ĐẾM HOẶC ƯỚC TÍNH
MỘT LƯU Ý VỀ VIỆC BIÊN SOẠN ƯỚC TÍNH DÂN SỐ TỪ DỮ LIỆU HÀNH CHÍNH

John Dunne, Francesca Kay and Timothy Linehan
Cơ quan Thống kê Trung ương Ireland

Tóm tắt

Giống như nhiều quốc gia khác, Ireland đã nghiên cứu các hệ thống ước tính dân số mới được biên soạn bằng dữ liệu hành chính. Ireland không có đăng ký dân số trung ương để có thể biên soạn các ước tính.

Bước chính trong việc biên soạn ước tính dân số từ dữ liệu hành chính là trước tiên xây dựng một Bộ dữ liệu dân số thống kê (SPD). Lý tưởng nhất là một SPD sẽ có một bản ghi cho mỗi người dân chứa các thuộc tính có liên quan. Sau đó, SPD sẽ cho phép biên soạn số liệu thống kê bằng cách chỉ cần đếm qua các bản ghi.

Trên thực tế, việc biên soạn SPD dễ xảy ra sai số. Các sai số này có thể được phân loại thành 4 loại sai số: sai số quá phạm vi, sai số thiếu phạm vi, sai số phân loại và lỗi liên kết.

Cho đến nay, Ireland đã nghiên cứu 2 cách tiếp cận khác nhau để biên soạn ước tính dân số từ dữ liệu hành chính. Cách đầu tiên là *phương pháp đếm đơn giản*, cách thứ hai là *phương pháp ước tính*. Bài viết này sẽ khám phá những ưu điểm và nhược điểm của cả hai phương pháp trước khi xem xét cách chúng có thể được tích hợp để loại bỏ những nhược điểm.

Nhiều Cơ quan thống kê quốc gia sẽ cân nhắc những thách thức tương tự khi biên soạn Điều tra dân số hàng năm như ước tính dân số và bài viết này nhằm mục đích đóng góp vào cuộc thảo luận đó.

1. Giới thiệu

Các cơ quan thống kê ở nhiều quốc gia đang nghiên cứu các phương pháp thay thế các hệ thống ước tính dân số dựa trên điều tra dân số truyền thống. Không phải quốc gia nào cũng có đăng ký dân số trung ương (CPR) có thể dễ dàng sử dụng làm cơ sở cho số liệu thống kê dân số được biên soạn trực tiếp. Ireland là một quốc gia như vậy. Cơ quan Thống kê Trung ương (CSO) Ireland, giống như nhiều cơ quan thống kê khác, đã đầu tư nguồn lực đáng kể vào việc khai thác các nguồn dữ liệu hành chính cho mục đích thống kê. Là một phần của nỗ lực này, CSO đã nghiên cứu các phương pháp mới để biên soạn ước tính dân số.

Bước đầu tiên trong việc biên soạn ước tính dân số là biên soạn SPD từ các nguồn dữ liệu hành chính. Ý tưởng đơn giản đằng sau SPD là nó có thể được sử dụng thay cho CPR để đếm số người trong dân số cho một điểm tham chiếu hoặc thời kỳ tham chiếu nhất định.

SPD lý tưởng sẽ có bản ghi cho từng đơn vị thống kê (người) trong dân số mục tiêu - mỗi đơn vị được xác định bằng một số nhận dạng duy nhất. Dân số mục tiêu để ước tính dân số yêu cầu một người đang sống trong Tiểu bang. Sẽ có những biến thể được định nghĩa cơ bản, trên thực tế, trên nhận định, đã đăng ký,... nhưng tiền đề cơ bản là người đó phải sống trong Tiểu bang. Khi biên soạn SPD từ nhiều

nguồn dữ liệu, có thể phát sinh bốn loại sai số chính liên quan đến dân số mục tiêu:

● Sai số quá phạm vi: Khi SPD có các đơn vị không thuộc nhóm dân số mục tiêu.

● Sai số thiếu phạm vi: Trường hợp SPD thiếu các đơn vị thuộc nhóm dân số mục tiêu.

● Sai số liên kết: Khi các đơn vị được xác định không chính xác là các đơn vị khác, ví dụ khi Mã số nhận dạng cá nhân (PIN) không chính xác.

● Phân loại sai: Khi một thuộc tính có giá trị không chính xác cho một đơn vị. Điều này có thể xảy ra khi các thuộc tính giống nhau hoặc tương tự trên các nguồn dữ liệu đóng góp khác nhau có giá trị xung đột.

