DỮ LIỆU TỔNG HỢP

Giải pháp mới cho người dùng truy cập dữ liệu thống kê

Quản lý quyền truy cập của người dùng vào nguồn dữ liệu thu thập là một nhiệm vụ phức tạp của các cơ quan thống kê quốc gia (NSOs). Họ vừa phải duy trì tính bảo mật của người cung cấp thông tin, vừa phải chịu áp lực công bố những bộ dữ liệu chi tiết để đáp ứng nhu cầu ngày càng cao của người dùng.

Thông thường, các cơ quan thống kê quốc gia cung cấp cho những người dùng đáng tin cậy (chẳng hạn như các học giả, nhà nghiên cứu,…) quyền truy cập vào một số dữ liệu vi mô của cá nhân hoặc doanh nghiệp, đồng thời xuất bản các bảng thống kê tổng hợp cho những người dùng khác. Tuy nhiên, đây vẫn chưa phải là giải pháp hoàn hảo để quản lý quyền truy cập vào dữ liệu, vì nhiều người dùng sẽ không nhận được lượng thông tin chi tiết mà họ đang tìm kiếm, trong khi việc kiểm tra và quản lý người dùng đáng tin cậy lại mất nhiều thời gian mà chưa thực sự đảm bảo rằng họ có lạm dụng nguồn dữ liệu đó hay không.

Thật may, có một cách khác để cung cấp cho người dùng thông tin chi tiết, đó chính là cung cấp cho họ Dữ liệu tổng hợp. Dữ liệu tổng hợp có thể được mô phỏng theo cách có nhiều thuộc tính giống như tập dữ liệu gốc và cho phép trích xuất các kết quả và thông tin chi tiết giống nhau, nhưng với rủi ro tiết lộ thông tin về các cá nhân liên quan đến dữ liệu đó thấp hơn nhiều.

Nếu bạn tham gia vào việc quản lý quyền truy cập của người dùng vào số liệu thống kê chính thức và muốn có một giải pháp khác để giải quyết các tình huống khó xử về quyền truy cập dữ liệu, thì tài liệu ***"Synthetic Data for Official Statistics - A Starter Guide"*** – ***"Dữ liệu tổng hợp cho thống kê chính thức- Hướng dẫn cho người bắt đầu"*** sẽ cung cấp cho bạn những gì bạn cần thực hiện.

1. Tác giả:

"Synthetic Data for Official Statistics- A Starter Guide" – "Dữ liệu tổng hợp cho thống kê chính thức- Hướng dẫn cho người bắt đầu" được xuất bản vào Tháng 1- 2023. Tài liệu này được xây dựng dựa trên kết quả từ Dự án Dữ liệu tổng hợp của Nhóm cấp cao UNECE về Hiện đại hóa thống kê chính thức (UNECE High-Level Group on Modernisation of Official Statistics) (HLG-MOS) (2020-2021) và những nghiên cứu của Mạng lưới Tư duy Bầu Trời Xanh (Blue Skies Thinking Network) về dữ liệu tổng hợp được phê duyệt bởi HLG-MOS.

Các chuyên gia xây dựng tài liệu này đến từ các Cơ quan Thống kê quốc gia của

Canada, New Zealand, Hà Lan, Thụy Điển và một số trường đại học như: Endinburgh Napier University, Delft Technical University, Radboud University, bên cạnh đó là sự tham gia của Trung tâm nghiên cứu dữ liệu hành chính Scotland, Cơ quan doanh thu Canada.

2. Nội dung của tài liệu:

2.1. Kết cấu

Nội dung chính của tài liệu được chia ra thành 5 chương, cụ thể như sau: Chương 1: Giới thiệu; Chương 2: Tác dụng của dữ liệu tổng hợp; Chương 3: Các phương pháp tạo ra dữ liệu tổng hợp; Chương 4: Những cân nhắc khi phổ biến dữ liệu tổng hợp; Chương 5: Các thước đo tiện ích đánh giá dữ liệu tổng hợp.

2.2. Những nội dung cơ bản

2.2.1. Chương 1: Giới thiệu

Dữ liệu là một nguồn tài nguyên quý giá, cung cấp thông tin đầu vào quan trọng cho các nhà thống kê, nhà kinh tế học và nhà khoa học dữ liệu để tạo ra những hiểu biết chi tiết và kịp thời, nhằm đáp ứng nhu cầu thông tin của các bên liên quan. Các NSO vừa phải đáp ứng lời kêu gọi về tính mở và minh bạch của dữ liệu vừa phải cam kết bảo vệ tính bảo mật và quyền riêng tư của những dữ liệu ấy. Nhưng, mọi người đều phải thừa nhận rằng việc phổ biến thông tin vừa hữu ích vừa an toàn là điều khó có thể đạt được.

Rủi ro khi phổ biến dữ liệu là rủi ro hoặc khả năng xảy ra rủi ro khi phổ biến bất kỳ dữ liệu nào, bao gồm cả dữ liệu tổng hợp. Nếu kết quả của dữ liệu tổng hợp càng mô phỏng chặt chẽ kết quả dữ liệu gốc thì rủi ro phổ biến những thông tin nhạy cảm trong dữ liệu gốc càng cao, do vậy người ta thường phải cân bằng giữa những tiện ích mà dữ liệu tổng hợp mang lại và rủi ro phổ biến thông tin khi phát hành dữ liệu tổng hợp đó.

Khi nghiên cứu về các giải pháp truy cập dữ liệu, các NSO thường dựa vào khung Năm An Toàn (Desai et al. 2016) để có thể quản lý được những dữ liệu bảo mật, và những bộ dữ liệu tổng hợp chính là một trong những giải pháp truy cập được đánh giá cao trên thang điểm Dữ liệu An toàn.

Để đưa ra được những khẳng định như vậy, nhóm tác giả đã so sánh những giải pháp phổ biến dữ liệu thường thấy ở các NSO như Tệp giả, Trung tâm nghiên cứu dữ liệu, Tệp dữ liệu vi mô công cộng (PUMF), hay những Giải pháp truy cập từ xa. Với các giải pháp truyền thống này, việc cân bằng giữa tiện ích mà người dùng tin có được với rủi ro phổ biến dữ liệu là điều vô cùng khó khăn. Và dữ liệu tổng hợp chính là một chiến lược thay thế khả thi, tạo điều kiện thuận lợi cho việc truy cập dữ liệu, đặc biệt là trong các trường hợp liên quan đến dữ liệu nhạy cảm.

"Dữ liệu tổng hợp được định nghĩa là những dữ liệu được tạo ra một cách ngẫu nhiên, vừa có giá trị phân tích vừa duy trì kiểm soát phổ biến dữ liệu ở mức độ cao. Dữ liệu tổng hợp có nguồn gốc từ các phương pháp chỉnh sửa và quy nạp dữ liệu, và gần đây, nó đã trở nên phát triển hơn với những thành tựu của khoa học dữ liệu và khoa học máy tính, cũng giống như nỗ lực của các NSO nhằm chia sẻ dữ liệu cởi mở và minh bạch hơn."

