xác thực hoặc đối chiếu dữ liệu được thực hiện định kỳ mỗi kỳ một lần hoặc hai lần một năm.

Dữ liệu sử dụng trong cuộc thống kê dân số thử nghiệm là kết quả của việc đối chiếu dữ liệu trong kỳ hai của năm 2018. Do đó, ngoài thực tế là những thay đổi trong dữ liệu cá nhân không phải lúc nào cũng được báo cáo ngay lập tức cho các cấp chính quyền, một số biến riêng lẻ có thể không còn hợp lệ tại thời điểm thực hiện điều tra dân số vào tháng 2 năm 2019. Ví dụ, tình trạng hôn nhân hoặc trình độ học vấn cao nhất của một người có thể đã thay đổi sau lần báo cáo cuối cùng trong hệ thống.

2.2 Truyền dữ liệu

Cơ sở pháp lý cho phép cơ quan thống kê Indonesia truy cập cơ sở dữ liệu hành chính đã được quy định trong Luật Thống kê số 16/1997. Tuy nhiên, ở các cấp cơ sở, các quy định kỹ thuật là cần thiết để có thêm cơ chế thực hiện chi tiết về truyền dữ liệu và sử dụng dữ liệu. Tại thời điểm điều tra dân số, quy trình đăng ký sao chép đầy đủ các hồ sơ hành chính

từ Bộ Nội vụ đến Cơ quan Thống kê Indonesia (BPS) chưa được hoàn thành. Do đó, vì mục tiêu của cuộc thống kê dân số thử nghiệm và dựa trên biên bản thỏa thuận giữa hai đơn vị, BPS đã được truy cập dữ liệu dân cư đã được đăng ký đến cấp thấp nhất của chính quyền địa phương ví dụ như Kota Bandung và Barito Kuala. Dữ liệu hành chính thu được từ cả hai lĩnh vực của cuộc điều tra thử sau đó được sao chép từ máy chủ của Bộ Nội vụ đến máy chủ của BPS. Quá trình này đã sử dụng Giao diện lập trình ứng dụng (API) ở dạng một dịch vụ web được thiết kế để hỗ trợ khả năng tương tác giữa máy với máy qua mạng. Dữ liệu thu được liên quan đến kết quả được đối chiếu từ kỳ II năm 2018.

2.3 Làm sạch dữ liệu

Sau khi dữ liệu được sao chép sang máy chủ của BPS, các biến dữ liệu được làm sạch vì vậy chúng có thể được sử dụng cho quy trình điều tra dân số tiếp theo. Các biến thu được từ dữ liệu gồm: tỉnh, quận, huyện, làng, địa chỉ, số thẻ gia đình, số định danh, tên, giới tính, ngày sinh, nơi sinh, quan hệ với chủ hộ, tên mẹ, trình độ giáo dục, tình trạng hôn nhân, tôn giáo và nghề nghiệp.

Quy trình làm sạch dữ liệu gồm những bước sau:

1. Kiểm tra và khớp các đơn vị hành chính

Dữ liệu từ Bộ Nội vụ có vấn đề do mã và tên đơn vị hành chính không chuẩn, đặc biệt là đối với cấp vùng cụ thể hơn. Do đó, việc kiểm tra và hiệu chỉnh đã được thực hiện để đảm bảo sự liên kết chặt chẽ của từng dữ liệu cá nhân. Việc kiểm tra được thực hiện trên các biến tỉnh, thành phố/ quận, huyện và làng. Kết quả cho thấy rằng có một số dữ liệu đơn vị hành chính không khớp với mã phân chia có sẵn ở BPS, đặc biệt là ở cấp làng xã. Sự khác biệt là do thông tin về dữ liệu hành chính ở các làng không được cập nhật cũng như đặt tên không chuẩn với những quy chuẩn cho làng. Hơn nữa, các chỉnh sửa đã được thực hiện đối với các trường hợp không khớp thông tin sau khi xác nhận dữ liệu vào Bộ Nội vụ.