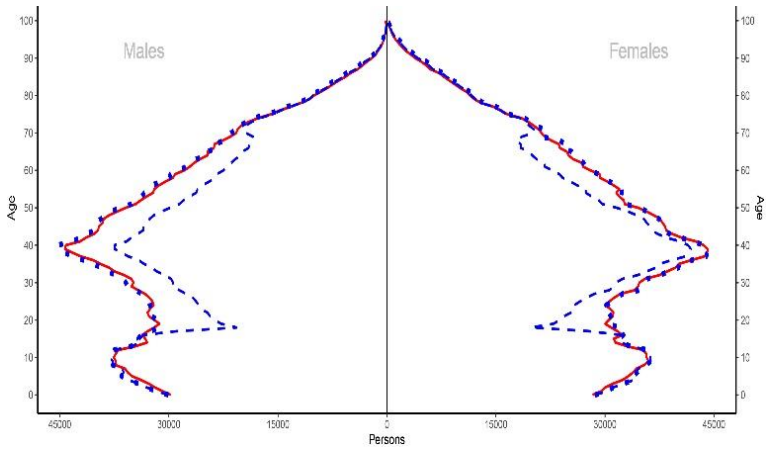
Cho đến nay, Ireland đã nghiên cứu hai cách tiếp cận khác nhau để biên soạn ước tính dân số từ dữ liệu hành chính. Cách tiếp cận đầu tiên, được trình bày trong bài viết này là *phương pháp đếm đơn giản*, dựa trên việc xây dựng một SPD giúp giảm thiểu tổng số sai số bản ghi cá nhân sao cho các số đếm đơn giản từ SPD sẽ cung cấp ước tính dân số. Cách tiếp cận thứ hai, được trình bày trong bài viết này là *phương pháp ước tính*, dựa trên việc xây dựng một SPD nhằm mục đích loại bỏ mọi loại sai số ngoại trừ sai số quá phạm vi và sau đó điều chỉnh số đếm sao cho việc thiếu phạm vi bằng các phương pháp Ước tính hệ thống kép (DSE) để có được ước tính dân số.

2. Phương pháp

2.1. Phương pháp đếm đơn giản

Phương pháp đếm đơn giản được sử dụng để biên soạn ước tính dân số cho năm tham chiếu 2020. Dân số ước tính là 5,2 triệu. Sự phân chia độ tuổi theo giới tính được cung cấp trong Hình 1, trong đó phương pháp đếm đơn giản và ước tính được so sánh cho năm tham chiếu 2020.

Hình 1. So sánh phương pháp đếm đơn giản với phương pháp ước tính khi sử dụng để biên soạn ước tính dân số theo giới tính và tuổi, năm 2020.



Đường liên tục màu đỏ biểu thị ước tính sử dụng phương pháp đếm đơn giản - SPD biên soạn dựa trên việc giảm thiểu số sai số. Đường chấm màu xanh biểu thị ước tính sử dụng phương pháp ước tính - số đếm điều chỉnh từ SPD biên soạn dựa trên việc giới hạn loại sai số thiếu phạm vi - SPD cũng được biểu thị bằng đường đứt nét màu xanh.

Phương pháp này áp dụng cách tiếp cận Dấu hiệu sự sống (SoL) để biên soạn SPD. Các quy tắc hỗ trợ SoL được chọn theo cách trực quan để nhắm mục tiêu vào một bản ghi trong SPD cho mỗi người trong dân số. Cách tiếp cận này dựa trên việc giảm thiểu số sai số khi đếm các bản ghi trong SPD để ước tính dân số. Trong thực tế, sẽ có sai số trong SPD và những sai số này cũng sẽ triệt tiêu lẫn nhau ở một mức độ nào đó.

Theo một nghĩa nào đó, cách tiếp cận này tương tự như cách tiếp cận dựa trên đăng ký

dân số của Bắc Âu. Ở đây, người ta chỉ đơn giản cho rằng tất cả mọi người dân đều tham gia vào một trong những hệ thống hành chính công cơ bản và việc ước tính dân số chỉ đơn giản là đếm những người tham gia vào các hệ thống hành chính công. Trong hệ thống dựa trên đăng ký của Bắc Âu, dân số được coi là dân số đã đăng ký - toàn bộ dân số và chỉ dân số được đăng ký. Người ta thường chấp nhận rằng các sai số quá phạm vi là không đáng kể và số lượng đã đăng ký là đủ. Những sai số này trước đây đã được ước tính và cung cấp sự an tâm cho người dùng đã sử dụng cách tiếp cận kiểu SoL trong các nguồn dữ liệu hành chính để khám phá các vấn đề quá phạm vi và chứng minh rằng khả năng quá phạm vi trong CPR ít hơn đáng kể so với 1% dân số trong trường hợp của Thụy Điển.

Các nguồn dữ liệu liên quan đến các khoản thanh toán trợ cấp trẻ em phổ cập, tuyển sinh tiểu học, tuyển sinh sau tiểu học, tuyển sinh trung học và giáo dục nâng cao, tự kinh doanh, việc làm, phúc lợi xã hội và thanh toán lương hưu đã được sử dụng để xác định những người được đưa vào SPD.

Vị trí hoặc nơi cư trú sau đó được chỉ định bằng cách sử dụng phương pháp tiếp cận dựa trên quy tắc để làm nổi bật hơn các nguồn dữ liệu được coi là có chất lượng cao hơn. Các nguồn dữ liệu được sử dụng để chỉ định địa lý bao gồm đăng ký cho thuê, thuế tài sản địa phương cho chủ sở hữu bất động sản và danh sách các địa chỉ do Bộ Bảo vệ Xã hội quản lý.

Một số lượng hạn chế các thuộc tính cũng được bao gồm trong SPD cho mỗi người dân. Những thuộc tính này bao gồm lĩnh vực kinh tế của người sử dụng lao động đối với người lao động, quốc tịch và liệu một người có nhận được khoản thanh toán phúc lợi hay không cùng với các thuộc tính cốt lõi như tuổi tác, giới tính và quốc tịch.