2.2.2. Chương 2: Tác dụng của dữ liệu tổng hợp

Dữ liệu tổng hợp chính là một giải pháp có thể giải quyết được các vấn đề về phổ biến dữ

liệu thống kê mà các NSO phải đối mặt. Tuy nhiên, giá trị của dữ liệu tổng hợp sẽ thay đổi tùy theo bản chất của vấn đề gặp phải.

Chương này đề cập tới bốn tác dụng của dữ liệu tổng hợp, bao gồm: Phát hành dữ liệu vi mô tổng hợp cho cộng đồng; Phân tích thử nghiệm; Giáo dục và Thử nghiệm công nghệ. Với mỗi tác dụng, các tác giả đưa ra ví dụ thực tế của cơ quan thống kê các quốc gia: New Zealand, Canada, Scotland, Úc hay Văn phòng Thống kê Quốc gia Anh. Những ví dụ này giúp cho người đọc có thể dễ dàng hình dung ra cách thức mà các NSO đã sử dụng dữ liệu tổng hợp cho mục đích của họ như thế nào.

Một ví dụ (nói cách khác, chính là sự chia sẻ kinh nghiệm) về việc sử dụng dữ liệu tổng hợp để Phát hành dữ liệu vi mô tổng hợp cho cộng đồng của Cơ quan Thống kê New Zealand như sau: Họ công bố những dữ liệu chi tiết thông qua những Tệp hồ sơ đơn vị tổng hợp (SURF). SURF được tạo ra bởi một mô hình toán học dựa trên các dữ liệu gốc nhưng không giống với dữ liệu gốc.

Chẳng hạn, các tệp SURF được phát hành năm 2007 dựa trên dữ liệu của cuộc điều tra về thu nhập từ quý 2 năm 2003. Mỗi tệp này chứa hơn 11.000 hồ sơ và có 100 tệp đã được công bố vào năm 2007, đại diện cho 100 mẫu của dân số New Zealand- những người có việc làm và được trả lương trong độ tuổi 25 đến 64. Các biến bao gồm tuổi tác, giới tính, dân tộc, trình độ học vấn cao nhất, số giờ làm việc hàng tuần và thu nhập hàng tuần. Tuy nhiên, những dữ liệu được công bố thường là các dữ liệu dân số đã bị thêm "nhiễu". Do đó, quyền riêng tư của dữ liệu được đảm bảo ngay cả khi một số dữ liệu tổng hợp giống với dữ liệu gốc. Và khi phát hành các tệp SURF, sẽ có khuyến nghị rõ ràng từ cơ quan thống kê quốc gia cho mục đích sử dụng của người dùng.

Bảng 1. Tổng hợp các trường hợp sử dụng, những lưu ý quan trọng và mức cân bằng giữa tiện ích với bảo mật

Trường hợp sử dụng	Lưu ý	Sự cân bằng giữa tiện ích và bảo mật
Phát hành dữ liệu vi mô tổng hợp cho cộng đồng	Những người tổng hợp dữ liệu sẽ không biết ai sử dụng và sử dụng dữ liệu đó như thế nào	Tính bảo mật cao và tiện ích cũng cao
Phân tích thử nghiệm	Những phân tích cụ thể hoặc phân phối các biến bắt buộc phải duy trì *có thể* bị nhận diện tại thời điểm tổng hợp	Tính bảo mật cao và tiện ích cũng cao
Giáo dục	Những người tổng hợp *có thể* biết phân tích sẽ thực hiện và người dùng *có thể* có giấy tờ hoặc thỏa thuận bảo mật với NSO, tuy nhiên điều ngược lại cũng có thể đúng	Tiện ích cao với từng mức độ bảo mật khác nhau
Thử nghiệm công nghệ	Giá trị của dữ liệu tổng hợp phụ thuộc vào mức độ phức tạp của hệ thống và mức độ phức tạp của dữ liệu thử nghiệm. Nhiều phương pháp tạo dữ liệu tổng hợp có thể quá nặng về mặt tính toán	Tiện ích trung bình và bảo mật trung bình

Tương tự như vậy, với các tác dụng của dữ liệu tổng hợp, người đọc sẽ được tiếp cận với kinh nghiệm thực tế được chia sẻ từ các NSO. Và ở cuối chương, các tác giả đã đưa ra một bảng tổng hợp để người đọc có những đánh giá tổng quan về 4 cách sử dụng dữ liệu tổng hợp, những lưu ý khi sử dụng dữ liệu tổng hợp tùy theo mục đích, và sự cân bằng giữa tiện ích và tính bảo mật của 4 cách sử dụng đó. Các tác giả gọi đây là "Mẹo để bắt đầu"!

Việc đúc kết và đưa ra đánh giá khái quát về những tiện ích cũng như tính bảo mật của các trường hợp sử dụng dữ liệu tổng hợp như trên là một điểm sáng của tài liệu, cho dù người đọc là người đã có nhiều năm kinh nghiệm hay là người mới tiếp xúc với công việc, khi đọc bảng tổng hợp trên họ cũng hình dung được một cách dễ dàng, nhanh chóng. Và đặc biệt hơn, cách mà các tác giả đặt tên là "Mẹo để bắt đầu" rất thu hút và tạo nên sự chú ý cho người đọc. Từ các chương sau trở đi, mỗi chương đều có mục này!

2.2.3. Chương 3: Các phương pháp tạo ra dữ liệu tổng hợp

Có nhiều phương pháp tạo dữ liệu tổng hợp, nhưng để xác định nên sử dụng phương pháp nào thì trước hết phải xác định được loại dữ liệu tổng hợp được yêu cầu là gì và chúng sẽ được sử dụng cho mục đích nào.

Chương này nhằm mục đích cung cấp cái nhìn tổng quát về các phương pháp đã có sẵn để tạo ra dữ liệu tổng hợp và đưa ra các khuyến nghị để người đọc có thể lựa chọn được phương pháp phù hợp nhất. Các phương pháp được chia ra thành 3 nhóm:

- Mô hình tuần tự
- Dữ liệu mô phỏng
- Phương pháp học sâu

Mục tiêu chính của chương này là làm nổi bật khả năng ứng dụng của các phương pháp đã được các NSO tìm ra, đồng thời phác thảo những ưu điểm và nhược điểm của chúng.

Điểm qua về kết cấu của chương, mỗi nhóm phương pháp sẽ đưa ra từ 1-2 phương pháp sau đó mô tả các phương pháp này; nêu ưu, nhược điểm và những lưu ý khi sử dụng; các công cụ thực hiện và ví dụ thực tế của một số NSO. Ngoài ra, người đọc còn có thể tiếp cận các kiến thức một cách kỹ lưỡng hơn qua tài liệu tham khảo mà các tác giả đã chú thích ở cuối mỗi trang và cuối tài liệu.

Dưới đây là một đoạn trích dẫn về nhóm "Phương pháp học sâu" trong tài liệu:

"Học sâu là một tập hợp con của học máy và là một thể loại đang phát triển trong lĩnh vực Khoa học dữ liệu và Trí tuệ nhân tạo. Các phương pháp này đang trở nên phổ biến hơn trong lĩnh vực dữ liệu tổng hợp vì các nhà tổng hợp đang xử lý ngày càng nhiều các tập dữ liệu lớn. Tại thời điểm viết hướng dẫn này, Mạng đối nghịch phát sinh (GAN) là một phương pháp học sâu được các NSO sử dụng để tạo dữ liệu tổng hợp.