2. Kiểm tra và mã hóa dữ liệu

Dữ liệu thu được từ Bộ Nội vụ vẫn ở dạng ký tự cho từng biến. Nhìn chung, việc hoàn tất các nội dung truy nhập thông tin là tốt nhưng có một số truy nhập không tương thích có thể làm phức tạp quá trình mã hóa. Ví dụ, trong biến Giáo dục, có một số mục nhập trong cơ sở dữ liệu dẫn đến có cùng trình độ học vấn. Do đó, cần thiết phải thiết lập một từ điển dữ liệu cho tất cả các thuật ngữ khác nhau trong cơ sở dữ liệu để mã hóa quy trình có thể được thực hiện đúng cách.

3. Thiết lập dữ liệu tổng thể của đơn vị hành chính địa phương (SLS Master)

Hình thành SLS master về cơ bản là nhóm các dữ liệu dân số dựa trên thông tin của các đơn vị hành chính địa phương có trong các biến địa chỉ cá nhân. Điều này rất quan trọng, vì mỗi cá nhân dự kiến sẽ được xác định trong một đơn vị hành chính địa phương nhỏ nhất (SLS). Tuy nhiên, Bộ Nội vụ không có SLS bởi vì việc thành lập SLS nhỏ nhất thuộc thẩm quyền của chính quyền tự trị. Ngoài ra, không có tiêu chuẩn quốc gia về sự hình thành SLS. Do đó, Việc thành lập SLS tổng thể phải được thực hiện đầy đủ để nó có thể bao gồm tất cả SLS ở tất cả các tỉnh/ thành phố tự trị.

4. Chọn lựa các ghi chép cá nhân

Cuộc điều tra thử gồm 2 giai đoạn. Giai đoạn thứ nhất là giai đoạn xác minh để xác định độ phủ. Điều này có thể được thực hiện bằng cách người đứng đầu đơn vị hành chính địa phương nhỏ nhất xác minh từng cá nhân

trong danh mục đơn vị hành chính địa phương nhỏ nhất (SLS). Điều tra viên đến gặp người đứng đầu SLS để xác định liệu mỗi cá nhân vẫn đang sinh sống tại địa phương đó hay không. Sau đó, có 26292 cá nhân được xác định là cư dân cư trú tại 3 ngôi làng trong tổng số 27169 người được liệt kê trong cơ sở dữ liệu của 3 ngôi làng này. Những danh sách cá nhân này kết hợp những người đã có trong danh sách và những người không được ghi tên là cư dân trong những ngôi làng này nhưng được coi là cư dân thường trú ở những khu vực đó. Bảng số 1 cho thấy số cá nhân ở 3 làng trong dữ liệu hành chính và sau giai đoạn xác minh.

Bảng 1: Số người phân theo địa điểm, dữ liệu hành chính và giai đoạn xác minh trong cuộc điều tra dân số thử nghiệm

STT	Tên làng	Tỉnh/ Thành phố tự trị	Số hồ sơ	
			Dữ liệu hành chính	Giai đoạn xác minh trong cuộc điều tra thử
1	Indahsari	Barito Kuala	1.585	1.466
2	Mekarsari	Barito Kuala	3.306	3.077
3	Sukamiskin	Bandung City	22.397	21.630
	Tổng		**27,169**	**26,292**

Bảng 2: Số lượng và tỷ lệ phần trăm người trả lời phù hợp với hồ sơ dữ liệu hành chính dựa trên phương thức điều tra

STT	Loại	Số người trả lời trong giai đoạn điều tra	Bản ghi khớp với dữ liệu hành chính	
			Người trả lời	%
1	CAPI	1,801	997	55%
2	CAWI	886	671	76%
	Tổng	**2,693**	**1,668**	**62%**

Sau giai đoạn xác minh, công tác điều tra đã thực hiện được 3 phần: phỏng vấn của điều tra viên sử dụng máy tính bảng (CAPI); tự điền thông tin thông qua website (CAWI); và tự điền thông tin thông qua bảng hỏi giấy (DOPU). Vì giai đoạn này không được thiết kế để hoàn thành việc điều tra dân số toàn diện tại các khu vực điều tra thử, do đó, chỉ có 2793 người trả lời (10,62% dân số được xác định ở 3 làng) đã hoàn thành tất cả 3 phần trên.

Trong cuộc điều tra thử, không có quy trình xử lý dữ liệu đối với bảng hỏi giấy hoặc từ DOPU bởi vì tại thời điểm đó, hệ thống vẫn đang trong quá trình chuẩn bị. Các tài liệu sẽ được sử dụng để kiểm tra hệ thống xử lý dữ liệu bằng cách sử dụng chức năng chụp lại để nhận dạng sau này. Do đó, đã có 100 người trả lời từ DOPU đã được loại trừ khỏi việc đối sánh với dữ liệu hành chính.

