

DỮ LIỆU LỚN LÀM THAY ĐỔI KIỂU MÔ HÌNH TRONG THỐNG KÊ NHÀ NƯỚC

Barteld Braaksma và Kees Zeelenberg, Cơ quan Thống kê Hà Lan

Tóm tắt

Dữ liệu lớn đem đến nhiều cơ hội trong hoạt động thống kê nhà nước như: đưa ra nhiều giải pháp tăng cường, đúng lúc kịp thời hơn và cho ra những sản phẩm thống kê mới. Tuy nhiên, dữ liệu lớn cũng mang lại rất nhiều thách thức như: tình trạng không kiểm soát những thay đổi về nguồn dữ liệu ảnh hưởng đến tính liên tục, khó định hình để kết nối với khung tổng thể, và dữ liệu gián tiếp ngụ ý những hiện tượng hấp dẫn đối với ngành Thống kê. Dưới đây là hai giải pháp tương ứng với những thách thức và cơ hội đó.

Trước hết, chúng ta có thể xem dữ liệu lớn là những điều không hoàn hảo, tuy nhiên rất đúng lúc, kịp thời, là chỉ tiêu của các hiện tượng trong xã hội. Những dữ liệu này đang tồn tại và đó chính là lý do tại sao chúng ta đang tò mò về chúng. Thứ hai, chúng ta có thể nghiên cứu sâu hơn về giải pháp này bằng phổ biến các mô hình. Một số phương pháp mới như sử dụng các kỹ năng học máy móc có lẽ ưu thế hơn các phương pháp truyền thống như của Bayes.

Các cơ quan thống kê quốc gia (NSI) vẫn luôn do dự khi sử dụng các mô hình, ngoại trừ một số trường hợp cụ thể như ước lượng diện tích nhỏ. Chúng ta đang tranh luận rằng NSI không nên ngại sử dụng các mô hình, nên công khai sử dụng các mô hình trong tài liệu và minh bạch công khai trước người dùng. Ngoài ra, mục đích chính của NSI là mô phỏng xã hội; chúng ta nên tìm hiểu các hoạt động dự báo. Do đó, những mô hình được chọn sử dụng nên phụ thuộc vào những dữ liệu quan sát thực tế và mang giá trị quan trọng.

Từ khóa: dữ liệu lớn, thống kê sử dụng mô hình.

“Re-Make/Re-Model” một bài hát do Bryan Ferry sáng tác là bài đầu tiên trong cuốn album nổi tiếng của Roxy Music.

Bài hát mở đầu bằng một lời giới thiệu rất cụ thể, một chút âm thanh của tiệc cocktail, trước khi ra mắt chính thức năm 1950. Trong khi phần hòa nhạc của guitar, piano acoustic, guitar bass, kèn saxophone và trống chơi theo lối tương đối đơn giản và truyền thống thì sự hòa tấu của các yếu tố khác lại khá độc, lạ mang hơi hướng tương lai. Eno thì liên tục quật giọng từng hồi như cơn gió với âm điệu bất định trong Studio điện tử VCS3 của mình trong khi phong cách xướng âm của Ferry là nổi bật quẩn và đau khổ phiêu trên giai điệu. (trích Wikipedia).

1. Giới thiệu

Dữ liệu lớn xuất hiện với khối lượng lớn, vận tốc nhanh, và đa thể loại; ví dụ như: lướt web, tin nhắn Twitter, chi tiết các cuộc gọi qua di động, dữ liệu về giao thông, các giao dịch ngân hàng. Điều này đem đến nhiều cơ hội mới cho ngành Thống kê hoặc tái cấu trúc hoạt động thống kê hiện hành. Sự xuất hiện với khối lượng lớn này có thể giúp công tác thống kê chính xác hơn, chi tiết cụ thể hơn; xuất hiện với vận tốc nhanh giúp các ước lượng thống kê kịp thời hơn, với tần suất cao hơn; xuất hiện đa thể loại có thể hướng thống kê đến những lĩnh vực mới.

Trong khi đó, dữ liệu lớn có thể linh hoạt thay đổi và có tính lựa chọn: bao quát cả tổng thể mà chúng ngụ ý, có thể thay đổi từ ngày này sang ngày khác, dẫn đến bước nhảy không thể giải thích trong chuỗi thời gian. Thông thường, các quan sát đơn lẻ trong bộ dữ liệu lớn thiếu các biên liên kết và do đó không thể kết nối với các bộ dữ liệu khác hoặc các khung tổng thể. Điều này rõ ràng hạn chế khả năng thay đổi các lựa chọn và hạn chế sự thay đổi.

Do đó, sử dụng dữ liệu lớn trong hoạt động thống kê nhà nước cũng đòi hỏi nhiều phương pháp tương ứng khác. Chúng tôi xin được thảo luận hai giải pháp như sau:

Trước hết, chúng ta có thể xem dữ liệu lớn là những điều: không hoàn hảo, kịp thời, là chỉ tiêu của các hiện tượng trong xã hội. Với tư duy đó, đây chính là những gì NSI thường làm: chúng ta thu thập những dữ liệu được hoàn thành bởi đối tượng được điều tra và tại sao lại vậy, thậm chí sự thật là chúng được hoàn thành với cùng một lý do: tại sao chúng lại trở nên thú vị đối với xã hội, đối với một NSO. Hay nói ngắn gọn, chúng ta có thể nói rằng: những dữ liệu

này đang tồn tại và đó chính là lý do tại sao chúng ta thấy chúng thật thú vị.

Thứ hai, chúng ta có thể chính thức phát triển phương pháp này bằng cách trực tiếp mô hình hóa những dữ liệu này. Trong những năm gần đây, rất nhiều nhà thống kê toán ứng dụng đã và đang phát triển các phương pháp mới giải quyết tình trạng dữ liệu lớn.

Trong chương 2, chúng tôi lược tả về dữ liệu lớn và những khả năng sử dụng cũng như một số ví dụ thực tế. Trong chương 3, trước hết, chúng tôi đề cập đến cách dữ liệu lớn được sử dụng: được thu thập, được lắp ráp, ví dụ được thống kê đúng nghĩa. Trong chương 4, chúng tôi thảo luận về cách sử dụng mô hình để hình thành thông tin từ các nguồn dữ liệu lớn với điều kiện NSI có thể sử dụng các mô hình trong hoạt động thống kê nhà nước.