Có một số nhược điểm. Theo cách tiếp cận này, việc điều chỉnh các quy tắc để đưa một người vào SPD có thể tác động trực tiếp đến số lượng dân số. Tuy nhiên, nếu các quy tắc được áp dụng theo cách nhất quán từ năm này sang năm khác và các nguồn dữ liệu cơ bản ổn định và mạnh mẽ trong hoạt động của chúng, thì có thể lập luận rằng các sai số được đưa vào là có hệ thống và do đó, không nên tác động đến các quan sát về bản chất thay đổi của dân số từ năm này sang năm khác. Những thay đổi về bản chất của dữ liệu trong các nguồn dữ liệu cơ bản hoặc những thay đổi liên quan đến tính khả dụng của các nguồn dữ liệu cũng sẽ là một điểm yếu hoặc nhược điểm của cách tiếp cận này. Những thay đổi trong một nguồn dữ liệu cơ bản có thể xảy ra do sự thay đổi trong hành vi của dân số liên quan đến các tương tác với hệ thống hành chính công tương ứng, sự thay đổi trong các quy tắc (hoặc việc thực hiện các quy tắc) để vận hành một hệ thống hành chính công hoặc một số lý do khác.

Phương pháp này có ưu điểm là có thể biên soạn các bảng chéo thống nhất về ước tính cho dân số chỉ bằng cách đếm các bản ghi cho mỗi ô bảng. Điều này có thể được thực hiện đối với bất kỳ thuộc tính nào có nguồn gốc từ các biến có trong các nguồn dữ liệu đóng góp vào SPD. Tại thời điểm viết bài này, phương pháp này chỉ được áp dụng cho một năm tham chiếu - 2020.

2.2. Phương pháp ước tính

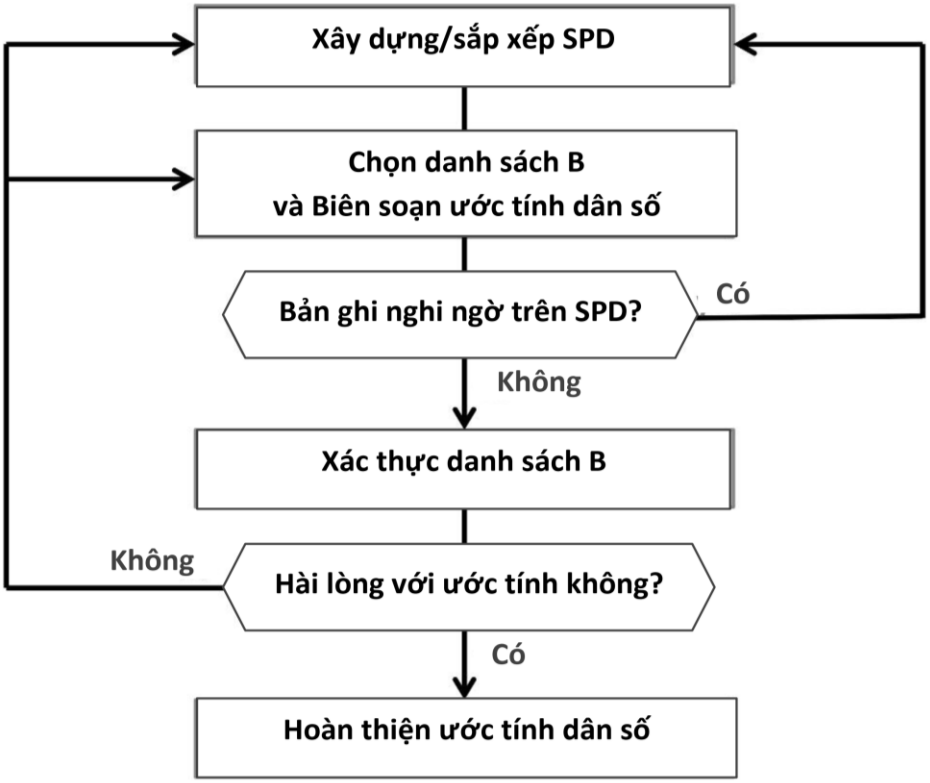
2.2.1. Tổng quan về phương pháp ước lượng

Phương pháp ước tính này xuất phát từ dự án PECADO (Ước tính dân số chỉ được biên soạn từ dữ liệu hành chính) của Ireland.

Phương pháp ước tính, khi áp dụng bằng cách sử dụng các nguồn dữ liệu hành chính được chọn cho năm tham chiếu 2020, ước tính dân số Ireland là 5,3 triệu người. Phân tích độ tuổi theo giới tính được cung cấp trong Hình 1, trong đó phương pháp đếm đơn giản và ước tính được so sánh cho năm tham chiếu 2020.

Ở mức đơn giản nhất, phương pháp ước tính là một quy trình gồm 2 bước.

Hình 2. Bản đồ quy trình cấp cao để biên soạn ước tính dân số trong dự án PECADO của Ireland.