Có thể đối chiếu những người trả lời khác thu được từ CAPI và CAWI với hồ sơ hành chính. Quá trình đối chiếu được thực hiện bằng cách sử dụng số định danh duy nhất, được gọi là NIK, làm chìa khóa để đối chiếu. Kết quả là, tỷ lệ thành công là 62% (xem Bảng 2). Những trường hợp thông tin không khớp nhau có thể do 2 khả năng: Một là, một số cá nhân có thể đã được lưu trữ thông tin trong dữ liệu hành chính nhưng đã được xác định là cư dân tại các khu tự trị khác nằm ngoài Kota Bandung hoặc Barito Kuala. Điều này là do quy định sao chép dữ liệu hành chính ở cấp quốc gia vẫn chưa hoàn thiện tại thời điểm thực hiện cuộc điều tra thử; kết quả là, cơ sở dữ liệu chỉ chứa các hồ

sơ dữ liệu từ hai tỉnh trên. Hai là, một số cá nhân có thể mới sinh hoặc thậm chí là người lớn nhưng chưa bao giờ được lưu trữ trong cơ sở dữ liệu hành chính. Do đó, căn cứ mục đích của cuộc nghiên cứu này, tổng số đã có 1668 bộ dữ liệu đã được đối chiếu khớp với nhau có thể được phân tích thêm.

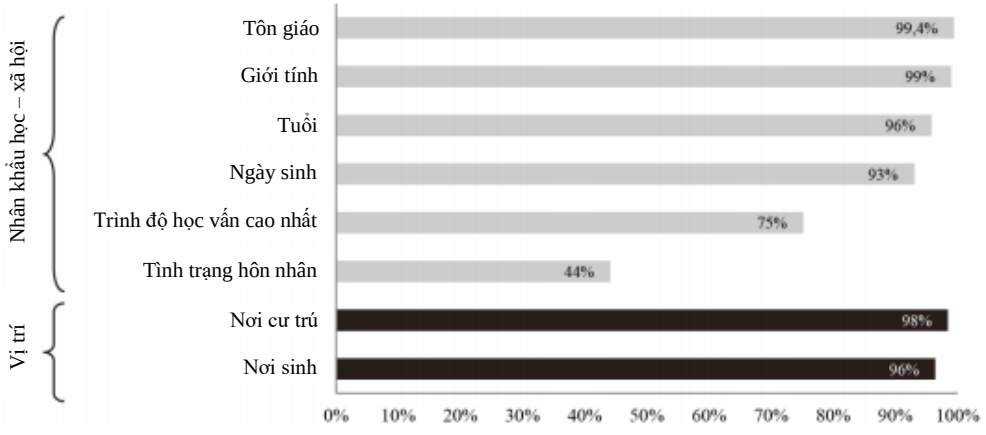
2.4 Phân tích thống kê

Nghiên cứu này nhằm mục đích xác định liệu có sự khác biệt giữa dữ liệu của các biến được chọn lấy từ kết quả cuộc điều tra thử và dữ liệu hành chính đối với cùng một cá nhân. Các biến có thể được kiểm tra từ 2 nguồn dữ liệu là: giới tính, tôn giáo, tình trạng hôn nhân, ngày sinh, nơi cư trú, nơi sinh, trình độ giáo dục, tuổi. Sáu biến này được chọn vì chúng tồn tại ở trong cả hai nguồn dữ liệu.

Nghiên cứu đã sử dụng các phương pháp phân tích thống kê mô tả và phi tham số để xác định xem liệu có sự khác biệt trong tám biến số thu được từ dữ liệu hành chính và từ cuộc điều tra dân số thử nghiệm. Phi tham số phân tích đã được sử dụng vì các giả định về các tham số phân phối không tương thích với các biến. Phân tích thống kê các cặp mẫu được sử dụng để so sánh thông tin rút ra từ hồ sơ hành chính và dữ liệu cuộc điều tra dân số thử cho cùng một cá nhân đã được xác định bởi từ khóa NIK. Việc xử lý sự khác biệt đối với thử nghiệm cặp có thể được xem như là sự khác biệt của các nguồn dữ liệu. Giả thiết rằng dữ liệu thử nghiệm điều tra là cập nhật hơn dữ liệu từ nguồn hành chính.