2. Dữ liệu lớn

Dữ liệu lớn xuất hiện với khối lượng lớn, vận tốc nhanh và đa hình thức. Trong chương này, chúng ta sẽ xem xét một số ví dụ trong chương trình nghiên cứu và cải tiến của Cơ quan Thống kê Hà Lan: các tin nhắn qua mạng xã hội, dữ liệu vòng lặp giao thông, dữ liệu điện thoại di động. Đặc biệt, chúng ta sẽ thảo luận về những cách sử dụng thực tế hoặc có thể trong thống kê nhà nước và một số vấn đề phát sinh khi phân tích nguồn dữ liệu dưới góc nhìn của thống kê nhà nước. Một số ví dụ khác chúng ta sẽ không đề cập đến như: lướt web, dữ liệu quét qua máy scan, các hình ảnh vệ tinh và các giao dịch ngân hàng.

2.1. Dữ liệu vòng lặp giao thông

Ở Hà Lan, theo ghi chép, gần 100 triệu lượt kiểm tra giao thông được thực hiện mỗi ngày. Đặc

Chủ đề: Nguồn dữ liệu cho thống kê nhà nước

Thống kê nhà nước phải được tiến hành dựa trên những gì quan sát được: thông thường dữ liệu thô sau khi thu thập cần được xử lý để đưa ra những thông tin chính xác, tin cậy, kịp thời.

Từ nhiều năm nay, những người sản xuất dữ liệu thống kê nhà nước thường phụ thuộc vào dữ liệu do chính họ thu thập được, sử dụng bảng hỏi bằng giấy, phỏng vấn trực tiếp hoặc qua điện thoại hoặc một vài phương thức khác ít mang tính truyền thống hơn như điều tra trực tuyến qua các trang web. Phương pháp truyền thống này bắt nguồn từ thời kỳ dữ liệu còn khan hiếm, khi các cơ quan thống kê nhà nước là một trong số rất ít các cơ quan có khả năng thu thập dữ liệu và phổ biến thông tin. Ưu điểm lớn nhất của phương pháp điều tra này là khả năng sự bao quát tất cả các câu hỏi được hỏi và tổng thể cần nghiên cứu.

Gần đây, các cơ quan thống kê bắt đầu sử dụng dữ liệu hành chính (chủ yếu là cơ quan chính phủ) như một nguồn dữ liệu thứ cấp. Sử dụng nguồn dữ liệu thứ cấp làm giảm tính kiểm soát các dữ liệu sẵn có và tổng thể hành chính thường không phù hợp hoàn toàn với tổng thể thống kê. Tuy nhiên, chi phí thu thập những dữ liệu này rẻ hơn nhiều so với tiến hành một cuộc điều tra như chúng ta thường làm. Ở một số quốc gia, truy cập và sử dụng nguồn dữ liệu thứ cấp được quy định rõ ràng trong pháp luật.

Nguồn dữ liệu lớn thậm chí còn ít bị kiểm soát hơn nhiều. Chúng vốn là những dữ liệu “hữu cơ”[1] được thu thập bởi những người khác- những người không hề có mục đích thống kê. Ví dụ, một cơ quan thống kê muốn sử dụng dữ liệu về các giao dịch bán lẻ, lấy thông tin giá cả phục vụ thống kê chỉ số giá tiêu dùng trong khi đó những người tạo ra dữ liệu này thì chỉ thấy chúng như một công cụ để kiểm tra hàng tồn kho và doanh số.

biệt, hơn 12 nghìn lượt kiểm tra trên các tuyến đường Hà Lan, số lưu lượng xe ô tô luôn cập nhật từng phút. Dữ liệu này được thu thập và lưu trữ bởi Kho lưu trữ dữ liệu thông tin giao thông quốc gia - cơ quan cung cấp dữ liệu miễn phí cho Thống kê Hà Lan. Điểm đáng chú ý của những vòng lặp này phân biệt chiều dài của từng loại từ đó có thể cho thấy sự khác biệt giữa chúng ví dụ như xe ô tô và xe tải. Những dữ liệu này cho thấy sự khác biệt rõ ràng trong điều khiển hành vi.

Thu thập lượng lớn dữ liệu vô cùng lớn này chính là thách thức lớn nhất đối với ngành Thống kê; nhưng nó cũng có thể cho kết quả nhanh hơn, chứa nhiều thông tin chi tiết cấp vùng hơn và nhiều giải pháp tăng cường trong bối cảnh đó. Suy rộng ra, đây

có thể là ngụ ý về những thay đổi trong hoạt động kinh tế.

Một vấn đề đặt ra là nguồn dữ liệu này không có tính bao quát nhưng lại có tính chất chọn lựa. Số các phương tiện được kiểm tra không được lưu trữ từng phút do lỗi hệ thống và không phải tất cả các con đường của Hà Lan đều có vòng lặp kiểm tra. Thật may chúng ta có thể xử lý điều này bằng cách điền các dữ liệu thiếu bằng dữ liệu đã được ghi lại bởi cùng vòng lặp trong khoảng thời gian 5 phút trước hoặc sau đó (xem hình 1). Theo thời gian, tính bao quát cũng dần được cải thiện. Dần dần, ngày càng nhiều con đường có vòng lặp kiểm tra, bao phủ toàn diện hơn các con đường quan trọng nhất của Hà Lan đồng thời giảm tính lựa chọn. Trong vòng 1 năm, đã có thêm hơn 2 nghìn vòng lặp.

2.2. Tin nhắn qua các phương tiện xã hội

Phương tiện xã hội là một nguồn dữ liệu mà mọi người sẵn sàng chia sẻ thông tin, thảo luận các chủ đề mình thích cũng như các mối quan hệ gia đình, bạn bè. Hàng ngày, ở Hà Lan, hơn 3 triệu tin nhắn công khai trên các phương tiện xã hội. Đối với bất kỳ ai truy cập mạng internet, những tin nhắn này luôn sẵn có nhưng thu thập tất cả chúng rõ ràng là một vấn đề vô cùng lớn. Công ty Coosto thực hiện nhiệm vụ thu thập tất cả các tin nhắn trên các phương tiện xã hội cung cấp cho Cơ quan Thống kê Hà Lan phân tích. Ngoài ra, họ cũng cung cấp thêm một số thông tin, như chấm điểm theo cảm tính cho những tin nhắn cá nhân hoặc thêm các thông tin về xuất xứ của một tin nhắn.