Với những cải tiến về công nghệ và khả năng tính toán, việc triển khai các quy trình học máy đã trở nên dễ dàng và dễ tiếp cận hơn. Do đó, các phương pháp học máy ngày càng được sử dụng nhiều hơn để tạo ra các tệp dữ liệu tổng hợp. Cụ thể hơn, việc sử dụng các mô hình học sâu đã trở nên hấp dẫn vì các mô hình dự đoán dựa trên các tệp dữ liệu lớn có khả năng phát triển mạnh.

3.3.1. Mạng đối nghịch phát sinh (GAN)

Mạng đối nghịch phát sinh (GAN) (Goodfellow et al., 2014) là một mô hình nổi bật được sử dụng để tạo ra dữ liệu tổng hợp. Mô hình này cố gắng tìm hiểu cấu trúc cơ bản của dữ liệu gốc bằng cách tạo dữ liệu mới (cụ thể hơn là các mẫu mới) từ cùng một phân phối thống kê giống như dữ liệu gốc, với hai mạng thần kinh cạnh tranh nhau trong một trò chơi. Bởi vì GAN dựa trên các mạng thần kinh, điều đó có nghĩa là phương pháp này có thể được sử dụng để tạo dữ liệu tổng hợp rời rạc, liên tục hoặc dạng văn bản.

Trong một GAN có hai mô hình mạng thần kinh cạnh tranh:

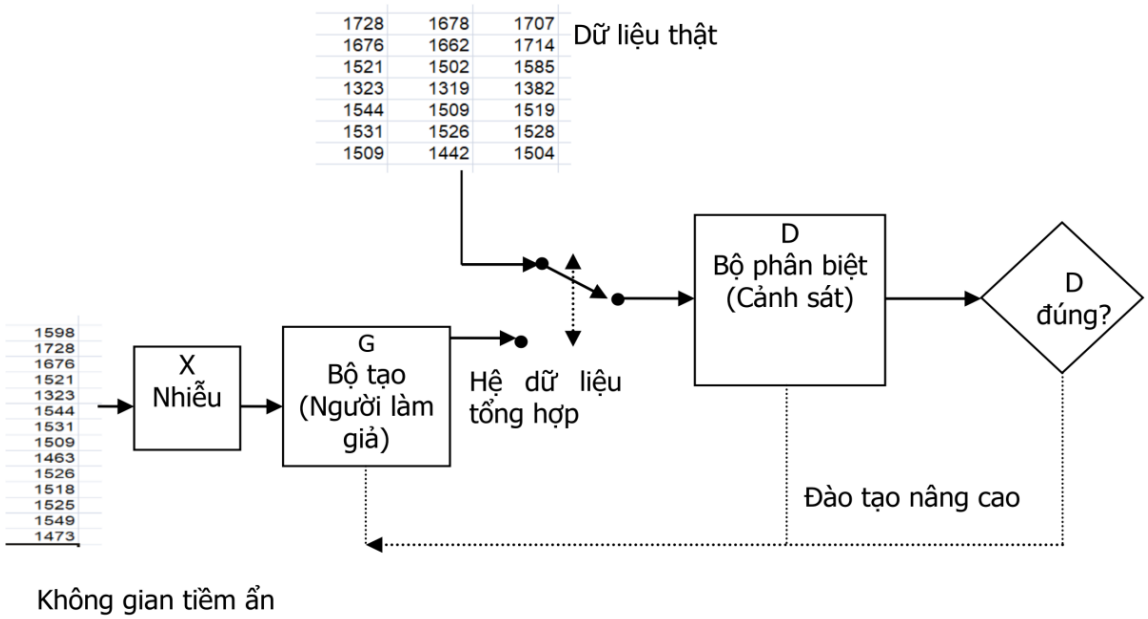
- Một được gọi là bộ tạo và lấy nhiễu (hoặc giá trị ngẫu nhiên) làm đầu vào và tạo mẫu;

- Hai là bộ phân biệt, nhận các mẫu từ cả bộ tạo và bộ dữ liệu huấn luyện cố gắng phân biệt hai nguồn.

Bộ phân biệt phục vụ chức năng tương tự như bộ phân loại nhị phân, sẽ lấy cả dữ liệu thực (hoặc gốc) cũng như dữ liệu được tạo (hoặc tổng hợp) làm đầu vào và sẽ tính toán giá trị xác suất giả sẽ được so sánh với giá trị ngưỡng cố định để phân loại đầu vào từ bộ tạo là được tạo ra hay là dữ liệu thật.

Như được hiển thị trong Hình 4 của tài liệu, quá trình tạo là một quá trình lặp đi lặp lại, trong đó hai mạng chơi một trò chơi đang diễn ra mà bộ tạo và lấy nhiễu đang học cách tạo mẫu thực tế hơn, trong khi đó, bộ phân biệt đang học để phân biệt một cách tốt hơn những dữ liệu được tạo với dữ liệu thật. Sự tương tác giữa hai mạng này là cần thiết cho sự thành công của GAN vì cả hai mạng đều học hỏi lẫn nhau, và cuối cùng đạt được trạng thái cân bằng.

Hình 4. Minh họa đào tạo GAN



Nguồn: Kaloskampis et al. (2020)"...

Có thể hiểu một cách đơn giản từ sơ đồ trên như sau:

Bài toán đặt ra là dùng GAN để tạo ra tiền giả, còn dữ liệu thật chính là tiền thật. Bộ tạo và lấy nhiễu (Generator- G) giống như người làm tiền giả, còn bộ phân biệt (Discriminator-D) giống như cảnh sát. Người làm tiền giả sẽ cố gắng làm ra những đồng tiền giả giống thật nhất có thể để cảnh sát không thể phân biệt được. Còn cảnh sát sẽ cố gắng phân biệt đâu là tiền thật và đâu là tiền giả. Mục tiêu cuối cùng là người làm tiền giả sẽ làm ra được tiền mà cảnh sát không thể phân biệt được đâu là thật, đâu là giả, và tiền giả sẽ được mang đi để tiêu dùng. Trong quá trình huấn luyện GAN thì cảnh sát có 2 việc: một, là học cách phân biệt giả/thật; hai, là nói cho người làm giả biết tiền mà người đó làm ra vẫn chưa thể qua mắt được và cần cải thiện nhiều hơn. Từ đó, người làm giả sẽ dần dần làm tiền giả giống tiền thật nhiều nhất có thể, còn cảnh sát cũng sẽ thành thạo hơn trong việc phân biệt.

Cách trình bày của các tác giả giúp những người đầu tiên tiếp xúc với các kiến thức này cũng có được cái nhìn tổng quát về các phương pháp được sử dụng để tạo dữ liệu tổng hợp. Với tất cả các phương pháp được đưa ra, các tác giả cố gắng giải thích một cách chi tiết nhất có thể. Đặc biệt, cách sử dụng những sơ đồ, hay đưa các ví dụ thực tế vào tài liệu khiến cho người đọc cảm thấy hài lòng.