Bảng 3: Biến, loại dữ liệu và kiểm tra thống kê

STT	Biến	Loại dữ liệu	Kiểm định thống kê
1	Giới tính	Định danh với 2 danh mục	Kiểm định McNemar
2	Tôn giáo, tình trạng hôn nhân, nơi cư trú, nơi sinh	Định danh với nhiều hơn 2 danh mục	Kiểm định Marginal Homogeneity
3	Trình độ học vấn cao nhất, độ tuổi	Thứ tự, khoảng thời gian	Kiểm định Wilcoxon



Hình 1: Tỉ lệ cân bằng của 8 biến số được chọn từ cuộc điều tra thử và từ hồ sơ dữ liệu hành chính

Giải thiết ngây ngô của thử nghiệm thống kê cho rằng hai xác suất cận biên của mỗi đầu ra là giống nhau đối với các cặp mẫu từ nguồn hành chính và thử nghiệm điều tra. Bởi vì có các kiểu dữ liệu khác nhau của 8 biến số, các thử nghiệm thống kê sử dụng mẫu cặp đôi được điều chỉnh theo kiểu dữ liệu của mỗi biến (xem Bảng 3).

3. Các kết quả

3.1 Các thống kê mô tả

Tương ứng với mục đích của nghiên cứu này, tám biến được chọn từ thử nghiệm điều tra và hồ sơ hành chính được so sánh với nhau: giới tính, tôn giáo, tình trạng hôn nhân, ngày sinh, nơi cư trú, nơi sinh, bằng cấp giáo dục đạt được và tuổi. Tỷ lệ bình đẳng giữa thử nghiệm điều tra và hồ sơ dữ liệu hành chính đối với mỗi biến được chỉ ra trong Hình 1. Các kết quả so sánh chỉ ra rằng bốn biến số nhân khẩu học – xã hội như tôn giáo, giới tính, tuổi, ngày sinh có tỷ lệ thống nhất cao. Hơn 90% người được phỏng vấn theo cặp có thông tin trùng khớp đối với các biến số này. Kết quả chỉ rằng hồ sơ hành chính của bốn biến số nhân khẩu học – xã hội có độ chính xác cao. Bằng cấp giáo dục cao nhất có độ chính xác thấp hơn với khoảng 75%. Mặt khác, tình trạng hôn nhân là biến với tỷ lệ trùng khớp thấp nhất với khoảng 44%. Các kết quả này chỉ ra rằng thông tin cá nhân về tình trạng hôn nhân và bằng cấp giáo dục không được cập nhật trong hệ thống quản lý hành chính. Thông tin liên quan đến nơi ở và nơi sinh ở cấp độ địa phương (tỉnh) (mức độ cao nhất của chính quyền địa phương ở Indonesia) cũng có tỷ lệ trùng khớp cao. Khoảng 98% và 96% của các cá nhân cặp đôi có cùng thông tin về tỉnh sinh sống và tỉnh nơi sinh ra tương ứng.