Để tìm ra rằng liệu phương tiện xã hội có phải là một nguồn dữ liệu hấp dẫn với thống kê hay không, ta tiến hành nghiên cứu các tin nhắn qua phương tiện xã hội dưới hai góc độ: nội dung và cảm tính. Theo các nghiên cứu về nội dung các tin nhắn qua mạng Twitter của người dân Hà Lan (tin nhắn đáng chú ý trên các phương tiện xã hội tại thời điểm đó) thì gần 50% các tin nhắn đó chỉ là những điều “nhảm nhí vô định” (xem hình 2). Nội dung được thảo luận trong phần còn lại đó là các hoạt động rảnh rỗi (10%), công việc (7%), phương tiện truyền thông (5%) và chính trị (3%). Về việc sử dụng những tin nhắn này còn nghiêm trọng hơn khi các tin nhắn này bị hãm lại bởi các tin nhắn nhảm nhí thiếu nghiêm túc. Cuối cùng cũng sẽ gây tác động xấu đến hoạt động nghiên cứu khai thác sử dụng nội dung tin nhắn.

Yếu tố cảm tính trong các tin nhắn qua các phương tiện xã hội cho thấy mối tương quan lớn với niềm tin của người tiêu dùng [2]. Facebook đã cho thấy một cái nhìn toàn diện nhất. Yếu tố cảm tính theo quan sát đã hiển thị trên cơ sở dữ liệu đều đặn hàng tuần, hàng tháng nhưng số liệu hàng

ngày lại thể hiện hành vi bất ổn định. Do đó, ta có thể sản xuất các chỉ tiêu cảm tính hàng tuần, thậm chí là ngày làm việc đầu tiên sau tuần tiền hành nghiên cứu.

2.3. Dữ liệu qua mạng di động

Ngày nay, mọi người mang điện thoại di động đi khắp nơi và sử dụng chúng cả ngày. Để quản lý lưu lượng điện thoại, rất nhiều dữ liệu cần được xử lý thông qua các công ty điện thoại di động. Dữ liệu này liên quan chặt chẽ đến hành vi của con người; mà hành vi của con người lại chính là đối tượng quan tâm của thống kê nhà nước. Ví dụ, lưu lượng điện thoại được chuyển tiếp thông qua hệ thống cột ăng ten phân bố theo địa lý - hệ thống có thể xác định vị trí của người sử dụng điện thoại.

Một vài dữ liệu sử dụng trong thống kê nhà nước có thể dễ dàng thấy như tổng lượng khách du lịch đến và tổng thể thời gian ban ngày. Xác định địa điểm vào ban ngày là một chủ đề mà trước đây chúng ta biết rất ít về nó do thiếu các nguồn hỗ trợ; trái lại “tổng thể thời gian ban đêm” lại phụ thuộc vào những phần đăng ký chính thức.

2.4. Diễn giải dữ liệu

Trích dẫn ý nghĩa thống kê của các nguồn dữ liệu lớn không phải chuyện đơn giản, ví dụ diễn dữ liệu về các cuộc điện thoại di động bị chặn bởi một vài lý do nào đó. Các cuộc điện thoại thường xuyên có liên quan đến con người nhưng bằng cách nào để giải thích rằng những điều đó là hiển nhiên. Con người có thể mang rất nhiều điện thoại hoặc không, trẻ con sử dụng điện thoại do cha mẹ đăng ký, các điện thoại có thể bị tắt máy,... Ngoài ra, cách mọi người sử dụng điện thoại cũng có thể thay đổi theo thời gian, phụ thuộc vào sự thay đổi của hóa đơn, những hỗ trợ về kỹ thuật và sự yêu thích đối với các công cụ truyền thông so với những công cụ khác. Đối với các tin nhắn qua phương tiện truyền thông,

những vấn đề tương tự có thể phát sinh khi cố gắng xác định tính cách của người soạn tin nhắn.

Rất nhiều nguồn dữ liệu lớn được soạn thảo gồm những dữ liệu quan sát được nhưng không được thiết kế vì mục đích phân tích dữ liệu. Chúng bị thiếu tổng thể mục tiêu, cấu trúc dữ liệu và sự đảm bảo về chất lượng. Điều này gây ra khó khăn nhất định khi áp dụng các phương pháp thống kê truyền thống chủ yếu dựa trên lý thuyết mẫu. Ví dụ, đánh giá các vấn đề có tính chọn lựa có thể chứng minh vấn đề. Do đó, rất nhiều nguồn dữ liệu được soạn thảo dưới dạng tin nhắn, nhu cầu trích dẫn thông tin từ các văn bản này cũng tăng theo. Đây gọi là lỗi khả năng khai thác tin nhắn và kỹ năng học hỏi máy móc, nhưng vẫn chưa thể quen thuộc với các thông kê viên nhà nước.

3. Dữ liệu lớn như các số liệu thống kê

Dữ liệu lớn có thể đem đến nhiều cơ hội mới cho ngành Thống kê mới hoặc tái cấu trúc ngành Thống kê hiện hành. Khối lượng dữ liệu lớn có thể cho kết quả chính xác hơn, chi tiết cụ thể hơn; vận tốc nhanh có thể đem lại những ước lượng thống kê thường xuyên và kịp thời; và đa dạng thông tin có thể đem đến nhiều cơ hội cho thống kê trong những lĩnh vực mới.

Trong khi đó, dữ liệu lớn có thể có tính lựa chọn và tính bất ổn định: tính bao trùm của tổng thể mà chúng đại diện có thể thay đổi hàng ngày, dẫn đến những bước nhảy không thể giải thích trong chuỗi thời gian. Thông thường, các quan sát cá nhân trong những bộ dữ liệu lớn này thiếu các biến liên kết và do đó không thể kết nối với các bộ dữ liệu khác hoặc các khung tổng thể. Điều này có thể hạn chế tính bất ổn định và các khả năng điều chỉnh chọn lựa khi sử dụng phương pháp truyền thống.

