# VIỆC SỬ DỤNG CÁC BẰNG CHỨNG HOẠT ĐỘNG WEB NHẰM TĂNG TÍNH KỊP THỜI CÁC CHỈ TIÊU THỐNG KÊ CHÍNH THỨC

*Fernando Reis, Pedro Ferreira và Vittorio Perduca, Ủy ban Thống kê châu Âu*

***Tóm tắt***

Cộng đồng thống kê chính thức phản ứng với những cơ hội và thách thức được cung cấp bởi dữ liệu lớn. Ở châu Âu, Thủ trưởng các Viện thống kê quốc gia và Ủy ban Thống kê châu Âu (Eurostat) đã nhất trí về biên bản ghi nhớ giải quyết các vấn đề về nguồn dữ liệu lớn. Một trong những nguồn dữ liệu lớn sẵn có của thống kê chính thức là các dấu vết điện tử để lại khi người sử dụng truy cập vào các dịch vụ web. Nhiều dịch vụ cung cấp dữ liệu dựa trên các dấu vết để lại ở thời gian thực hoặc khoảng thời gian ngắn. Nhiều hoạt động của con người được đo lường bằng số liệu thống kê chính thức có liên quan chặt chẽ đến hành vi của người dùng trực tuyến, dữ liệu hoạt động web cung cấp tiềm năng để báo các chỉ tiêu kinh tế-xã hội với mục đích tăng tính kịp thời của số liệu thống kê. Nhiều thí nghiệm được tiến hành gần đây cho thấy những dự báo này có thể thực hiện được. Tuy nhiên, có mô hình dự báo tốt là chưa đủ để sản xuất ra số liệu thống kê chính thức. Nếu muốn đánh giá khả năng sử dụng nguồn dữ liệu lớn thì chúng ta cần phải suy nghĩ về tính minh bạch, tính liên tục, chất lượng và tiềm năng được tích hợp với các phương pháp thống kê truyền thống, cũng nghiên cứu chi tiết hơn về mối quan hệ giữa hoạt động web với các hiện tượng được dự báo.

Từ khóa: Dữ liệu lớn, hiện đại hóa, web, dự báo, ước tính nhanh

## 1. Giới thiệu

Dữ liệu lớn làm cho cộng đồng thống kê chính thức chú ý đến sự tồn tại của nhiều nguồn dữ liệu mới có khả năng sử dụng trong sản xuất ra số liệu thống kê. Một trong những nguồn đó là các dấu vết để lại bởi người sử dụng các dịch vụ web, liên quan đến các khía cạnh khác trong đời sống xã hội của người sử dụng và được đo bằng số liệu thống kê chính thức. Ví dụ, khi đối mặt với sự thất bại trong công việc, người dùng tìm kiếm thông tin về việc làm mới trên mạng, tham khảo các trang web liên quan đến việc làm và đăng bài viết trên Facebook hay Twitter.

Người dùng sử dụng dữ liệu web do nó có khả năng cung cấp rất nhanh vì các dịch vụ web là dịch vụ điện tử được hỗ trợ hoàn toàn bởi các hệ thống IT và ở mức độ tự động hóa cao. Những dữ liệu này sẽ tự động lưu trữ trong cơ sở dữ liệu hỗ trợ các dịch vụ web hoặc các bản ghi trên máy chủ. Một số dữ liệu này là công cộng (ví dụ Twitter) hoặc là mẩu tin quảng cáo (dưới dạng tổng hợp) bởi các dịch vụ web (ví dụ Google).

Hiện đã có một số kinh nghiệm trong việc sử dụng dữ liệu hoạt động web để dự báo các chỉ tiêu thống kê kinh tế-xã hội, như tỷ lệ mắc bệnh cúm, thất nghiệp, du lịch và các luồng di cư. Một số cơ quan thống kê đã tiến hành các nghiên cứu.