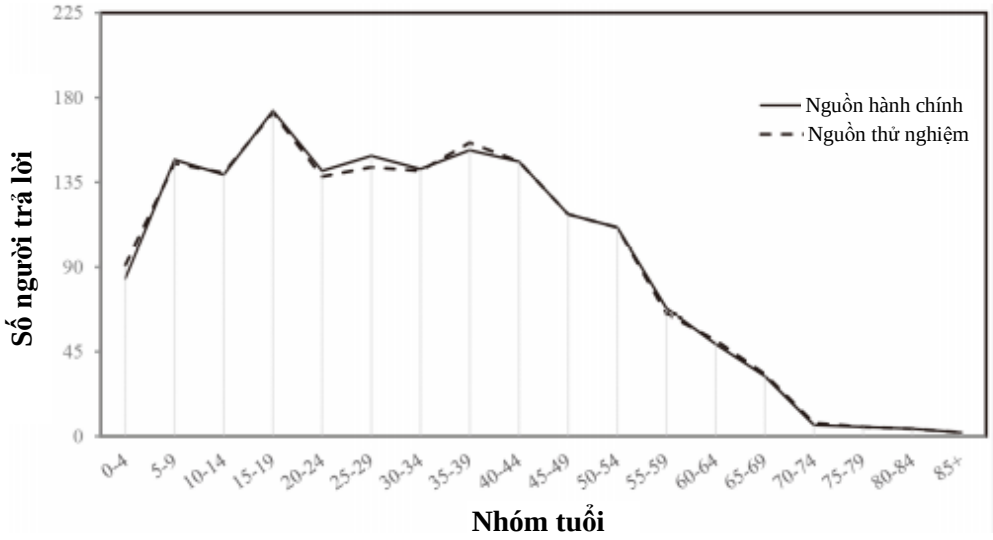
Chất lượng của dữ liệu hành chính có thể được phân tích bằng cách so sánh tần số của biến Tuổi giữa nguồn dữ liệu hành chính và nguồn dữ liệu điều tra thử nghiệm. Tuổi và giới tính là các đặc điểm cá nhân cơ bản của dân số loài người, chúng có thể ảnh hưởng tới tất cả các khía cạnh của quá tình kinh tế - xã hội và địa lý. Bởi vậy, việc so sánh tần số tuổi giữa hai nguồn dữ liệu nên được so sánh bởi giới tính. Các đường biểu đồ chồng lấn chỉ ra các thông tin rất giống nhau về biến tuổi giữa hai nguồn dữ liệu.

Hình 2 chỉ ra rằng hai nguồn dữ liệu, được trình bày bằng hai đường con, là rất gần nhau về mức độ và xu hướng. Những sự khác biệt nhẹ giữa hai nguồn dữ liệu được thấy trong nhóm tuổi từ 25-29 và 35-39. Khi dữ liệu được tổng hợp lại bởi giới tính (xem Hình 3), nó có thể được xem là dữ liệu của nam giới từ cả hai nguồn có sự tương đồng hơn ở các nhóm tuổi.

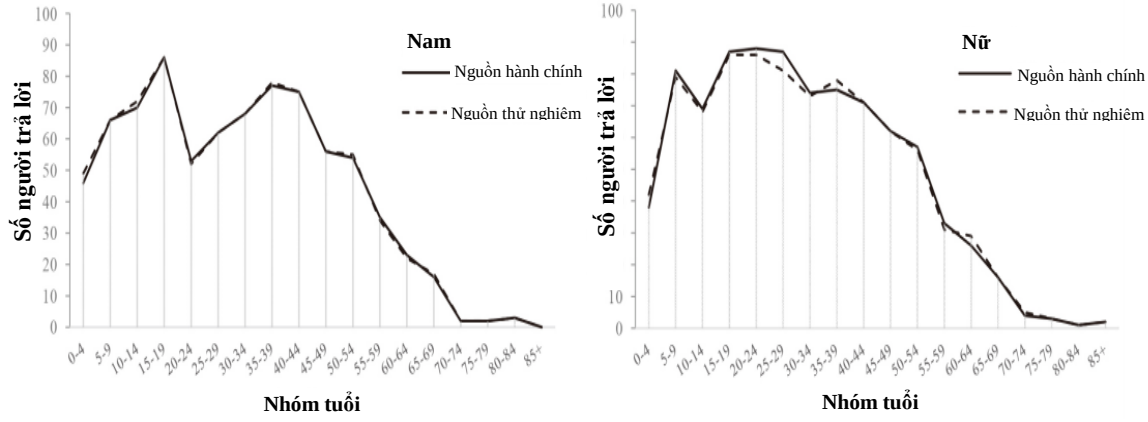
Mặt khác, dữ liệu về phụ nữ có sự khác biệt nhẹ tại một vài nhóm tuổi. Dữ liệu hành chính có nhiều nữ giới hơn trong nhóm tuổi từ 25-29, trong khi đó có ít nữ giới hơn trong nhóm tuổi từ 35-39 và 60-64 so với dữ liệu từ cuộc điều tra thử nghiệm.