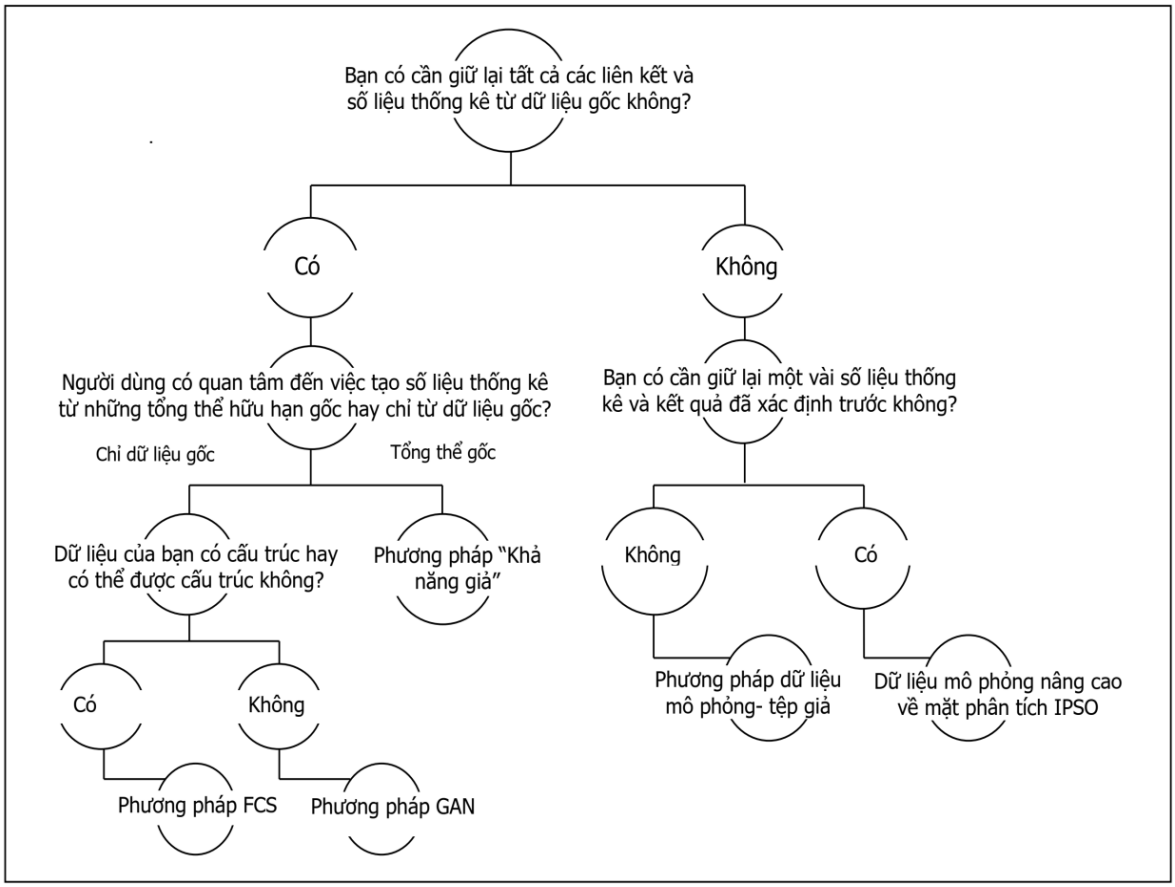
Điều thú vị ở đây là, các chương đều có mối liên hệ chặt chẽ với nhau, tài liệu này được các tác giả viết dưới dạng từ tổng quan đến chi tiết, tức là, những chương sau, mặc dù nội dung có vẻ như được viết riêng biệt, nhưng thực tế, những nội dung đó lại liên kết với nội dung của chương trước. Ví dụ, trong mục "3.5. Mẹo để bắt đầu" của chương 3, mục này cung cấp những khuyến nghị về các phương pháp tạo dữ liệu tổng hợp, dựa trên những trường hợp sử dụng dữ liệu tổng hợp mà trước đó đã trình bày ở chương 2. Cách viết này giúp người đọc hệ thống lại các kiến thức quan trọng của từng chương mà không cần phải mất nhiều thời gian để xem lại.

Dưới đây là một phần trích dẫn của "Bảng 7. Đề xuất phương pháp theo trường hợp sử dụng":

		Các trường hợp sử dụng			
Phương pháp		**Phát hành dữ liệu vi mô tổng hợp cho công chúng/ Phân tích thử nghiệm**	**Giáo dục**	**Công nghệ thử nghiệm**	**Lưu ý**
Phương pháp học sâu	Mạng đối nghịch phát sinh (GAN)	Khuyến khích sử dụng khi có dữ liệu dạng văn bản hoặc dữ liệu phi cấu trúc	Có thể được sử dụng, mặc dù nếu các phân tích được thực hiện và kết luận thống kê được xác định trước, nó có thể tốn thời gian hơn nhiều so với các phương pháp khác	Có thể được sử dụng nhưng phương pháp này có lẽ quá tiên tiến so với nhu cầu phân tích thực tế	Về mặt lý thuyết, phương pháp này nhằm mục đích duy trì mối quan hệ giữa các biến từ dữ liệu gốc. Đây là phương pháp duy nhất xử lý dữ liệu dạng văn bản và phi cấu trúc. Rủi ro phổ biến thông tin và giá trị phân tích cần được đánh giá theo quy trình

Ngoài ra, ở cuối chương này, vẫn nằm trong mục "3.5. Mẹo để bắt đầu", các tác giả đưa ra một sơ đồ hình cây, gọi là "Cây quyết định" để giúp người đọc xác định được phương pháp nào phù hợp nhất với mục đích của người tạo dữ liệu tổng hợp.

Hình 6. Cây quyết định phương pháp tạo dữ liệu tổng hợp



Sơ đồ này vừa là khuyến nghị của các tác giả, đồng thời cũng thể hiện mục đích hệ thống lại những kiến thức mà các tác giả đã viết ở chương này.

2.2.4. Chương 4: Những cân nhắc khi phổ biến dữ liệu tổng hợp

Chương 4 và chương 5 là hai chương cuối cùng của tài liệu, và theo ý kiến cá nhân của tôi, hai chương này là hai chương có nội dung phức tạp nhất, đặc biệt là chương 4. Để hiểu được nội dung của chương này, người đọc cần có kiến thức về toán học, khoa học dữ liệu và khoa học máy tính. Tuy nhiên, đối với người đọc nếu chưa có khả năng hiểu hết được những nội dung trong chương này, thì các tác giả cũng đã cung cấp rất nhiều tài liệu tham khảo liên quan để người đọc nghiên cứu sâu hơn.

Ở phần Giới thiệu của tài liệu, các tác giả đã viết: *"Mọi người đều thừa nhận rằng việc phổ biến thông tin vừa hữu ích vừa an toàn là không thể đạt được như ý muốn".*

Văn phòng Ủy viên Quyền riêng tư của Canada viết: *"Cần lưu ý rằng không có rủi ro [phổ biến] bằng không khi phát hành dữ liệu"*.