Trong chương này, chúng ta sẽ thảo luận NSI xử lý các vấn đề thống kê như thế nào, liệu chúng ta

có thể coi việc tăng cường dữ liệu lớn như là một hoạt động thống kê theo đúng luật. Chúng ta có thể chấp nhận coi dữ liệu lớn chỉ là một chỉ tiêu phát triển xã hội: chưa hoàn hảo tuy nhiên phản ánh kịp thời. Nhìn chung, những việc mà NSI thường làm như sau: chúng ta thu thập dữ liệu đã được thu thập từ đối tượng được điều tra và lý do tại sao, thậm chí sự thật là chúng chỉ được lắp ráp lại với nhau vì cùng lý do tại sao chúng hấp dẫn với xã hội, với NSI. Tóm lại, câu hỏi chúng ta tranh luận ở đây là: những dữ liệu này đang tồn tại và tại sao chúng lại trở nên hấp dẫn đối với chúng ta.

Có lẽ đây là điều hiển nhiên nhất đối với các tin nhắn qua các phương tiện xã hội, và các chỉ tiêu được suy ra từ đó. Các ý kiến được thể hiện qua Twitter và Facebook cũng đóng một vai trò và đôi khi là vai trò quan trọng trong các cuộc tranh luận công khai. Ví dụ, trên web của hệ thống đài và ti vi Hà Lan thường có mục các tin nhắn Twitter gửi công khai và do đó những tin nhắn này trở thành một phần của bản tin và mục thảo luận công chúng.

Tuy nhiên, chỉ tiêu cảm tính dựa trên các tin nhắn trên phương tiện xã hội, được thảo luận trong phần trước là một ví dụ. Điều đó cho thấy, chỉ tiêu này có mối tương quan chặt chẽ với các ước lượng truyền thống về niềm tin người tiêu dùng. Do đó, chúng ta có thể kết luận rằng chỉ tiêu này có mối tương quan. Tuy nhiên, chỉ tiêu cảm tính dựa trên phương tiện xã hội không theo sát chỉ tiêu truyền thống. Nói cách khác, phương pháp truyền thống thống kê về niềm tin người tiêu dùng là mở một cuộc điều tra các cuộc điện thoại, những thống kê này chứa rất nhiều sai số mẫu, và có lẽ tệ hơn, cũng có khi không có sai số mẫu nào. Điều quan trọng ở đây là: chỉ tiêu niềm tin người tiêu dùng truyền thống không phải là phương pháp phản ánh chính xác về niềm tin người tiêu dùng do có nhiều sai số mẫu, thậm chí có thể có hiện tượng chệch hoặc không có

sai số mẫu. Do đó, chúng ta có thể nói cách hợp lý hơn rằng chỉ tiêu tình cảm qua phương tiện xã hội và chỉ tiêu truyền thông đều là một phương pháp ước lượng “tâm trạng của quốc gia” và chúng ta không nên xem xét một trong hai phương pháp này đâu là phương án chính xác và tối ưu.

Một điều không thể bỏ qua ngoài tính đúng đắn là chất lượng: tính liên quan, tính kịp thời, dễ dàng truy cập, tính tương quan so sánh và tính liên kết. Do đó chỉ tiêu phương tiện xã hội rõ ràng có thể được tạo ra thường xuyên và kịp thời hơn, ghi điểm về phương diện tính kịp thời. Vì vậy, dù chỉ tiêu tình cảm qua phương tiện xã hội có thể ít liên quan hơn hoặc tính chính xác thì nó vẫn hữu ích với xã hội nếu NSI sản xuất chúng như một hoạt động thống kê nhà nước.

4. Các số liệu thống kê chính thức từ mô hình sử dụng dữ liệu lớn

Trong chương này, chúng ta thảo luận về cách sử dụng mô hình để hình thành thông tin từ các nguồn dữ liệu lớn, và với điều kiện NSI có thể sử dụng các mô hình phục vụ công tác thống kê nhà nước.

4.1. Thiết kế mô hình, các phương pháp hỗ trợ mô hình và các phương pháp dựa trên mô hình

Chúng ta sẽ bàn về việc phân biệt giữa: các phương pháp dựa trên thiết kế, các phương pháp hỗ trợ mô hình và các phương pháp dựa trên mô hình. Các phương pháp dựa trên thiết kế là một phương pháp phù hợp với mô hình mà đối tượng được điều tra được lấy mẫu theo xác suất đã biết, và thống kê viên sử dụng những xác suất này để tính 1 ước lượng chệch của một số đặc tính trong tổng thể như: thu nhập trung bình. Các phương pháp hỗ trợ mô hình sử dụng một mô hình để nắm bắt trước những thông tin về tổng thể nhằm tăng tính chính xác của ước lượng. Tuy nhiên, nếu mô hình này không đúng thì sau đó

các ước lượng này vẫn không bị chệch khi đưa duy nhất một thiết kế vào tính toán.

Các phương pháp dựa trên mô hình vẫn phụ thuộc vào tính không chính xác của mô hình: các ước lượng bị chệch nếu mô hình không đúng. Như ví dụ chúng ta đã đề cập, chúng ta muốn ước lượng tổng doanh thu của các doanh nghiệp trong một giai đoạn nhất định và đó chính là ví dụ của các doanh nghiệp mà doanh thu - quan sát được theo khái niệm thống kê nhưng một bản kê khai cho tất cả các doanh nghiệp trong tổng thể với 1 lượng thuế đã trả và doanh thu đó tuân theo Luật Thuế - Doanh thu (ví dụ: doanh thu không bao gồm: doanh thu được miễn giảm thuế hoặc doanh thu không được ghi chép lại do lỗi).

Phương pháp hỗ trợ mô hình là phương pháp sử dụng các dữ liệu về doanh thu cá nhân- thuế như các biên giả trong ước lượng hồi quy. Thậm chí nếu mô hình này không tương thích với doanh thu bị đánh thuế thì ước lượng kết quả cũng sẽ xấp xỉ mức không chệch trong thiết kế mẫu. Một ví dụ đơn giản về ước lượng dựa trên mô hình sẽ minh chứng thêm rằng tất cả dữ liệu doanh thu cá nhân trong bản kê khai doanh thu - thuế, và sử dụng chúng như một ước lượng trong tổng doanh thu.