Để tìm ra nguyên nhân của sự khác biệt, thực hiện việc kiểm tra thấu đáo hơn đối với các cá nhân có độ tuổi khác nhau trong cuộc điều tra thử nghiệm và dữ liệu hành chính. Kiểm tra tính hợp lý cũng đã được thực hiện giữa các biến số về Tuổi và các biến số cá nhân khác, ví dụ mối quan hệ với người đứng đầu gia đình, tình trạng hôn nhân, bằng cấp cao nhất đạt được. Từ việc kiểm tra đã cho ra kết quả là đã có các lỗi của dữ liệu đầu vào bởi điều tra viên đối với biến số ngày sinh hoặc tuổi từ phương thức CAPI. Hơn nữa, lỗi trong điền biến số Giới tính cũng đóng góp vào sự khác biệt này, đặc biệt đối với nữ giới. Ngoài ra, bằng cách nhìn vào sự khác biệt ở đỉnh trong nhóm tuổi từ 25-29, sự

khác biệt chỉ là 6 trong khoảng 90 người. Thật khó để nói rằng sự khác biệt là lớn bởi vì mẫu là nhỏ đối với nhóm tuổi này.



Hình 2: Tần số nhóm tuổi dựa trên nguồn dữ liệu



Hình 3: Tần số nhóm tuổi dựa trên giới tính và nguồn dữ liệu

Sự xem xét sâu hơn được thực hiện để đảm bảo hiệu chỉnh dữ liệu biến số Giới tính. Mặc dù giới tính thay đổi thông qua phẫu thuật chuyển giới ngày nay là có thể thực hiện được, ví dụ trường hợp Indonesia 1% khác biệt thông tin về Giới tính vẫn được xem như đáng ngờ. Bởi vậy, việc kiểm tra dữ liệu sâu hơn về tên người được hỏi được thực hiện. Dựa trên cảm nhận chung và văn hoá Indonesia, có những cái tên có thể xác định cho giới tính cụ thể (xem Bảng 4). Các kết quả kiểm tra cho biết hầu hết sự bất nhất trong biến số Giới tính có kết quả từ các điều tra thử nghiệm hơn là những ghi chép về hành chính.

Sự kiểm tra cũng được thực hiện đối với sự bất nhất của biến số Bằng cấp cao nhất đạt được trong cả hai nguồn. Khoảng 4% người được điều tra có mức độ bằng cấp giáo dục đạt được thấp hơn các kết quả từ các cuộc điều tra thử nghiệm (xem Bảng 5). Giả thiết ban đầu

trong nghiên cứu này là dữ liệu từ cuộc điều tra thử nghiệm được cập nhật hơn so với dữ liệu hành chính. Bởi vậy, sự không thống nhất là một điều bất thường và thể hiện lỗi trong quá trình thử nghiệm thống kê.

Cả hai sự kiểm tra đối với Giới tính và Bằng cấp giáo dục đạt được cho biết rằng sự không thống nhất trong dữ liệu có thể tồn tại bởi vì lỗi không lấy mẫu. Nó có thể được gây ra bởi lỗi của con người hoặc lỗi hệ thống. Hơn nữa, thiết kế giao diện ứng dụng không phù hợp, đối với cả CAPI và CAWI, có thể dẫn tới các lỗi.

Bảng 4: Xác định giới tính dựa trên tên của cá nhân

STT	Tên	Hồ sơ hành chính	Kết quả kiểm tra điều tra dân số		Kết quả xác định
			CAPI	CAWI	
1	Meita xxxxxxxxx	Nữ		Nam	Nữ
2	Ahmadi xxxx xxxxxxx	Nữ		Nam	Nam
3	Nawafiansyah xxxxx xxxxx	Nam	Nữ		Nam
4	Kevin xxxxx xxxxxxxxxxx	Nam	Nữ		Nam
5	Syaiful xxx	Nam	Nữ		Nam
6	Winda xxxxxx xxxxxxx	Nam	Nữ		Nữ
7	M. Nashrullah	Nam	Nữ		Nam
8	Muhammad xxxxx	Nam	Nữ		Nam
9	Chaterine xxxxxxxx xxxxx	Nữ	Nam		Nữ
10	Vitra xxxxxxxxxxx	Nữ	Nam		Nữ
11	Dinara xxxxxx	Nữ	Nam		Nữ
12	Aqilah xxxxxx	Nữ	Nam		Nữ
13	Della xxxxxxxxx	Nữ	Nam		Nữ
14	Migawati	Nữ	Nam		Nữ
15	Nasywa xxxxxxxx xxxxx	Nữ	Nam		Nữ
16	Mira xxxxxx	Nữ	Nam		Nữ
17	Sariba	Nữ	Nam		Nữ