Khi dữ liệu được thay đổi để làm giảm khả năng bị phổ biến bằng cách tạo dữ liệu tổng hợp hoặc bằng các phương pháp khác thì sự đánh đổi giữa những rủi ro phổ biến dữ liệu với những tiện ích luôn là vấn đề cần phải thảo luận. Làm sao để bảo vệ quyền riêng tư của dữ liệu tổng hợp mà vẫn đáp ứng được nhu cầu của người dùng? - Đáp án nằm trong Chương 4.

Tiện ích mà dữ liệu tổng hợp mang lại chính là thước đo mức độ gần gũi của kết quả có được từ dữ liệu tổng hợp so với dữ liệu gốc. Và, như đã nói ở trên, sẽ chẳng bao giờ đạt được dữ liệu tổng hợp có tiện ích bằng 100 % mà không có rủi ro phổ biến, bởi vì việc thay đổi dữ liệu để bảo vệ khỏi rủi ro phổ biến luôn thay đổi giá trị của dữ liệu gốc. Trong khi đó, việc phát triển các thước đo rủi ro phổ biến dữ liệu tổng hợp thường khó khăn hơn nhiều so với các thước đo tiện ích dữ liệu.

Nội dung của chương này được chia làm các phần sau đây:

(1) Các biện pháp rủi ro phổ biến thông tin cho dữ liệu tổng hợp

Có hai loại rủi ro phổ biến chính: Phổ biến danh tính và Phổ biến thuộc tính.

Phổ biến danh tính: Phổ biến danh tính xảy ra khi một hồ sơ trong dữ liệu được phổ biến và được nhận dạng khớp với một cá nhân mà kẻ tấn công biết giá trị của dữ liệu được phát hành từ một nguồn khác. Và phổ biến danh tính chỉ liên quan đến dữ liệu tổng hợp một phần vì các hồ sơ trong dữ liệu tổng hợp đầy đủ sẽ không khớp trực tiếp với dữ liệu gốc. Việc phổ biến danh tính khó đo lường hơn, đặc biệt với bộ dữ liệu tổng hợp hoàn toàn không có hồ sơ nào là thật. Có một phương pháp đã được một NSO sử dụng để đánh giá việc phổ biến danh tính đó là tỷ lệ liên quan đến tái tạo cơ sở dữ liệu. Các tỷ lệ này đề cập đến tỷ lệ phần trăm trùng khớp mà một người nhận được từ dữ liệu tổng hợp so với dữ liệu gốc; với mục đích xác định dữ liệu gốc có thể được tạo ra dễ dàng như thế nào với thông tin có sẵn đã phát hành cho công chúng. Thế nhưng, tỷ lệ liên quan đến việc xây dựng lại những dữ liệu mà liên quan đến phổ biến danh tính này thực tế vẫn là ngụ ý về khả năng phổ biến thuộc tính.

Phổ biến thuộc tính: Phổ biến thuộc tính từ dữ liệu tổng hợp đầy đủ hoặc một phần đều có thể xảy ra mà không cần phổ biến danh tính. Kịch bản mà các tác giả hình dung là kẻ tấn công quan sát thấy một cá nhân trong dữ liệu tổng hợp có vẻ trùng khớp duy nhất với một cá nhân đã biết và chúng có thể sử dụng dữ liệu tổng hợp để lấy thông tin về một hoặc nhiều giá trị tổng hợp của họ. Các biện pháp phổ biến thuộc tính được đưa ra ở đây dựa trên ý tưởng về giá trị duy nhất được sao chép. Tức là, các giá trị được xác định từ tất cả các biến hoặc từ một tập hợp con đã được chọn trước đó được kết hợp với nhau và được sao chép thành một hồ sơ duy nhất trong dữ liệu tổng hợp đồng thời đó cũng là hồ sơ duy nhất trong dữ liệu gốc. Đây được gọi là các biến chính hoặc các biến định danh bán đại diện cho thông tin mà kẻ tấn công có thể sở hữu về một cá nhân. Sau đó, kẻ tấn công có thể cho rằng đây là cá nhân thực sự có những đặc điểm này, và tiến hành tìm giá trị của các mục khác trong hồ sơ của người đó. Tất nhiên, các giá trị như vậy có thể sai nhưng việc công khai kiểu dữ liệu như vậy lại có thể gây bất lợi cho NSO. Do đó, NSO có

thể thực hiện xóa các hồ sơ duy nhất hoặc giảm thiểu tỷ lệ các hồ sơ như vậy có trong dữ liệu tổng hợp, và trong thực tế, việc loại bỏ các hồ sơ duy nhất cũng chỉ ảnh hưởng rất nhỏ đến tiện ích của dữ liệu.

(*) Kỹ thuật bảo vệ quyền riêng tư:

Các kỹ thuật bảo vệ quyền riêng tư có thể được áp dụng cả trên dữ liệu gốc hoặc dữ liệu tổng hợp. Các NSO có thể chọn áp dụng các kỹ thuật bảo vệ quyền riêng tư cho dữ liệu tổng hợp căn cứ vào bối cảnh quốc gia, khuôn khổ luật pháp và hoạt động của họ. Tài liệu này giới thiệu về các kỹ thuật bảo vệ quyền riêng tư, bao gồm: Quyền riêng tư khác biệt (DP) (Differnential Privacy), K- Ẩn danh (K- Anonymity), ℓ- Đa dạng (ℓ- Diversity), t-Gần gũi (t- Closeness).

Trong đó; kỹ thuật DP là loại kỹ thuật được các tác giả mô tả chi tiết nhất. Được đánh giá là phương pháp quan trọng và hữu ích nhất trong các phương pháp nên DP đã được các tác giả mô tả và trình bày rất cụ thể, nội dung liên quan đến phương pháp này chiếm dung lượng khoảng 50% nội dung của chương 4. Ở đây tôi xin trích dẫn và dịch lại một vài ý chính mà các tác giả giới thiệu về phương pháp này để bạn đọc có thể hình dung bước đầu về DP:

"Quyền riêng tư khác biệt (DP) được giới thiệu vào năm 2006 trong khoa học máy tính bởi Cynthia Dwork, Frank McSherry, Kobbi Nissim và Adam Smith."

"... Mục đích sử dụng của DP trong số liệu thống kê chính thức là ngăn chặn phổ biến khi công bố thông tin thống kê thay vì giải quyết các mối lo ngại về quyền riêng tư khi thu thập dữ liệu cá nhân từ các cá nhân. Do đó, DP tự thể hiện rằng mình như một giải pháp thay thế cho khuôn khổ kiểm soát phát hành thông tin truyền thống đã được các NSO sử dụng trong nhiều thập kỷ nay.

Hai loại chính của DP là: ε-DP một tham số, DP- hai tham số yếu hơn (ε-δ). Tài liệu này tập trung vào DP một tham số, đây là dạng mà mọi người hay nghĩ đến khi nhắc tới DP. DP có thể được người quản lý dữ liệu sử dụng để cung cấp các đảm bảo về bảo vệ phổ biến thông tin và có thể chứng minh bằng toán học. Khi phổ biến thông tin thống kê thu được từ cá nhân, DP là khung kiểm soát công bố thông tin đầu tiên nêu rõ loại và mức độ bảo vệ mà nó cung cấp".