Bước đầu tiên liên quan đến việc biên soạn SPD chỉ có một loại sai số liên quan đến dân số - quá phạm vi. Với phương pháp này, SPD được biên soạn dựa trên việc áp dụng tiêu chí SoL nghiêm ngặt trên một tập hợp các nguồn dữ liệu tương tự như đối với phương pháp đếm. Sự khác biệt chính là tiêu chí được sử dụng cho SPD trong *phương pháp ước tính* là để giảm các loại sai số cần xử lý xuống còn một loại - đó là thiếu phạm vi. Các tiêu chí nghiêm ngặt có mục đích đảm bảo rằng tất cả các bản ghi có trong SPD đều đại diện cho một người trong dân số, nhưng SPD không nhất thiết phải chứa một bản ghi cho mọi người trong dân số. Thiếu phạm vi là điều được mong đợi.

Bước thứ hai yêu cầu điều chỉnh số lượng SPD cho tình trạng thiếu phạm vi. Sau khi SPD được biên soạn, một nguồn dữ liệu được chỉ định cố tình loại trừ khỏi quá trình biên soạn SPD được sử dụng làm danh sách thứ hai trong thiết lập DSE để điều chỉnh số lượng SPD cho tình trạng thiếu phạm vi nhằm có được ước tính dân số.

Trong thực tế, một quy trình lặp lại được sử dụng với bộ công cụ DSE mở rộng, được gọi là bộ công cụ PECADO, để đảm bảo các ước tính là mạnh mẽ và có thể được bảo vệ theo quan điểm phương pháp luận. Bộ công cụ cũng có thể được sử dụng để xử lý các bản ghi nghi ngờ và sai số quá phạm vi trong danh sách A. Hình 2 minh họa quy trình lặp lại được áp dụng để đảm bảo các ước tính là mạnh mẽ.

Các công cụ/phương pháp được sử dụng để biên soạn ước tính dân số (bộ công cụ PECADO) được mô tả bên dưới. Lưu ý rằng DSE và DSE được cắt giảm (TDSE) lần đầu tiên được trình bày ước tính dân số cho các năm tham chiếu từ 2011 đến 2016 cùng với lập luận hợp lý về tính vững chắc của chúng.

Nhược điểm của cách tiếp cận này là không có tập dữ liệu hoàn chỉnh cho tổng thể dân số. SPD được biên soạn như một phần của cách tiếp cận này có thể chứa các vấn đề thiếu hụt đáng kể và do đó không dễ để tạo ra nhiều bảng chéo khác nhau cho dân số theo cách mạch lạc.

2.2.2. Tóm tắt bộ công cụ PECADO

2.2.2.1. DSE được xem xét lại

Trong thiết lập DSE, hai danh sách được biểu thị là danh sách A (có kích thước *x*) và danh sách B (kích thước *n*) và kích thước của sự so khớp giữa hai danh sách là *m*. Ba giả định chính cần thiết là:

i) Không có bản ghi sai: Một tổng thể dân số chính xác nhất đảm bảo không có bản ghi nào từ bên ngoài nhưng chúng tôi cũng cho rằng không có bản ghi trùng lặp hoặc bản ghi được xác định không chính xác trong danh sách A hoặc danh sách B.

ii) Giả định khớp: Không có sai số liên kết khi khớp các bản ghi giữa danh sách A và danh sách B.

iii) Giữ đồng nhất đối với danh sách B: Mỗi đơn vị *i* trong tổng thể *U* có cơ hội bằng π được giữ trong danh sách B.

Những giả định này bây giờ cho phép chúng ta viết

$$\widehat{N} = \frac{nx}{m} \qquad (1)$$

Chúng tôi xem xét *x* như một hằng số cố định (danh sách A có thể là bất kỳ danh sách cố định nào) và chỉ *n* và *m* thay đổi do tính ngẫu nhiên trong danh sách B.

Giả định bổ sung được sử dụng để cho phép ước tính phương sai liên quan đến *việc lưu giữ độc lập* những người trong tổng thể trên danh sách B, nghĩa là, sự kiện một người bị lưu giữ trên danh sách B không phụ thuộc vào việc bất kỳ người nào khác có bị lưu giữ trên danh sách B hay không. Bộ ước tính phương sai giống như DSE tiêu chuẩn được mô tả trong sách giáo khoa theo công thức:

$$\hat{V}\left[\widehat{N}\right] = \frac{n(n-m)x(x-m)}{m^3} \qquad (2)$$

Giả định bổ sung cho phép sử dụng định lý nhị thức để có được các ước lượng cho các phương sai của *n* và *m* và cũng cho phép chúng ta cân bằng phương sai của *n* và *m* đến sự thay đổi của *m*.

2.2.2.2. TDSE và tìm kiếm các bản ghi sai số trong danh sách A

Coi như *r* là các bản ghi sai sót được chứa trong danh sách A, ước lượng Eq bây giờ sẽ ước tính quá phạm vi tổng thể. Nếu số lượng bản ghi sai sót *r* được biết, thì một ước lượng lý tưởng hoặc không chệch được đưa ra bởi

$$\widetilde{N} = \frac{n(x-r)}{m} \qquad (3)$$

Tuy nhiên, trong thực tế *r* chưa được biết.