Lưu ý: Một số tên không được hiển thị vì lý do bảo mật

Bảng 5: Ma trận so sánh về trình độ học vấn cao nhất từ dữ liệu hành chính và điều tra thử nghiệm

		Dữ liệu hành chính										
	Loại	Chưa đi học	Dưới tiểu học	Tiểu học	Trung học cơ sở	Trung học phổ thông	Chứng chỉ I/II	Chứng chỉ III	Đại học	Thạc sĩ	Tiến sĩ	Tổng
Dữ liệu điều tra thử nghiệm	Chưa đi học	143	4		1				1			149
	Dưới tiểu học	112	88	3	2	3						208
	Tiểu học	19	41	162	6	5			1			234
	Trung học cơ sở	4	39	7	182	23						255
	Trung học phổ thông		10	7	27	408	1	4	2			459
	Chứng chỉ I/II					1	10	1	1			13
	Chứng chỉ III				1	8	1	58	4			72

	Đại học		1	1	2	17		7	192	9	1	230
	Thạc sĩ							2	18	21		41
	Tiến sĩ									3	4	7
	Tổng	278	183	180	221	465	12	72	219	33	5	1668

Bảng 6: Giá trị P trong các kiểm định phi tham số ghép đôi của 8 biến trong dữ liệu hành chính và cuộc điều tra thử nghiệm

STT	Biến	Giá trị P
1	Giới tính	0.332
2	Tôn giáo	0.072
3	Trình độ giáo dục cao nhất	**0.000***
4	Tình trạng hôn nhân	**0.005***
5	Tuổi	0.357
6	Ngày sinh	0.377
7	Nơi cư trú	0.155
8	Nơi sinh	0.952

3.2. Các kiểm định thống kê

Để tìm ra sự khác biệt về mặt thống kê giữa dữ liệu từ hai nguồn, các kiểm định giả thuyết được thực hiện với tám biến số được lựa chọn. Bảng 6 chỉ ra giá trị p từ kiểm định phi tham số với mỗi biến. Kết quả kiểm định thống kê đối với 6 biến số: Giới tính, Tôn giáo, Tuổi, Ngày sinh, Địa phương cư trú, và Nơi sinh, không thể hiện có sự khác biệt lớn giữa cuộc điều tra thử nghiệm và dữ liệu hành chính. Mặt khác, các kiểm định thống kê về Bằng cấp cao nhất đạt được và Tình trạng hôn nhân chỉ ra rằng có những sự khác biệt về dữ liệu từ hai nguồn.

4. Kết luận và kiến nghị

Đánh giá chất lượng dữ liệu hành chính với mục đích điều tra được thực hiện bằng cách so sánh các dữ liệu từ các hồ sơ hành chính với dữ liệu từ một cuộc điều tra thử nghiệm. Các kết quả mô tả và thống kê đã nhất quán chỉ ra rằng dữ liệu hành chính có một độ chính xác tốt đối với một vài biến như giới tính, tôn giáo, tuổi, ngày sinh, nơi cư trú, nơi sinh. Tuy nhiên, có những sự khác biệt lớn đối với biến số Bằng cấp giáo dục đạt được và tình trạng hôn nhân.

Dựa trên các kết quả thử nghiệm, có thể kết luận rằng các hồ sơ hành chính từ Bộ Nội vụ có thể được sử dụng làm nguồn dữ liệu ban đầu sử dụng cho tổng điều tra theo phương pháp kết hợp tại Indonesia. Tại cuộc tổng điều tra dân số năm 2020, dữ liệu hành chính sẽ được sử dụng mục đích chính cho mục đích kiểm tra và xác nhận phạm vi bao phủ. Hơn nữa, các thông tin cá nhân về Giới tính từ dữ liệu hành chính có chất lượng tốt và có thể được mượn như là biến số trong điều tra nhằm giảm thiểu lỗi gây ra trong cuộc tổng điều tra dân số.