Những nội dung tiếp theo, các tác giả tập trung vào mô tả phương pháp này gồm: Một phương pháp tuân thủ DPε là gì? Tại sao lại gọi là Quyền riêng tư khác biệt? Những yêu cầu toán học bên dưới Quyền riêng tư khác biệt chính xác là những gì? Những đảm bảo bảo vệ phổ biến chính thức có nghĩa là gì? Một phương pháp tuân thủ DPε sẽ như thế nào? Liệu một phương pháp tuân thủ DPε sẽ luôn luôn tốt hơn phương pháp không tuân thủ DP? 2 ứng dụng của phương pháp tuân thủ DPε...

Bên cạnh các giải thích và mô tả phương pháp DP, các tác giả cũng giới thiệu khái quát về 3 phương pháp đó là: K- Ẩn danh (K- Anonymity), ℓ- Đa dạng (ℓ- Diversity), t-Gần gũi (t- Closeness) ở phần sau, tuy nhiên hàm lượng nội dung dành cho 3 phương pháp này chỉ dừng lại ở mức thấp hơn nhiều so với phương pháp DP, nhưng những ví dụ hay những lưu ý khi thực hiện 3 phương pháp này vẫn được viết đầy đủ và cụ thể.

(*) Các phương pháp đánh giá phổ biến thông tin khác:

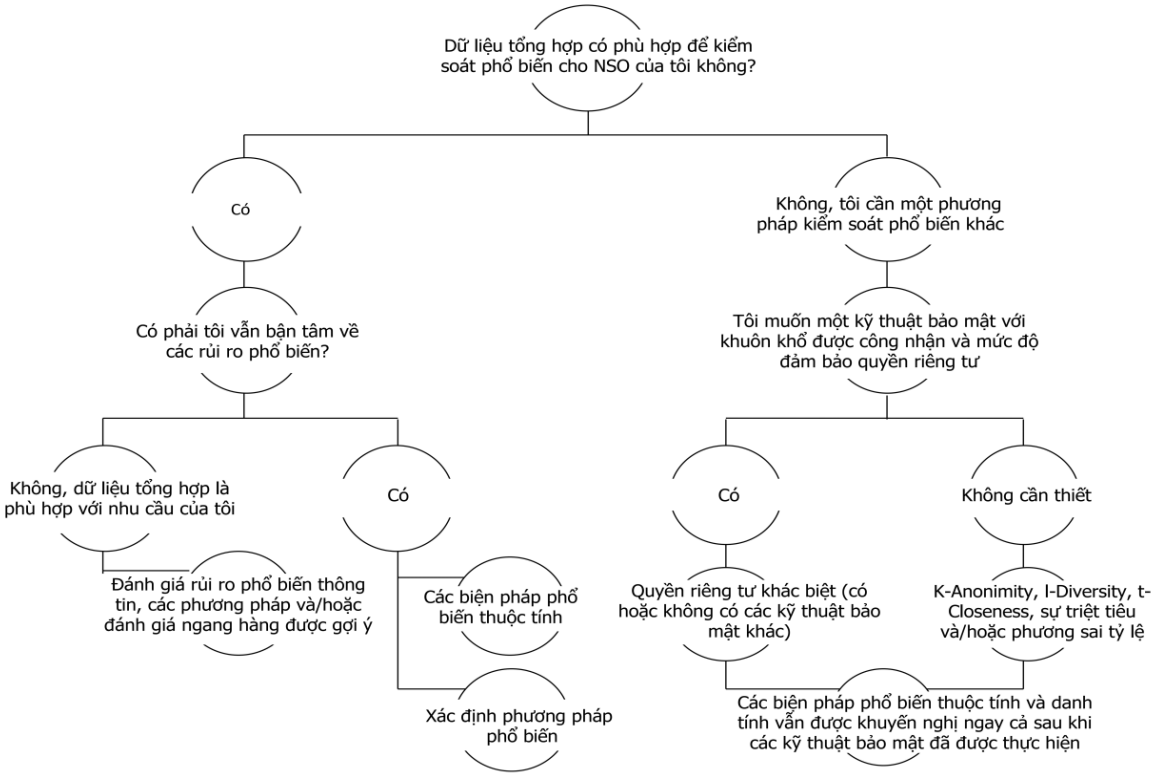
Phần này bao gồm 2 phương pháp: Đánh giá ngang hàng và Phương sai tỷ lệ.

Cũng như những phương pháp được trình bày ở trên, các tác giả mô tả khái quát sau đó đưa ví dụ thực tế của Cơ quan Thống kê New Zealand và 1 số bảng dữ liệu mô tả cách thực hiện phương pháp Phương sai tỷ lệ.

(*) Mẹo để bắt đầu:

Và cũng như các chương trước, một sơ đồ hình cây được các tác giả trình bày giúp người đọc hệ thống lại được toàn bộ nội dung của chương. Căn cứ vào bối cảnh, khuôn khổ hoạt động và tính pháp lý của từng quốc gia, các NSO có thể lựa chọn biện pháp và kỹ thuật phổ biến để áp dụng phù hợp với dữ liệu tổng hợp của họ.

Hình 10. Cây quyết định hỗ trợ xem xét kiểm soát phổ biến thông tin



Nếu bạn quyết định triển khai các kỹ thuật bảo vệ quyền riêng tư bổ sung, các biện pháp phổ biến thuộc tính và phổ biến danh tính vẫn hữu ích để đánh giá hiệu quả của các kỹ thuật. Bất kỳ biện pháp hoặc kỹ thuật phổ biến nào khi thực hiện cũng phải tính đến tiện ích của dữ liệu. Các phương pháp đo lường và đánh giá tiện ích của dữ liệu tổng hợp, có nên hay không nên có các kỹ thuật bảo mật bổ sung sẽ được trình bày ở chương tiếp theo.

2.2.5. Chương 5: Các thước đo tiện ích đánh giá dữ liệu tổng hợp

Tiện ích, hay nói cách khác, chính là giá trị của tệp dữ liệu tổng hợp, phản ánh mức độ hữu ích của tập dữ liệu đó đối với mục đích hoặc trường hợp sử dụng dữ liệu. Như

đã thảo luận trong chương hai, dữ liệu tổng hợp thường được sử dụng thay cho dữ liệu gốc hoặc dưới dạng phân tích sơ bộ, để dẫn đến kết quả cuối cùng sẽ chạy trên dữ liệu gốc. Trong cả hai trường hợp, tiện ích của dữ liệu tổng hợp dựa trên mức độ giống nhau của các kết luận giữa dữ liệu tổng hợp và dữ liệu gốc. Điều này cũng quan trọng không kém trong trường hợp sử dụng sau vì phân tích sự bộ vì dữ liệu tổng hợp sẽ hướng dẫn các mô hình cuối cùng được sử dụng.

Trong chương 3, các phương pháp tạo dữ liệu tổng hợp nên phụ thuộc vào trường hợp sử dụng, nội dung mà người tổng hợp muốn lưu giữ và dạng dữ liệu gốc. Khi bộ dữ liệu tổng hợp đã được tạo, tiện ích có thể được đánh giá.