Nếu chúng ta cắt *k* bản ghi từ danh sách A, *k* bản ghi có thể được phân vùng thành ba nhóm. Nhóm đầu tiên sẽ chứa các bản ghi sai số, nhóm thứ hai sẽ chứa các bản ghi hợp lệ không được chọn trong danh sách B và nhóm thứ ba sẽ chứa các bản ghi hợp lệ đã được chọn trong danh sách B, được ký hiệu là k_1. Bây giờ chúng ta có thể viết một ước lượng mới (TDSE) như sau:

$$\widehat{N_k} = n\frac{n-k}{m-k_1} \qquad (4)$$

Chúng ta có thể sử dụng TDSE, lưu ý các kết quả để đánh giá sự hiện diện của các bản ghi sai số trong các tập hợp con khác nhau của danh sách A

i) Nếu $k_1/m < k/x$, thì $\widehat{N_k} < \widehat{N_0}$. Có bằng chứng về các bản ghi sai trong phần tử đã cắt của danh sách A.

ii) Nếu $k_1/m = k/x$, thì $\widehat{N_k} = \widehat{N_0}$. Không có bằng chứng nào về các bản ghi sai trong phần tử đã cắt của danh sách A.

iii) Nếu $k_1/m > k/x$, thì $\widehat{N_k} > \widehat{N_0}$. Có bằng chứng về các bản ghi sai sót còn sót lại trong phần tử chưa cắt bớt của danh sách A.

iv) Nếu $k < r$, thì $\widetilde{N} < \widehat{N_k}$. Ước tính đã cắt không thể loại bỏ mọi sai lệch do bản ghi sai sót khi $k < r$.

v) Nếu tất cả các bản ghi sai số *r* đều nằm trong số *k* bản ghi bị cắt bớt, thì $\hat{E}\left[\widehat{N_k}\right] = \tilde{N}$.

Tóm lại, nếu các bản ghi sai số tồn tại trong danh sách A, miễn là người ta có thể cắt các bản ghi sai số trong danh sách A hiệu quả hơn so với khi cắt ngẫu nhiên các bản ghi, thì có thể mong đợi TDSE tại công thức (4) sẽ làm giảm độ lệch của DSE tại công thức (1) khi không có bản ghi sai số nào được giả định và đưa nó đến gần hơn với DSE lý tưởng tại công thức (3). Nếu việc cắt thành công trong việc loại bỏ tất cả các bản ghi sai số, thì kỳ vọng của TDSE sẽ trở nên gần giống với DSE lý tưởng tại công thức (3).

Nếu giả sử rằng tất cả các bản ghi sai số đã được xóa khỏi danh sách A, thì ước lượng phương sai cho TDSE có thể được viết như sau:

$$\hat{V}\left[\widehat{N_k}\right] = \frac{n(n-m_k)x_k(x_k-m_k)}{m_k^3} \qquad (5)$$

Một chiến lược cắt tỉa hiệu quả sẽ loại bỏ các bản ghi sai mà không làm tăng phương sai của ước lượng đến mức không còn hữu ích hoặc ổn định nữa. Có một sự đánh đổi khi cắt tỉa - số lượng bản ghi cắt tỉa càng nhiều thì ước lượng phương sai tăng càng nhiều, được biểu thị tại công thức 5.

2.2.2.3. Tác động đến ước tính khi giả định lưu giữ đồng nhất trong danh sách B

Dunne và Zhang xem xét tác động của sự vi phạm trong giả định về tỷ lệ lưu giữ đồng nhất đối với danh sách B. Điều này được thực hiện bằng cách xem xét phân vùng của dân số thành hai nhóm con hoặc tầng lớp và sau đó xem xét sự khác biệt giữa ước tính phân tầng ($\widehat{N'}$) và một ước lượng không phân tầng ($\widehat{N}$). Nếu ước tính dân số cho hai nhóm phụ (được biểu thị bằng chỉ số 1 và 2) được đưa ra bởi $\widehat{N_1} = n_1x_1/m_1$ và $\widehat{N_2} = n_2x_2/m_2$ trong đó giả định *chiếm giữ đồng nhất* được coi là giữ nguyên trong mỗi nhóm phụ và ước tính dân số cho ước tính không phân tầng được đưa ra bởi $\widehat{N} = (n_1+n_2)(x_1+x_2)/(m_1+m_2)$ sau đó sự khác biệt có thể được viết như sau:

$$D = \widehat{N} - \widehat{N'} = \left(\frac{n_1}{\widehat{N}_1} - \frac{n_2}{\widehat{N}_2}\right)\left(\frac{x_2}{\widehat{N}_2} - \frac{x_1}{\widehat{N}_1}\right)\frac{\widehat{N}_1\widehat{N}_2}{m} \qquad (6)$$

Từ công thức (6) ta thấy rằng D = 0 khi ngay $x_2/\widehat{N}_2 = x_1/\widehat{N}_1$ ngay cả khi xác suất lưu giữ danh sách B thay đổi trên cả hai phần. Nói cách khác, việc lưu giữ danh sách B không đồng nhất tự nó không nhất thiết gây ra sự chênh lệch lớn $\widehat{N}$ cũng cung cấp một bài kiểm tra để đánh giá tác động lên các ước tính do tính không đồng nhất giữa nhiều nhóm phụ.