Mô hình này ngụ ý rằng: doanh thu trong bản kê khai doanh thu - thuế bằng với khái niệm doanh thu trong thống kê. Nếu không dựng mô hình này thì ước lượng kết quả sẽ bị chệch. Tất nhiên, nếu trong thực tế chúng ta có cả hai loại dữ liệu này, mẫu và bản kê khai thì sẽ không hữu ích khi chỉ sử dụng bản kê khai và ước lượng dựa trên mô hình. Tuy nhiên, có thể chi phí sử dụng dữ liệu kê khai sẽ rẻ hơn và không cần phải lấy mẫu tất cả. Ở một số quốc gia có vấn đề chính trị bất ổn, gánh nặng hành chính của các doanh nghiệp có thể cũng là một rào cản khi lấy mẫu.

Các NSI luôn e ngại sử dụng các phương pháp dựa trên mô hình trong thống kê nhà nước. Họ thường phụ thuộc vào cá cược tổng điều tra hoặc điều tra, sử dụng các phương pháp dựa trên thiết kế hoặc các phương pháp hỗ trợ mô hình. Tuy nhiên, trong một số lĩnh vực thống kê cụ thể, NSI đã sử dụng các phương pháp dựa trên mô hình, ví dụ như trong các ước lượng khu vực nhỏ, trong việc xử lý dữ liệu trống và chọn lựa, trong tính toán điều chỉnh mùa vụ trong chuỗi thời gian và trong các ước lượng kinh tế vĩ mô. Thực tế, các kỹ năng thông thường như điền dữ liệu thiếu thường phụ thuộc vào một số giả định mô hình. Do đó, ta có thể nói, các mô hình đang được sử dụng trong thống kê nhà nước. Tuy nhiên, những mô hình này thường xuyên ngụ ý mà không được nhấn mạnh trong các tài liệu hoặc trong các bản công bố, phổ biến.

4.2. Tính bao phủ và chọn lựa

Dữ liệu lớn có thể linh hoạt thay đổi và có tính lựa chọn: dữ liệu lớn gồm cả tổng thể mà chúng ngụ ý, có thể thay đổi từ ngày này sang ngày khác, dẫn đến những bước nhảy khó giải thích trong chuỗi thời gian. Thông thường, các quan sát đơn lẻ trong bộ dữ liệu lớn thiếu các biến liên kết và do đó không thể kết nối với các bộ dữ liệu khác hoặc các khung tổng thể. Điều này rõ ràng hạn chế khả năng thay đổi các lựa chọn và hạn chế sự thay đổi. Hay nói cách khác, trong rất nhiều hiện tượng chúng ta có dữ liệu lớn, chúng ta cũng có các thông tin khác như dữ liệu điều tra cho một phần của tổng thể và các thông tin ưu tiên từ các nguồn khác.

Do đó có một phương pháp thực hiện khác là sử dụng dữ liệu lớn song song với thông tin phụ trợ và xem xét liệu chúng ta có thể dựng mô hình về hiện tượng mà chúng ta muốn mô tả. Trong những năm gần đây, có một cuộc đại phẫu trong ngành thống kê toán: phát triển những phương pháp mới sử dụng dữ liệu lớn. Chúng diễn ra ở rất nhiều lĩnh vực:

hồi quy đa biến, kỹ năng học qua máy, dựng mô hình đồ họa, khoa học dữ liệu, mạng lưới những người theo trường phái học thuyết Bayes [3,4,5]. Tuy nhiên, các phương pháp truyền thống như: kỹ năng Bayes, thuật toán chọn lọc và các mô hình đa cấp (phân cấp) được sử dụng nhiều hơn [6].

Một chiến lược khác lấy cảm hứng từ các tài khoản quốc gia được soạn thảo. Rất nhiều nguồn tự chúng vẫn chưa hoàn thiện, chưa hoàn hảo và/ hoặc phần nào chông chéo lên nhau được tổng hợp, sử dụng một khung khái niệm liên quan để có được bức tranh toàn diện về tổng thể nền kinh tế, trong khi áp dụng rất nhiều phương pháp kiểm tra và cân đối. Cũng theo cách đó, dữ liệu lớn và các nguồn khác tự chúng cũng đang không hoàn thiện, hoặc bị chệch có thể được kết nối cùng nhau để tạo nên một bức tranh hoàn thiện, không chệch, từ đó phản ánh một hiện thực chắc chắn.

Nhìn chung, chúng ta có thể nói rằng dữ liệu lớn là trường hợp mà chúng ta thiếu thông tin về quá trình hình thành dữ liệu. Do đó, các mô hình cũng rất hữu ích khi ước lượng dữ liệu thiếu.

4.3. Chất lượng, tính khách quan và độ tin cậy

Với tư cách là nhà sản xuất số liệu thống kê nhà nước, các NSI phải cẩn trọng khi áp dụng các phương pháp dựa trên mô hình. Công chúng cũng không nên lo lắng về chất lượng thống kê nhà nước như đã quy định trong tuyên ngôn về sứ mệnh của hệ thống Thống kê Châu Âu.

“Chúng tôi cung cấp cho Cộng đồng châu Âu, thế giới và công chúng nguồn thông tin độc lập chất lượng cao về nền kinh tế và xã hội châu Âu, các cấp quốc gia và khu vực đồng thời phổ biến thông tin rộng rãi để mọi người dễ dàng truy cập vì các mục đích hoạch định chính sách, nghiên cứu và tranh luận”

Tính khách quan và độ tin cậy là hai trong số các nguyên tắc thực hiện thống kê nhà nước đã quy định trong Luật Thống kê châu Âu (7) "... có nghĩa rằng: thống kê phải được phát triển, được sản xuất và được phân tán theo hệ thống, theo cách tin tưởng và không chệch". Và cơ quan Thực hiện thống kê châu Âu cho rằng "Thống kê châu Âu phác họa thực tế một cách chính xác và tin cậy". Ngoài ra, còn rất nhiều tuyên ngôn quốc tế đã đề ra trong ISI [9] và UN [10], nhưng tất cả các Luật Thống kê quốc gia của Hà Lan đều có chung một nguyên tắc.

Khi sử dụng mô hình, chúng ta có thể diễn giải hai nguyên tắc này như sau. Nguyên tắc về tính khách quan nghĩa là dữ liệu được sử dụng để ước lượng mô hình nên phản ánh hiện tượng mà nó mô tả; hay nói cách khác, mục đích và tổng thể mẫu cũng phải tương ứng với hiện tượng thống kê. Dữ liệu trong quá khứ có thể được sử dụng để ước lượng mô hình nhưng ước lượng dựa trên mô hình chưa bao giờ vượt xa quãng thời gian hiện tại; do đó hiện nay, tổng hợp thì được phép nhưng không dự báo và phân tích chính sách.