Có hai loại thước đo tiện ích chính: thước đo tiện ích "chung" và thước đo tiện ích "cụ thể". Các thước đo cụ thể sẽ hữu ích khi đánh giá một phân tích cụ thể, tuy nhiên chúng lại không hữu ích cho các phương pháp tổng hợp điều chỉnh- sửa đổi dữ liệu để cải thiện tiện ích, vì bộ tổng hợp thường không biết phân tích sẽ được tiến hành bằng cách sử dụng dữ liệu tổng hợp khi quá trình điều chỉnh diễn ra.

Có hai lý do chính để đánh giá tiện ích chung của dữ liệu tổng hợp:

- So sánh các phương pháp tổng hợp khác nhau cho cùng một tập dữ liệu để tạo ra tập dữ liệu tổng hợp hữu ích nhất cho người dùng.

- Xác định nơi phân phối dữ liệu gốc và dữ liệu tổng hợp khác nhau và từ đó điều chỉnh các phương pháp tổng hợp để cải thiện tiện ích của dữ liệu tổng hợp.

Các tác giả đã liệt kê các phương pháp đánh giá như bảng dưới đây:

Bảng 20. Tổng hợp các thước đo tiện ích

Phương pháp	Thước đo hay các phương tiện thực hiện	Tài liệu tham khảo	Viết tắt
Cụ thể	Khoảng tin cậy chồng chéo	Karr và cộng sự (2006)	
	Đạt kết quả cụ thể	Kaloskampis và cộng sự (2020) Jordan và cộng sự (2018) Slokom và cộng sự (2021)	
Chung	Điểm xu hướng trung bình hay chính là sai số bình phương	Karr và cộng sự (2006) Woo và cộng sự (2009)	pMSF
Thước đo đơn từ điểm xu hướng	Thống kê Kolmogorov- Smirnov so sánh điểm xu hướng cho dữ liệu gốc và dữ liệu tổng hợp	Bowen và cộng sự (2021)	SPECKS
	Các so sánh khác về điểm xu hướng cho dữ liệu gốc và tổng hợp	Raab và cộng sự (2021)	U
	Tỷ lệ phần trăm trên 50% hồ sơ kết hợp được dự đoán chính xác bởi điểm xu hướng	Raab và cộng sự (2021)	PO50

Chung Thước đo đơn từ các bảng	Thống kê Voas-Williamson	Voas và Williamson (2001)	VW
	Freeman-Tukey (FT)	Voas và Williamson (2001)	FR
	Thống kê tỷ lệ khả năng từ các bảng và các nhân tố khác của gia đình phân kỳ	Voas và Williamson (2001)	G
	Phân kỳ Jensen-Shannon	Fuglede và Topsoe (2004)	JSD
	Số liệu Bhattacharyya	Bhattacharyya (1943)	dBhatt
	Độ lệch tuyệt đối trung bình về mật độ	Raab và cộng sự (2021)	WabsDD
	Sự khác biệt của ma trận tương quan	Kaloskampis và cộng sự (2019)	
	Độ lệch tuyệt đối trung bình có trọng số về mật độ	Raab và cộng sự (2021)	WMabsDD
Các phương pháp khám phá tiện ích	So sánh biểu đồ bằng các biện pháp trực quan và tóm tắt	http://www.synthpop.org.uk Kaloskampis và cộng sự (2020)	
	So sánh bảng chéo từ phân phối cận biên	Raab (2011) và NIST (2021)	
	So sánh và trực quan hóa các số liệu thống kê tóm tắt khác (ví dụ: Tương quan Pearson)	Beaulieu-Jones và cộng sự (2019) Kaloskampis và cộng sự (2020)	

Bảng trên cung cấp tóm tắt các thước đo được thảo luận trong chương 5 và các thước đo này được phân ra thành các nhóm như sau: Thước đo tiện ích cụ thể; Thước đo tiện ích chung theo điểm xu hướng hoặc theo bảng; và Thước đo khám phá tiện ích.

Nhóm các Thước đo tiện ích cụ thể: Các thước đo này so sánh kết quả của các mô hình thống kê phù hợp với dữ liệu tổng hợp và dữ liệu gốc. Bất kỳ một phân tích thống kê nào cũng có thể được sử dụng để tạo thước đo tiện ích, chẳng hạn như: sự khác biệt về giá trị trung bình của các biến, sự khác biệt về mối tương quan, bảng và bảng chéo, thậm chí phân tích kết quả hoặc đầu ra. Cách sử dụng rộng rãi nhất trong số này là khoảng tin cậy chồng chéo, nó cung cấp cả phép đo tóm tắt và trực quan hóa kết quả của mô hình thống kê từ hai nguồn dữ liệu. Phần này các tác giả giới thiệu về 2 thước đo: *Khoảng tin cậy chồng chéo và Đạt kết quả cụ thể*.

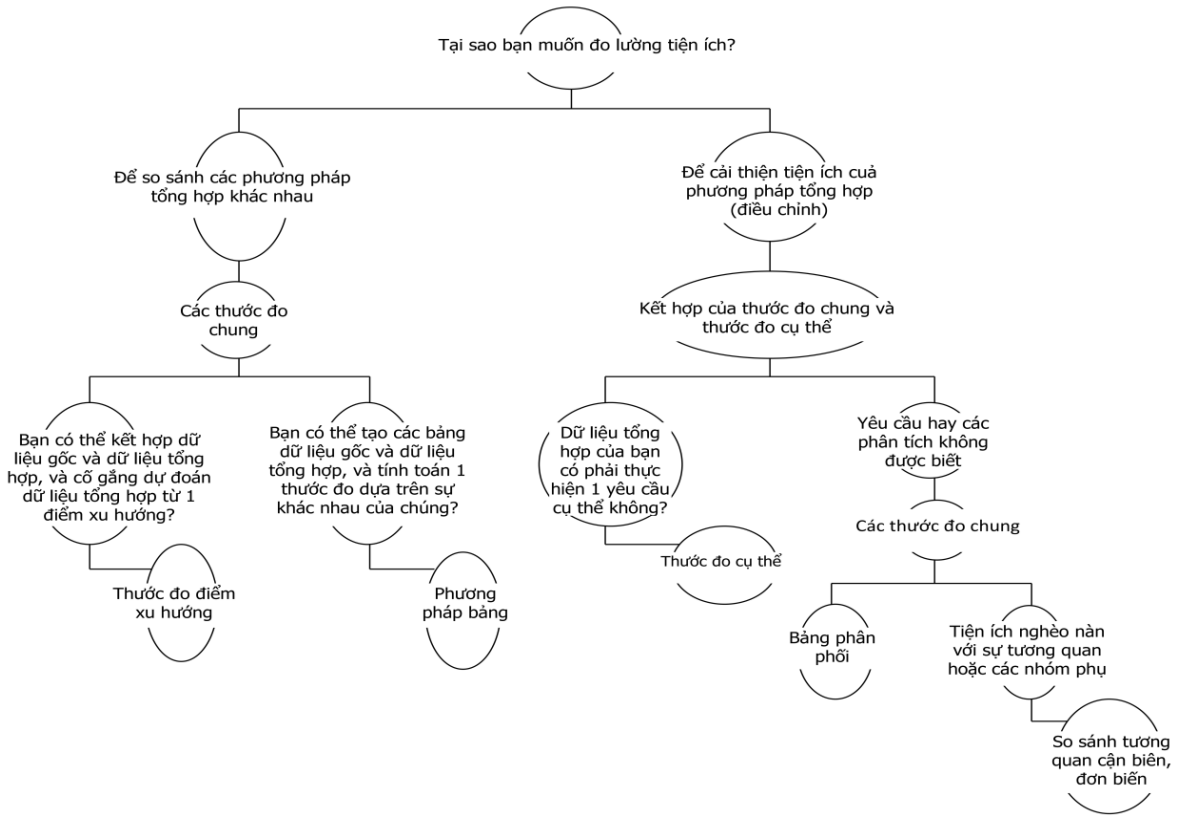
Nhóm các Thước đo tiện ích chung: Các thước đo mà chúng so sánh toàn bộ phân phối của dữ liệu tổng hợp với dữ liệu gốc này, được gọi là các thước đo tiện ích chung. Những thước đo này chủ yếu dựa trên hai phương pháp, thứ nhất là kết hợp dữ liệu gốc và dữ liệu tổng hợp đồng thời tính toán điểm xu hướng, xác suất mà bất kỳ hồ sơ nào là tổng hợp và thứ hai là bằng cách so sánh các bản dữ liệu gốc và dữ liệu tổng hợp. Hai phương pháp này có thể được coi là giống