2.2.2.4. Một lưu ý nữa về bộ công cụ PECADO

Một khía cạnh cải tiến của bộ công cụ *PECADO* là nó xem xét lại thiết lập DSE, đặc biệt là các giả định được sử dụng và nêu các phương pháp DSE sao cho các giả định được nới lỏng và nêu lại thành ba giả định chính với giả định thứ tư được đưa vào để cho phép ước tính phương sai.

Bộ công cụ PECADO cũng mở rộng các phương pháp DSE truyền thống sao cho các phần của SPD có thể được đánh giá cho các bản ghi sai bao gồm cả việc quá phạm vi. Đây là một phần mở rộng quan trọng vì giờ đây nó cho phép xác thực giả định không có bản ghi sai khi biên soạn ước tính. Nói chung hơn, phần mở rộng này cho phép sử dụng các phương pháp DSE để xử lý các sai số quá phạm vi.

Việc nới lỏng các giả định và khả năng có thể mở rộng các phương pháp tự thân cung cấp một phạm vi ứng dụng rộng hơn nhiều cho các phương pháp DSE. Một ứng dụng khả thi là thay thế khảo sát sau khi liệt kê truyền thống như một phần của Điều tra dân số truyền thống bằng cách sử dụng danh sách hành chính đơn giản với việc áp dụng các phương pháp DSE. Ngược lại với trực giác, Điều tra dân số có thể được coi là một cuộc khảo sát phạm vi bao phủ lớn cho bất kỳ danh sách hành chính nào (chỉ có phạm vi bao phủ không đầy đủ) trong thiết lập

DSE. Nếu một ứng dụng như vậy khả thi thì sẽ có những lợi ích đáng kể về mặt chi phí, tính kịp thời và giảm độ phức tạp của các hoạt động trong mô hình Điều tra dân số truyền thống.

2.2.3. Sơ lược về cách áp dụng bộ công cụ trong dự án PECADO

Hình 2 cung cấp tổng quan hoặc sơ đồ quy trình cấp cao về cách ước tính đã được biên soạn bằng cách sử dụng bộ công cụ và có thể được sử dụng để cung cấp sự đảm bảo xung quanh các ước tính. Chúng tôi mô tả ngắn gọn quy trình này bên dưới. Để xem xét sâu hơn về quy trình này và tính khả thi của nó, chúng tôi giới thiệu người đọc đến tài liệu tham khảo 4.

Để bắt đầu, một phương pháp tiếp cận SoL được sử dụng để xây dựng SPD ban đầu bằng cách sử dụng các nguồn dữ liệu đã chọn. Một nguồn dữ liệu ứng viên được loại trừ khỏi lựa chọn này và được chỉ định là danh sách B trong thiết lập DSE. Về lý thuyết, SPD này chỉ nên có phạm vi bao phủ thấp đối với dân số - việc sử dụng các số nhận dạng chất lượng cao giúp giảm thiểu sai số liên kết trong dự án PECADO.

Trong dự án PECADO có 2 nguồn dữ liệu được chọn cho danh sách B. Nguồn dữ liệu đầu tiên được tạo bằng hoạt động (gia hạn hoặc nộp đơn) trên hệ thống Dữ liệu Giấy phép lái xe Ireland (DLD) và nguồn dữ liệu thứ hai là khảo sát hộ gia đình được thiết kế để chọn những cá nhân từ dân số có xác suất bằng nhau. Nguồn DLD được coi là nguồn dữ liệu chính cho danh sách B vì phạm vi bao phủ của danh sách này phải đảm bảo ước tính chính xác (sai số chuẩn thấp). Xem xét công thức (6), chúng tôi cũng đưa ra giả định rằng phạm vi bao phủ dân số trên SPD không khác biệt đáng kể đối với người lái xe và người không lái xe. Phân tầng sau theo nhóm tuổi, giới tính và quốc tịch riêng lẻ được triển khai để giảm thiểu thêm bất kỳ sự chênh lệch nào do vi phạm giả định lưu giữ đồng nhất.

Ước tính dân số và khoảng tin cậy hiện được biên soạn theo tầng bằng cách sử dụng các phương trình (1) và (2).

Trên thực tế, người ta có thể nghi ngờ rằng một hoặc nhiều nguồn dữ liệu cơ bản (được sử dụng để biên soạn SPD) chứa các bản ghi sai và do đó sẽ dẫn đến ước tính quá cao về quy mô dân số, lưu ý công thức (3). Nếu các quy tắc SoL có hiệu quả thì sẽ không có bản ghi sai nào trên SPD. Để kiểm tra các quy tắc SoL, các phương pháp TDSE được mô tả ở trên được sử dụng để săn tìm các bản ghi sai bằng cách loại bỏ từng nguồn dữ liệu cơ bản theo lượt khi xây dựng SPD và so sánh các ước tính.

Trên thực tế, người ta có thể đưa ra lập luận rằng danh sách B (DLD) vi phạm giả định về *phạm vi thu thập đồng nhất* và sẽ dẫn đến sai lệch đáng kể. Để kiểm tra lập luận này, các ước tính dân số được biên soạn lại bằng cách sử dụng nguồn dữ liệu thứ hai (một cuộc khảo sát hộ gia đình) và các ước tính được so sánh. Kết quả cho thấy một tập hợp các ước tính nhất quán (lưu ý rằng tập hợp các ước tính thứ hai sử dụng khảo sát hộ gia đình sẽ có khoảng tin cậy lớn hơn) và do đó, không có bằng chứng nào cho thấy giả định rằng phạm vi dân số trên SPD không khác biệt đáng kể đối với người lái xe và người không lái xe là không hợp lệ.