Nguyên tắc độ tin cậy được hiểu là chúng ta phải tránh sửa đổi dữ liệu thống kê chỉ bởi vì những thay đổi mô hình, vd: bởi vì sập mô hình 9 (lỗi mô hình). Đặc biệt, đối với các mô hình chuỗi thời gian, chúng ta phải bảo vệ cẩn thận, vì sập mô hình có thể dẫn đến xác định sai các điểm thay đổi trong chuỗi thời gian

Tuy nhiên, chúng ta cũng nên hạn chế sử dụng các mô hình dựa trên hành vi, bởi vì những mô hình này là nguyên nhân dẫn đến sụp đổ mô hình: gần như chắc chắn rằng sẽ có lúc trong tương lai, bất kỳ mô hình hành vi nào cũng sẽ trở nên không đúng bởi vì hành vi kinh tế và hành vi các đơn vị trong xã hội đều đang thay đổi. Một lý do khác, để tránh các mô hình hành vi, chúng ta phải ngăn các tình huống mà một nghiên cứu viên đầu

vào thấy kết quả tốt khi khớp mô hình, nhưng lại không biết rằng cùng mô hình đó NSO đã sử dụng để tạo ra những dữ liệu khác nhau đã được nghiên cứu bên ngoài sử dụng.

Nguyên tắc khách quan và độ tin cậy cũng dẫn đến một số nguyên tắc phương pháp dựa trên mô hình. Đặc biệt, mô hình dựng nên được thực hiện kèm với các kiểm định thống số chuyên sâu để đảm bảo tính khách quan của mô hình.

Căn cứ vào những nguyên tắc này, Cơ quan Thống kê Hà Lan vừa mới phát triển bản hướng dẫn [11] về việc sử dụng mô hình trong thống kê nhà nước. Rất nhiều, nếu không muốn nói là hầu hết các ví dụ mô hình được sử dụng trong thống kê nhà nước, đi trước cả những hướng dẫn này. Do đó, mặc dù cảnh cáo, nhưng chúng ta tin rằng vẫn có chỗ để sử dụng các mô hình trong công tác sản xuất dữ liệu thống kê từ dữ liệu lớn.

4.4. Ví dụ

Dưới đây là một vài ví dụ về các phương pháp dựa trên mô hình sử dụng dữ liệu lớn. Lưu ý rằng tất cả các ví dụ hiện vẫn đang trong giai đoạn nghiên cứu. Tác giả bài viết này cũng không biết về các trường hợp các phương pháp tương tự đã được sử dụng trong hoạt động sản xuất thống kê nhà nước thường xuyên.

Phân tích vòng lặp giao thông cá nhân: Ở cấp độ các vòng lặp cá nhân, số lượng phương tiện giao thông hiển thị hành vi thay đổi. Điều này phần nhiều là do tính khó dự đoán của giao thông ở mức độ các phương tiện cá nhân. Các kỹ năng phức tạp khác cũng cần thiết để xác định rõ các mẫu và sản xuất số liệu thống kê có ý nghĩa. Một phương pháp hiện đã được thực hiện bởi Cơ quan Thống kê Hà Lan là coi công thức Bayes là bộ lọc đệ quy, giả sử thực hiện dữ liệu vòng giao thông thô tuân theo phân phối Poisson (xem hình 3).

Dữ liệu vòng lặp giao thông và các hoạt động kinh tế vùng: Liệu tăng cường giao thông có liên quan gì đến thông tin về các hoạt động kinh tế vùng hay không? Đây là một câu hỏi rất thú vị, đã được kiểm định bằng cách sử dụng dữ liệu vòng lặp giao thông ở vùng Eindhoven, khu sản xuất quan trọng của Hà Lan [12]. Dữ liệu từ cuộc điều tra tâm lý trong khu sản xuất) được sử dụng làm khung chuẩn, bởi vì nó được biết đến như một chỉ tiêu có tính chu kỳ về tình hình kinh doanh tốt có mối tương quan mạnh mẽ đã được chứng minh trong giai đoạn phát triển kinh tế ngắn hạn. Kết quả cuộc điều tra đã công bố đến từng tỉnh, và Eindhoven là khu vực thuộc vùng Noord-Brabant. Điều này có nghĩa rằng dữ liệu từ cuộc điều tra này nên có tính liên kết chặt chẽ với hoạt động kinh tế trong vùng Eindhoven.

Hoạt động phân tích này được thực hiện bởi 3 kỹ năng khác nhau: lựa chọn và tổng hợp dữ liệu trực tiếp, thuật phân tích thành phần phụ thuộc (ICA) và thuật phân rã chế độ dựa theo kinh nghiệm. Cả 3 kỹ năng này đều cho cùng một kết quả giống nhau nhưng thuật thứ 3 (EMD) cho kết quả biểu diễn chung tốt nhất (xem hình 4).

Sự phát triển của chỉ tiêu tăng cường giao thông thật ngạc nhiên luôn theo sát sự phát triển sản xuất kỳ vọng. Điểm cao nhất và thấp nhất trùng khớp với nhau có nghĩa là chỉ số tăng cường giao thông có thể là điểm ngoặt quan trọng trong hoạt động kinh tế.

Tính liên quan giữa hai chuỗi có thể được cải thiện thậm chí nâng cao hơn nữa nếu xử lý thêm vài thủ thuật, đặc biệt là điều chỉnh mùa vụ. Một lựa chọn quan trọng khác là thực hiện phân rã chu kỳ theo xu hướng, có thể giúp tập trung vào yếu tố chu kỳ kinh tế và loại bỏ một số yếu tố khác. Thật không may, chuỗi tăng cường giao thông quá ngắn tại thời điểm cả hai loại đang lọc.

Xu hướng tra Google (Google Trends) để dự báo tức thời. Trong mục (13), tác giả chỉ ra cách sử dụng dữ liệu về động cơ nghiên cứu từ Google Trends để “dự báo hiện tại” hay còn gọi là “dự báo tức thời”. Chúng cho thấy rất nhiều ví dụ khác nhau về các chỉ tiêu kinh tế gồm doanh thu tự động hóa, khiếu nại thất nghiệp, kế hoạch điểm đến du lịch, và niềm tin khách hàng.