Trong bài viết này, chúng tôi thấy đây là cách tương đối dễ dàng để tích hợp một số dữ liệu về hoạt động tìm kiếm web của người sử dụng nhằm tăng độ chính xác của mô hình dự báo đơn giản, như trong trường hợp thất nghiệp. Tuy nhiên, nếu thống kê chính thức sử dụng các dữ liệu hoạt động web để ước tính nhanh các chỉ tiêu kinh tế-xã hội thì không nên làm điều đó bằng cách tái tạo những gì người khác có thể làm, nhưng thay vì làm điều đó làm cho việc sử dụng các lợi thế so sánh cụ thể của nó. Để tích hợp loại nguồn tính toán các ước tính nhanh của các chỉ tiêu kinh tế-xã hội chính thức, cơ quan thống kê cần phải giải quyết một số thách thức. Những kinh nghiệm cung cấp bài học quan trọng giúp giải quyết những thách thức này.

Phần 2 bài viết tóm tắt những cơ hội và thách thức của dữ liệu lớn trong thống kê chính thức và mô tả các hành động được thực hiện bởi hệ thống thống kê châu Âu; Phần 3 mô tả công việc trước đây của các nhà nghiên cứu và các nhà thống kê chính thức về dự báo các chỉ tiêu kinh tế-xã hội dựa trên hoạt động web; Phần 4 là ví dụ về mô hình rất đơn giản nhằm cải thiện tính kịp thời của số liệu thống kê thất nghiệp dựa trên cả dữ liệu chính thức và dữ liệu ở Google Trends; Phần 5 minh họa kinh nghiệm của Eurostat trong ước tính nhanh dựa trên các dữ liệu thứ cấp và làm thế nào để phát triển các sản phẩm thống kê mới dựa trên dữ liệu lớn; Phần 6 giới thiệu dữ liệu hoạt động web trong việc tính toán các ước tính nhanh.

## 2. Đối phó với những thách thức dữ liệu lớn của Hệ thống thống kê châu Âu[14]

### *2.1. Dữ liệu lớn, dữ liệu mới*

Sau nhiều thế kỷ, đầu tiên duy nhất và tại đó người thu thập dữ liệu về kinh tế và xã hội đã vượt qua sự độc quyền của cơ quan thống kê. Bây giờ, dữ liệu ở tất cả xung quanh chúng ta. Những gì đã từng khan hiếm và phải thu thập một cách đắt đỏ thì nay trở thành tài nguyên dồi dào sẵn có.

Dữ liệu lớn nghĩa là trước tiên và dữ liệu mới hết mức tối đa để số liệu thống kê chính thức bao gồm dữ liệu các loại mới và có đặc điểm khác với những nguồn dữ liệu truyền thống. Thêm vào các phép đo định lượng truyền thống và đặc điểm định tính của các cá nhân và doanh nghiệp, dữ liệu lớn mang lại sự thừa nhận rằng có thể được tìm thấy giá trị trong bất kỳ loại dữ liệu nào. Điều này bao gồm dữ liệu mạng (ví dụ mạng xã hội và truyền thông điện thoại di động), văn bản (ví dụ Twitter), hình ảnh, âm thanh và video. Bằng chứng hoạt động web bao gồm các dấu vết để lại bởi những người sử dụng các dịch vụ web được đăng ký tại tập tin ghi nhận sự kiện (log file) của các Web server (đôi khi được biên soạn dưới dạng tổng hợp và được cung cấp bởi các nhà cung cấp) và các thông tin (thường là văn bản) được nhập vào bởi người sử dụng sẵn có trong trang web.

Các nguồn dữ liệu mới này đưa ra thách thức đặc biệt cho số liệu thống kê chính thức. Thứ nhất, đôi khi các tổ chức nắm giữ dữ liệu nằm ngoài thẩm quyền của các nhà chức trách thống kê (ví dụ khi họ là các công ty nước ngoài, chẳng hạn Google, Facebook). Thứ hai, thứ tự tầm quan trọng của dữ liệu có thể được thu thập bởi các Viện thống kê quốc gia (NSI) từ chủ sở dữ liệu cao hơn nhiều so với các bộ sưu tập dữ liệu truyền thống. Điều này gây ra hai hậu quả. Một mặt, điều đó không còn hợp lý và để

[14] Các nội dung của chương này trích từ một phần bài báo (Reis, Demunter, "Công việc Eurostat trong dữ liệu lớn và Biên bản ghi nhớ Scheveningen") gửi Tạp chí quốc tế về Thông tin khoa học địa lý, vấn đề đặc biệt mang tên "Địa lý Mobility: các ứng dụng của dữ liệu Location Based".

lại gánh nặng cho việc biên soạn và truyền dẫn dữ liệu đến nhà cung cấp dữ liệu. Mặt khác, tỷ lệ thông tin không có ý nghĩa tăng đáng kể. Thứ ba, trong một số trường hợp, cơ quan thống kê quan tâm đến các dữ liệu có giá trị thương mại cho các nhà cung cấp dữ liệu khi chúng có giá trị cốt lõi trong mô hình kinh doanh của mình (ví dụ Google, Facebook).