Để đối chiếu các ước tính với số liệu thống kê dân số hiện có (năm tham chiếu 2016), bộ công cụ được triển khai trong một kịch bản mà danh sách Điều tra dân số được coi là danh sách B (một cuộc khảo sát phạm vi bao phủ đáng kể) cho một danh sách hành chính (danh sách A) được biên soạn từ những người nhận được khoản thanh toán phúc lợi xã hội trong tháng Điều tra dân số được thực hiện. Có tính đến các khái niệm dân số cơ bản, sự khác biệt trong hai bộ ước tính có thể được đối chiếu hợp lý ở cấp độ khái niệm. Dự án PECADO sử dụng khái niệm loại Dân số thường trú hàng năm (một người được coi là thường trú tại bất kỳ thời điểm nào trong năm) khi ước tính quy mô dân số trong khi biện pháp khái niệm về số liệu Điều tra dân số của Ireland dựa trên việc thường trú vào thời điểm Điều tra dân số. Sự khác biệt về khái niệm có thể được giải thích

một cách hợp lý bằng các ước tính di cư. Ý tưởng sử dụng một khái niệm cơ sở phù hợp chặt chẽ với cách ước tính dân số tốt nhất không phải là mới, các quốc gia Bắc Âu sử dụng khái niệm dân số đã đăng ký để củng cố số liệu thống kê dân số của họ và cung cấp một sự cân nhắc sâu sắc về các khái niệm dân số khác nhau và ý nghĩa của chúng.

3. Thảo luận thêm và đề xuất kết hợp các phương pháp

Khi xem xét so sánh các ước tính từ hai phương pháp, phương pháp đếm đơn giản và phương pháp ước tính, trong Hình 1 chúng ta thấy hai tập ước tính này về cơ bản là tương đương nhau.

Điểm mạnh tương đối của phương pháp ước tính là nó có thể được bảo vệ như một tập hợp ước tính mạnh mẽ theo quan điểm phương pháp luận trong khi điểm mạnh tương đối của phương pháp đếm đơn giản là bảng chéo chỉ đơn giản được rút ra bằng cách đếm các chiều khác nhau trong SPD.

Việc xem xét các phương pháp cơ bản và thế mạnh so sánh của chúng dẫn đến một đề xuất kết hợp cả hai phương pháp để tận dụng thế mạnh so sánh của chúng. Đề xuất, ở dạng đơn giản, có thể được mô tả như sau:

Đầu tiên, biên soạn SPD từ các nguồn dữ liệu cơ bản (giữ lại một nguồn dữ liệu phù hợp để sử dụng làm danh sách B trong thiết lập DSE) cũng có thuộc tính chấm điểm từng bản ghi về việc bạn có coi bản ghi đó là *chắc chắn* hay không (100% tự tin rằng nó thuộc về dân số) hoặc *có thể* (< 100% bản ghi chắc chắn thuộc về dân số nhưng có một số khả năng là như vậy). Ví dụ, một SPD có thể được biên soạn với 1.100 bản ghi trong đó 900 bản ghi được đánh dấu là *chắc chắn* và 200 bản ghi được đánh dấu là *có thể*.

Thứ hai, biên soạn ước tính dân số chuẩn bằng phương pháp DSE trong đó danh sách A là tập hợp con của SPD, trong đó tất cả các bản ghi được đánh dấu là *chắc chắn* và danh sách B liên quan đến nguồn dữ liệu đã bị loại khỏi quá trình biên soạn SPD gốc. Trong ví dụ của chúng tôi, ước tính dân số hiện có thể được biên soạn với danh sách B và danh sách A phù hợp chứa 900 bản ghi *chắc chắn* để có được ước tính dân số là 1.000.

Thứ ba, bổ sung các bản ghi trong danh sách A vào ước tính dân số chuẩn bằng cách sử dụng lựa chọn dựa trên xác suất các bản ghi được đánh dấu là *có thể* từ SPD. Điều này tạo ra một SPD mới hiện có thể được sử dụng để lập bảng chéo trong khi tổng hợp thành các ước tính dân số có thể được bảo vệ theo quan điểm phương pháp luận. Trong ví dụ của chúng tôi, danh sách A hiện có thể được bổ sung bằng cách chọn từ 200 bản ghi *có thể* trong SPD với xác suất 0,5 = (1.000 – 900) / (1.100 – 900). Trong thực tế, một số hệ thống tính điểm có thể được triển khai để đánh giá khả năng đưa các bản ghi *có thể có* vào SPD.

Phạm vi dân số ngầm định của SPD càng lớn (bao gồm 100% bản ghi đáng tin cậy) thì sự an tâm cho người dùng càng lớn. Khi phạm vi bao phủ tăng lên thì độ chính xác của ước tính cũng tăng lên, phạm vi sai lệch do vi phạm giả định lưu giữ đồng nhất cũng giảm xuống và số lượng bản ghi (có thuộc tính liên quan) cần quy kết cũng giảm xuống.