Trong hầu hết các trường hợp, họ áp dụng mô hình tự hồi quy đơn kết hợp với Google Trends tìm kiếm thuật ngữ giống như một thiết bị dự đoán. Để dự báo thức thời niềm tin người tiêu dùng, họ sử dụng một mô hình hồi quy Bayes, vì vậy trong trường hợp này, rõ ràng không biết phải tìm thuật ngữ nghiên cứu chuyên sâu nào để sử dụng.

Họ thấy rằng các mô hình đơn giản gồm các biến Google Trends liên quan thường có xu hướng hình thành các mô hình không chứa các thuật ngữ dự báo từ 5% đến 20%. Mặt khác, chúng ta cũng nên cẩn trọng khi diễn giải, đọc kết quả dựa trên các thuật ngữ nghiên cứu.

Hai năm trước, đã có rất nhiều rất nhiều mối quan tâm đến Google Flu, nhưng càng gần về đây thì những biểu hiện dự báo tức thời của Google Flu đã có xu hướng giảm nghiêm trọng (14). Google đã chỉ trích vì sự không minh bạch này: họ không hề tiết lộ thuật ngữ nghiên cứu được sử dụng trong Google Flu, theo thường lệ một cuộc tranh luận giữa các nhà khoa học nổ ra và có sự xác nhận chéo giữa những người cùng cấp, ngang hàng với nhau.

5. Kết luận

Có 3 kết luận chính. Thứ nhất, dữ liệu lớn xuất hiện với khối lượng lớn, vận tốc nhanh và đa chủng loại. Điều này đã đem đến nhiều cơ hội mới để ngành Thống kê đổi mới hoặc tái cấu trúc hệ thống thống kê hiện tại:

- Khối lượng lớn có thể cung cấp chính xác hơn, chi tiết cụ thể hơn;

- Vận tốc lớn có thể đem đến những ước lượng thống kê thường xuyên và kịp thời hơn;

- Sự đa dạng trong dữ liệu lớn có thể đem lại nhiều cơ hội cho thống kê trong các lĩnh vực mới.

Thứ hai, ít nhất trong một số trường hợp, thống kê dựa trên dữ liệu lớn rất hữu dụng về mặt quyền lợi và ví dụ vì chúng được sử dụng trong hoạch định chính sách hoặc đóng vai trò trong thảo luận công chúng.

Thứ ba, các NSI nói chung không nên ngại sử dụng các mô hình trong sản xuất thống kê nhà nước bởi vì chúng cũng đã và đang được sử dụng thêm vào trước đây, và do đó chúng ta nên nghiên cứu kỹ hơn về cách sử dụng các mô hình để sản xuất dữ liệu thống kê nhà nước sử dụng dữ liệu lớn.

Tài liệu tham khảo

[1] Robert M. Groves, Three eras of survey research, *Public Opinion Quarterly* 75, 861–871, 2011, doi: [10.1093/pog/nfr057](https://doi.org/10.1093/pog/nfr057).

[2] Piet J.H. Daas and Marco J.H. Puts, Social media sentiment and consumer confidence, *Paper presented at the Workshop on using Big Data for Forecasting and Statistics*, Frankfurt, 2014.

[3] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen, High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2) (2014), 29-50, doi: [10.1257/jep.28.2.29](https://doi.org/10.1257/jep.28.2.29).

[4] David W. Nickerson and Todd Rogers, Political campaigns and big data, *Journal of Economic Perspectives*, 28(2) (2014), 51-74, doi: [10.1257/jep.28.2.51](https://doi.org/10.1257/jep.28.2.51).

[5] Hal R. Varian, Big data: new tricks for econometrics. *Journal of Economic Perspectives*, 28(2) (2014), 3-28, doi: [10.1257/jep.28.2.3](https://doi.org/10.1257/jep.28.2.3).

[6] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin, *Bayesian Data Analysis*, 3e, Chapman and Hall/CRC, 2013.

[7] European Union, Regulation on European statistics, *Official Journal of the European Union*, L 87 (31 March 2009), 164–173,

[8] European Union, Code of Practice for European Statistics, revised edition, Eurostat, Luxembourg, http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice, 2005/2011.

Đặc biệt là các phương pháp Bayes và các mô hình đa phân cấp đầy triển vọng.

Mặt khác, nên công khai sử dụng các mô hình. Chúng nên được đưa vào các tài liệu và được minh bạch công khai trước người dùng. Tuy nhiên, các mô hình không được sử dụng bừa bãi: chúng ta không nên quên rằng mục đích cơ bản của một NSI là mô tả chứ không ra lệnh hoặc kết tội. Do đó chúng ta không nên hạn chế việc đưa ra những dự đoán và thực hiện các mô hình hành vi trong sáng. Tuy nhiên, chúng ta cũng nên cẩn trọng tránh chọn sai mô hình khi giả định về nó sụp đổ. Do đó, bất kỳ mô hình nào cũng nên được xây dựng dựa trên dữ liệu thực tế quan sát được sau một thời gian quan sát, có liên quan đến tình hình kinh tế và hiện tượng xã hội mà chúng ta đang cố gắng mô tả bằng các ước lượng thống kê; và xây dựng mô hình nên thực hiện song song với kiểm định các thông số chuyên sâu.

<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32009R0223:EN:NOT>.

[9] International Statistical Institute, Declaration on Professional Ethics, revised edition, <http://www.isi-web.org/about-isi/professional-ethics>, 1985/2010.

[10] Statistical Commission of the United Nations, Fundamental Principles of Official Statistics. <http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx>, 1991/2014.

[11] Bart Buelens, Peter-Paul de Wolf, and Kees Zeelenberg, Model-based estimation at Statistics Netherlands. Discussion Paper, Statistics Netherlands, The Hague, 2014.