### *1.2. Cơ hội của dữ liệu lớn*

Các nguồn dữ liệu mới cung cấp nhiều cơ hội cho số liệu thống kê chính thức. Nhiều nguồn dữ liệu lớn bao gồm các bộ dữ liệu rất lớn có thể được NSI sử dụng để cung cấp số liệu thống kê chi tiết hơn nhiều so với phương pháp sản xuất thống kê truyền thống. Sự chi tiết này không chỉ thể hiện ở mức độ địa phương, mà còn để sản xuất số liệu thống kê cho nhóm dân số rất nhỏ mà chưa được thống kê chính thức đáp ứng.

Cơ hội khác là khả năng sử dụng dữ liệu đã có sẵn, chi phí thấp hơn so với các phương pháp truyền thống. Đây không phải là để nói rằng nguồn dữ liệu lớn là được miễn phí. Như đã đề cập trước đó, một số tập dữ liệu có thể lớn đến nỗi không còn hợp lý để rời khỏi trách nhiệm cung cấp dữ liệu thống kê đến một vài nhà cung cấp dữ liệu.

Cơ hội liên quan nhất đối với chúng tôi trong bài viết này là khả năng truy cập đến dữ liệu ngay sau khi các sự kiện xảy ra. Do các nguồn dữ liệu lớn thông thường bắt nguồn từ hệ thống tự động hóa nên không có độ trễ về thời gian của tập hợp dữ liệu. Trong trường hợp các hoạt động dịch vụ web của người sử dụng trên trang web được tự động đăng ký trong cơ sở dữ liệu hoặc trong các file log của web server. Trong trường hợp dịch vụ web cung cấp dữ liệu có nguồn gốc từ các hoạt động người dùng, họ có thể làm điều đó rất nhanh (xem thêm ví dụ của Google).

### *2.3. Tác động đến sản xuất số liệu thống kê chính thức*

Các hành động chủ yếu từ thiết kế ban đầu để tái sử dụng nguồn thứ cấp chắc chắn sẽ đòi hỏi sự biến đổi trong NSI. Thứ nhất, quá trình sản xuất số liệu thống kê thay đổi như thế nào và những kỹ năng của các nhà thống kê chính thức. Từ những người thiết kế duy nhất với mục đích sản xuất hệ thống thống kê nguyên tử cho các sản phẩm thống kê cụ thể, các nhà thống kê cần trở thành nhà thiết kế sản phẩm thống kê nhằm mục tiêu đáp ứng nhu cầu của xã hội hay các nhà hoạch định chính sách dựa trên vô số nguồn dữ liệu. Đây là sự thay đổi đã và đang xảy vì nhiều lý do khác nhau. Việc sử dụng nguồn dữ liệu hành chính đã tăng lên trong những thập kỷ qua, vì vậy việc sử dụng các nguồn thứ cấp không phải là mới với NSI. Sự cần thiết phải hiện đại hóa hệ thống sản xuất thống kê (để tăng tính hiệu quả và sự linh hoạt) cũng đã khởi xướng ra phong trào hướng tới việc hội nhập sản xuất số liệu thống kê ở lĩnh vực khác nhau.

Thứ hai, dữ liệu lớn cuối cùng có thể mang lại nhiệm vụ và trách nhiệm mới cho NSI. Cụ thể, thống kê chính thức có thể đảm nhận vai trò đảm bảo chất lượng của số liệu thống kê được sản xuất từ các nguồn dữ liệu lớn, tự mình hoặc các tổ chức khác, thông qua cơ chế kiểm định và chứng nhận chất lượng [2].