Điều quan trọng cần chú ý tới là chất lượng dữ liệu được cung cấp bởi điều tra viên sử dụng CAPI hoặc bởi người được phỏng vấn độc lập thông qua trang web. Dựa trên các phát hiện từ cuộc điều tra thử, vẫn có những lỗi dữ liệu xuất phát từ hai nguồn này. Các lỗi có thể bị gây ra do sự thiếu hiểu biết của điều tra viên hoặc người được phỏng vấn về các khái niệm xác định nào đó, hoặc bởi giao diện sử dụng của ứng dụng còn khiếm khuyết. Bởi vậy, việc đào tạo những điều tra viên trong lĩnh vực là một trong những nhân tố quan trọng để đảm bảo chất lượng dữ liệu. Đối với người được phỏng vấn điền các thông tin vào bảng hỏi thông qua trang web thì cần phải cung cấp giao diện người dùng đầy đủ. Cũng rất cần thiết cung cấp các khái niệm và định nghĩa hoặc các hướng dẫn một cách dễ hiểu cho công chúng.

Để có thể thực hiện việc điều tra dân số dựa trên đăng ký một cách đầy đủ trong tương lai, chất lượng của dữ liệu hành chính từ Bộ Nội

vụ phải được cải thiện hơn nữa. Sự chú ý đặc biệt phải được dành cho các biến số với độ chính xác thấp, ví dụ các biến số dễ bị thay đổi theo thời gian. Ngoài ra, khuyến nghị việc đánh giá các biến cá nhân trong các nguồn dữ liệu hành chính khác ở Indonesia. Thông tin từ nhiều nguồn dữ liệu hành chính khác nhau có thể được sử dụng để hoàn thiện dữ liệu cá nhân cho điều tra dựa trên đăng ký tại Indonesia.

Lời cảm ơn

Chúng tôi xin cảm ơn Ông Eunkoo Lee từ Viện Thống kê Liên hợp quốc về Châu Á Thái Bình Dương về sự hướng dẫn tận tình để hoàn thiện nghiên cứu này, Ông Matthew Shearing và Uỷ ban Xã hội và Kinh tế LQH về châu Á - Thái Bình Dương về tạo điều kiện hướng dẫn.

Tài liệu tham khảo

1. UNDESA Statistics Division. Principles and recommendations for population and housing censuses, revision 3. New York: United Nations; 2017.

2. UNECE. Guidelines on the use of registers and administrative data for population and housing censuses. Geneva: UNECE; 2018.

3. Statistics Korea. Introducing register-based census in Korea. KOSTAT. 2018 [cited 1 May 2020]. Available from http:// kostat.go.kr/iwsm/download/2018/S3-4.pdf.

4. Jialin C, Lip T. Challenge in the development of registerbased population statistics. Singapore: Statistics Singapore Newsletter; 2017.

5. Schulte Nordholt E. The usability of administrative data for register-based censuses. Statistical Journal of the IAOS. 2018 Jan 1; 34(4): 487–98. doi: 10.3233/SJI-180425.

6. Valle JLV, Jiménez AA, Julián MP. Moving towards a register based census in Spain. Statistical Journal of the IAOS. 2020 Jan 1; 187–192. doi: 10.3233/SJI-190516.

7. Kemendagri. Go digital siap wujudkan single identity number. Kementerian Dalam Negeri Republik Indonesia. [updated 9 February 2019; cited 1 May, 2020]. Available from https:// www.kemendagri.go.id/berita/baca/19131/Go-Digital-Dukca pil-Siap-Wujudkan-Single-Identity-Number.

8. Bachenheimer D, Baker D, Banerjee S, Chatfield C, Hyvonen I, Iyer A, Jha M, Kudaravalli S, Leong C, Madhav S, Malik R. Technology landscape for digital identification. World Bank; 2018.

9. Dygaszewicz J. Transition from traditional census to combined and registers based census. Statistical Journal of the IAOS. 2020 Jan 1; (36): 165–175. doi: 10.3233/SJI-190566

10. Lagarto S, Delgado A, Paulino P, Capelo J. When is administrative data good enough to replace statistical information? A quality indicator based on census comparison. Statistical Journal of the IAOS. 2017 Jan 1; 33(3): 749–53. doi: 10.3233/SJI160333.

11. Poston DL, Micklin M. Handbook of population. Springer US; 2005. doi: 10.1007/b100598.

Đậu Trang – Ngọc Mai (dịch)

Nguồn: https://content.iospress.com/download/statistical-journal-of-the-iaos/sji200744?id=statistical-journal-of-the-iaos/sji200744