nhau nhưng việc so sánh các bảng cho phép một số thứơc đo bổ sung mà phương pháp điểm xu hướng không có được. Trong phần này, các tác giả có phân tích mối quan hệ giữa các thước đo, một số thước đo trong Bảng 20 có liên quan tuyến tính. Cụ thể, VW và pMSE có cùng thước đo, hay SPECKS, PO50 và MabsDD cũng cùng thước đo. Các nghiên cứu thực nghiệm cho thấy rằng tất cả các thước đo đều tương quan với nhau khi so sánh với các bộ dữ liệu tổng hợp khác nhau. Việc mở rộng các thước đo tiện ích hay xác định các mô hình cho điểm xu hướng cũng được các tác giả đề cập trong nội dung này.

Nhóm các Thước đo khám phá tiện ích: Có 3 thước đo được giới thiệu ở đây, đó là So sánh đơn biến; So sánh cận biên và So sánh các số liệu thống kê khác.

Khi bắt đầu đánh giá dữ liệu tổng hợp dựa vào đo lường tiện ích, hãy nhớ lý do mà bạn muốn đo lường là để so sánh các phương pháp tổng hợp hay là để cải thiện quá trình tổng hợp dữ liệu. Sau đó, hãy nhớ đến người dùng cuối cùng và các yêu cầu đối với tệp dữ liệu tổng hợp của bạn. Nếu dữ liệu tổng hợp dành cho một nhiệm vụ hoặc loại phân tích cụ thể thì các phép đo tiện ích cụ thể sẽ là lựa chọn đầu tiên. Các phương pháp tiện ích chung có thể hữu ích cho cả việc so sánh các phương pháp tổng hợp và điều chỉnh tổng hợp đối với các thiếu sót được xác định. Đối với trường hợp thứ hai, các phép so sánh cận biên dựa trên điểm xu hướng có thể hữu ích để điều chỉnh lại quá trình tổng hợp dữ liệu của bạn. Nội dung này được các tác giả thể hiện trong "Cây quyết định đo lường tiện ích" thuộc phần "Mẹo để bắt đầu".

Hình 16. Cây quyết định đo lường tiện ích



Và ở cuối chương, các tác giả đưa ra gợi ý một số phương trình hữu ích cho một số thước đo pMSE, SPECKS, PO50, MabsDD và hai công cụ mã nguồn mở sẵn có được sử dụng phổ biến để tính toán các thước đo tiện ích này.

3. Ý kiến đánh giá chủ quan

Tài liệu "Synthetic Data for Official Statistics- A Starter Guide" là một tài liệu hướng dẫn rất phù hợp cho người làm công tác thống kê ở các cơ quan thống kê quốc gia, và cho mọi đối tượng nghiên cứu, làm việc chuyên ngành thống kê. Tài liệu này được trình bày rõ ràng, mạch lạc; nội dung các phần, các chương có tính logic cao. Ngoài ra, các tác giả thể hiện sự sáng tạo trong cách tổng hợp kiến thức cho người đọc (thể hiện ở việc xây dựng các bảng số liệu, sơ đồ dạng cây, sơ đồ giúp đưa ra quyết định, hay "Mẹo để bắt đầu", biểu đồ hình, bảng so sánh ưu nhược điểm,...). Điểm sáng của tài liệu là đưa các ví dụ thực tế, phân tích bài học kinh nghiệm của các NSO và đưa ra những khuyến nghị hữu ích, sát thực tế cho người đọc.

Tuy nhiên, khối lượng kiến thức về toán học, khoa học máy tính và khoa học dữ liệu trong tài liệu khá nhiều nên người đọc cần nghiên cứu sâu hơn các tài liệu tham khảo để có thể hiểu được đầy đủ nội dung mà các tác giả trình bày. Thêm nữa, việc trình bày các sơ đồ, bảng số liệu ở một số nội dung đôi khi diễn ra liên tục mà chưa có những bước giải thích cặn kẽ, đôi khi khiến người đọc cảm thấy lúng túng, khó có thể hiểu ngay. Hiện tại, tài liệu cũng chỉ được xuất bản với ngôn ngữ tiếng Anh, vậy nên, người đọc ở các quốc gia không sử dụng tiếng Anh sẽ mất nhiều thời gian để dịch và biên tập lại thậm chí phải tìm hiểu thêm để xác định chính xác những khái niệm, định nghĩa được diễn giải theo đúng ý của các tác giả.

4. Tài liệu này có thể tìm thấy ở đâu?

Bạn có thể download tài liệu này trên trang web (www.unece.org) dưới dạng PDF hoặc bạn có thể mua bản cứng từ Cửa hàng Sách của Liên hợp quốc (https://shop.un.org/) hay thông qua các nhà phân phối Ấn phẩm của Liên hợp quốc, Planetis.

5. Kết luận

Nếu bạn luôn mong muốn tìm được những lời chỉ dẫn từ những con người thực, có kiến thức chuyên sâu về thống kê và kinh nghiệm làm việc lâu năm, đặc biệt là những bài học họ đã khám phá và vận dụng nó thành công vào việc quản lý quyền truy cập của người dùng vào số liệu thống kê chính thức thì **"Synthetic Data for Official Statistics- A Starter Guide"** chính là tài liệu như vậy.

**Bài viết của tôi hoàn toàn là ý kiến đánh giá chủ quan của bản thân, thông qua một quá trình dịch và tìm hiểu về nội dung của tài liệu, tôi đã khái quát lại một vài nội dung (theo tôi đánh giá) là đặc sắc để bạn đọc có cái nhìn tổng quan về tài liệu này. Tuy nhiên, chắc chắn rằng trong quá trình trình bày, bài viết không thể tránh được những sai sót và hạn chế, thông qua phần kết luận này, tôi mong nhận được những đóng góp và phản hồi từ bạn đọc. Xin chân thành cám ơn!*

Ngọc Bích (Viết và tổng hợp)

Nguồn tham khảo: https://unece.org/info/publications/pub/373531