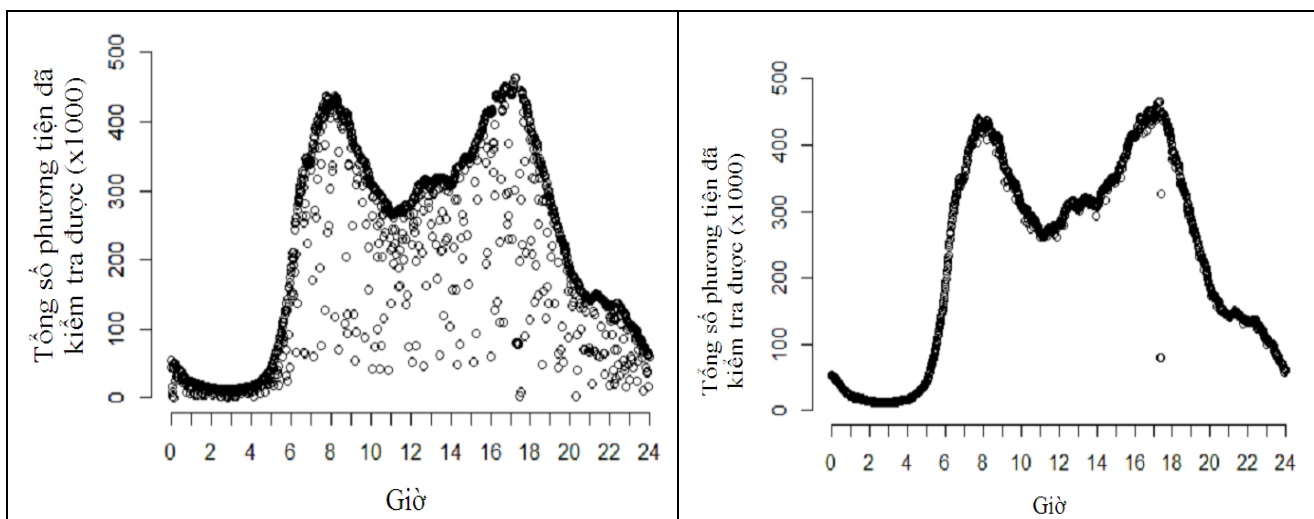
[12] Floris J. van Ruth. Traffic intensity as indicator of regional economic activity, Internal discussion paper, Statistics Netherlands, 2014.

[13] Hyunyoung Choi and Hal R. Varian, Predicting the present with Google trends, <http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf>, 2011.

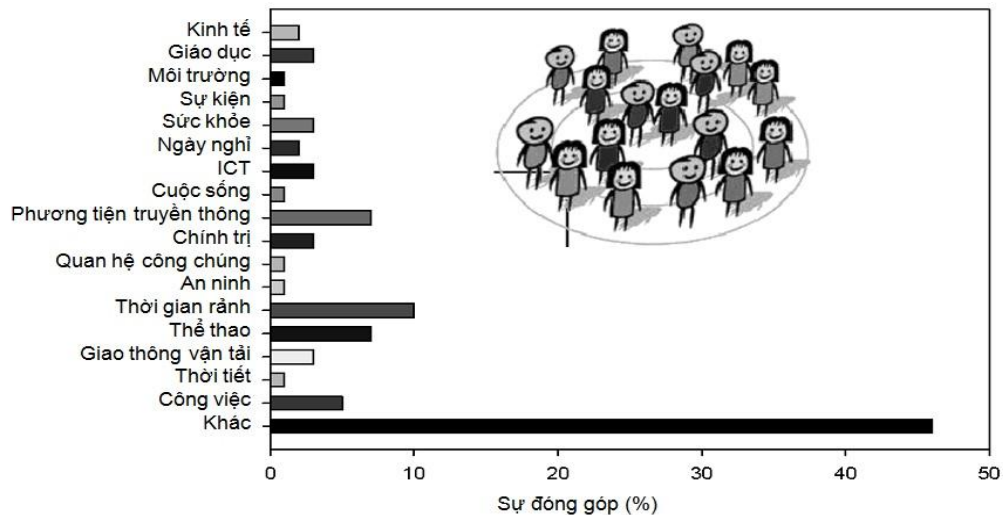
[14] David Lazer, Ryan Kennedy, Gary King and Alessandro Vespignani, The parable of Google flu: traps in big data analysis, Science 343(14) (2014), 1203-1205, doi: [10.1126/science.1248506](https://doi.org/10.1126/science.1248506)

Phụ lục hình

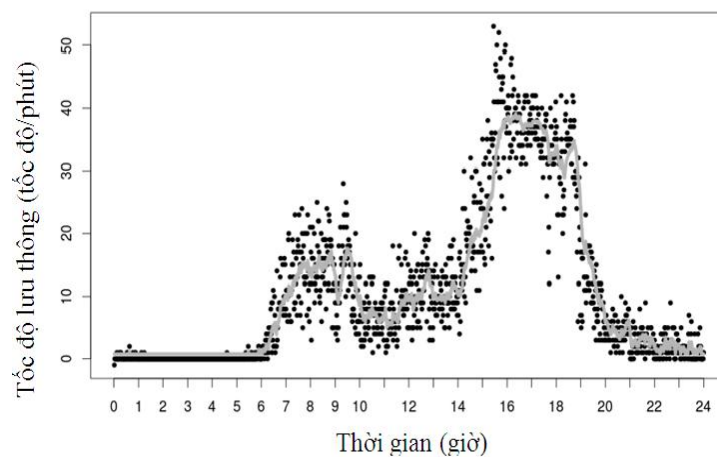
Hình 1. Mô hình phân phối giao thông trong một ngày (thứ 4, 01/09/2011) đã tổng hợp tất cả các vòng lặp giao thông trong 5 phút. Hình 1a thể hiện dữ liệu thô đã được ghi lại; Hình 1b thể hiện dữ liệu sau khi đã xử lý dữ liệu trống.



Hình 2: Phân phối tin nhắn qua Twitter của người Hà Lan theo chủ đề thống kê. Chủ đề thống kê là những vấn đề đã được xác định trong chương trình làm việc hàng năm của Thống kê Hà Lan; một chủ đề khác được thêm vào là “Phương tiện truyền thông” do mức độ liên quan của chủ đề này. Loại “chủ đề khác” này gồm các nội dung không liên quan đến bất kỳ chủ đề nào.



Hình 3. Kết quả (đường màu xám) của các ứng dụng của bộ lọc đệ quy Bayes đối với dữ liệu thô (các chấm đen) từ một vòng đơn kiểm tra giao thông, giả định rằng chúng tuân theo phân phối Poisson.



Hình 4. Chỉ tiêu hàng tháng đã lọc EMD về lưu lượng giao thông trung bình giờ cao điểm ở khu vực Eindhoven so với tốc độ phát triển sản xuất kỳ vọng của ngành công nghiệp sản xuất thuộc tỉnh Noord - Brahant. Hệ số tương quan là 0,523