Bất kể phương pháp nào được ưa chuộng, vẫn còn hai thách thức chính liên quan đến việc đáp ứng nhu cầu ước tính dân số chi tiết; phân tách địa lý chi tiết và thành phần hộ gia đình. Thông tin địa chỉ trên các nguồn dữ liệu hành chính có thể đã lỗi thời hoặc không nhất quán với các nguồn dữ liệu khác và do đó không phải lúc nào cũng chính xác hoặc cập nhật. Thật khó để triển khai các quy tắc để chỉ định người vào vị trí địa lý chi tiết khi chất lượng thông tin địa chỉ trên các nguồn dữ liệu hành chính không đồng đều, không nhất quán và không ổn định.

Việc mở rộng bộ công cụ để giải quyết tình trạng phân loại sai miền như một phần của *phương pháp ước tính* có những thách thức khi xử lý số lượng nhỏ liên quan đến sự cố địa

lý chi tiết. Nhìn chung, chỉ có các mối quan hệ hộ gia đình một phần được ghi lại trên hệ thống dữ liệu hành chính khi liên quan đến hình thức thanh toán hoặc cứu trợ ở Ireland; ví dụ, cha mẹ nhận được khoản thanh toán trợ cấp nuôi con sẽ có mối quan hệ cha mẹ - con cái được ghi lại (lưu ý, ngay cả những mối quan hệ hành chính này cũng có thể không phản ánh các thỏa thuận sinh hoạt trong thế giới thực - có thể trường hợp trẻ không sống với cha mẹ được chỉ định và thực tế có thể sống với cha mẹ/người giám hộ khác). Dân số bao gồm sinh viên là một nhóm đặc biệt khó xác định về vị trí địa lý và thành phần hộ gia đình, rất khó để xác định liệu họ có đang cư trú tại một số hình thức nhà ở cho sinh viên hay nhà thuê hay với cha mẹ của họ dựa trên thông tin có sẵn trong hệ thống hành chính ở Ireland hay không.

Điều tra dân số truyền thống có lợi thế là thu thập trực tiếp nhiều thuộc tính của từng cá nhân trong dân số và có thể dễ dàng phổ biến số liệu thống kê chi tiết dựa trên các thuộc tính này ở các cấp độ địa lý được phân tách cao. Nếu Điều tra dân số truyền thống được thay thế bằng Điều tra dân số chủ yếu sử dụng dữ liệu hành chính, thì cũng cần cân nhắc đáng kể về cách đáp ứng tốt nhất nhu cầu của người dùng số liệu thống kê đó. Điều tra dân số truyền thống thường dành sự chú ý đặc biệt để đếm các nhóm dân số khó tiếp cận (ví dụ: người không có giấy tờ, người vô gia cư) và bất kỳ sự thay thế nào đối với mô hình truyền thống có thể sẽ yêu cầu sự chú ý đặc biệt tương tự đối với các nhóm khó tiếp cận.

Tóm lại, các tác giả tin rằng có thể biên soạn ước tính dân số từ nguồn dữ liệu hành chính mà không cần phải có hệ thống hành chính công được hỗ trợ bởi đăng ký dân số trung ương chất lượng cao. Công việc được thực hiện cho đến nay cho thấy khả năng này, tuy nhiên, cần phải có nhiều công việc hơn nữa trong việc phát triển các phương pháp tương ứng để giải quyết những thách thức nổi bật, đáng chú ý nhất là cung cấp chi tiết thống kê về thành phần hộ gia đình và địa lý.

Tài liệu tham khảo

1. Dunne J. The Irish Statistical System and the Emerging Census Opportunity. Statistical Journal of the IAOS. 2015; 31(3): 391-400. Available from: http://www.medra.org/ servlet/aliasResolver? alias= iospress&doi= 10.3233/SJI-150915.

2. UNECE. Register-based Statistics in the Nordic Countries. Review of Best Practices with Focus on Population and Social Statistics. United Nations; 2007. Available from: http:// www.unece.org /fileadmin /DAM/ stats/ publications/ Registerbased _statistics_in_ Nordic _countries.pdf.

3. Bengtsson T, Rönning SÅ. Overcoverage in the Total Population Register. In: Nordiskt Statistikermöte – Statistics in a changing world. Towards 2020 and beyond. Stockholm: Statistics Sweden; 2016, p. 12. Available from: https://www.scb.se/Upload/NSM2016/ theme1/Tor Bengtsson – Stina Åsling Rönning.pdf.

4. Dunne J, Zhang LC. A system of population estimates compiled from administrative data only. Journal of the Royal Statistical Society Series A: Statistics in Society. 2023 (June). Available from: https://rss.org.uk/RSS/media/File-library/Events/ Discussion meetings/Dunne_preprint_27-June-2023.pdf.

5. Zhang LC, Dunne J. Trimmed Dual System Estimation. In: Bohning D, van der Heijden PGM, Bunge J, editors. Capturerecapture methods for the Social and Medical Sciences. CRC press; 2018, pp. 237-258.

6. Bishop Y, Feinberg S, Holland P. Discrete Multivariate Analysis. Springer; 1975.

Đỗ Ngát (lược dịch)

Nguồn: To count or to estimate: A note on compiling population estimates from administrative data