### *2.4. Phân tích thách thức*

Tuy nhiên, dữ liệu lớn đầy đủ hơn dữ liệu mới. Nó thể hiện sự thay đổi về quan điểm đối với dữ liệu. Trong khi một số công ty tư nhân xây dựng mô hình kinh doanh hoàn chỉnh dựa trên thăm dò dữ liệu thương mại (ví dụ Google, Facebook), thì có những người lại tìm cách kiếm tiền từ các dữ liệu đó (một số trường hợp đã làm việc công ty trong một thời

gian). Năng động trong việc tìm kiếm những cách thức sáng tạo để khám phá dữ liệu qua các phương pháp, công cụ phân tích dữ liệu và sự gia tăng ngoạn mục trong dữ liệu sẵn có (hoặc khả năng mới để thu thập dữ liệu), điều này dẫn đến diện mạo mới của các sản phẩm dữ liệu dựa trên nhiều hoặc một vài phân tích phức tạp, đặc biệt là phân tích dự báo trước.

Trong thế giới của dữ liệu lớn để thử nghiệm các phân tích, thống kê chính thức không thể tránh khỏi sự thiếu hụt trong phân tích. Do đó, dữ liệu lớn cũng đại diện cho thống kê chính thức với những thách thức đối với người sử dụng thống kê hiện tại với các sản phẩm thống kê mới khi mà họ đang quen với việc sử dụng ở nơi khác.

Loại sản phẩm thống kê nêu trong bài viết này là ví dụ về các sản phẩm phân tích mới. Dựa trên tính kịp thời cao của một số nguồn mới dựa trên hoạt động web cá nhân, có khả năng sử dụng mô hình dự báo để cung cấp cho người dùng ước tính nhanh các chỉ tiêu kinh tế-xã hội truyền thống trong thời gian ngắn.

### *2.5. Biên bản ghi nhớ Scheveningen và công việc tiếp theo*

Nhận thấy sự thay đổi về điều kiện và môi trường hoạt động của thống kê chính thức, cộng đồng quốc tế các nhà thống kê chính thức đã phản ứng lại.

Nhóm cấp cao UNECE về hiện đại hoá sản xuất và dịch vụ thống kê đã đưa ra trong tầm nhìn chiến lược của mình (UNECE, 2010), việc tạo ra các sản phẩm thống kê mới dựa trên thăm dò hoạt động của các nguồn dữ liệu mới như là một yếu tố then chốt của hiện đại hóa số liệu thống kê chính thức (Nhóm cấp cao UNECE về hiện đại hoá sản xuất và dịch vụ thống kê, 2011).

Nhận thấy tầm quan trọng của chiến lược về dữ liệu lớn cho Eurostat, Giám đốc Viện Thống kê quốc gia châu Âu đã nhất trí về một bản ghi nhớ địa chỉ dữ liệu lớn được chính thức thông qua bởi ESSC ở Scheveningen tháng 9/2013.

Biên bản ghi nhớ Scheveningen ghi nhận rằng mức độ gia tăng số hóa xã hội, để lại dấu vết số hóa khi người rời đi, cung cấp một cơ hội cho việc biên soạn số liệu thống kê dựa trên các khái niệm của thống kê chính thức. Đặc biệt, cung cấp giải pháp thay thế để đối phó với những thách thức phải đối mặt hiện nay, chẳng hạn như tỷ lệ đáp ứng và sự cần thiết phải nâng cao hiệu quả tổng thể của hệ thống sản xuất thống kê.

Tuy nhiên, Biên bản ghi nhớ Scheveningen cũng công nhận việc sử dụng dữ liệu lớn đặt ra thách thức cho Eurostat. Do đó, đây là sự kiểm tra khả năng và chiến lược phát triển của thống kê chính thức từ dữ liệu lớn. Để đưa ra được chiến lược và lộ trình như vậy, Eurostat đã hình thành lực lượng đặc nhiệm gồm nhiều người từ Eurostat, NSI, các tổ chức quốc tế khác và học viện.

Mặc dù rất dễ nhận ra dữ liệu lớn có khả năng gây tác động lớn, nhưng ở giai đoạn này không dễ dàng xem xét dữ liệu lớn có ý nghĩa như thế nào đối với thống kê chính thức. Các nguồn dữ liệu mới có thể sẵn sàng cho việc sản xuất số liệu thống kê, nhưng nhiều khả năng mỗi nguồn dữ liệu mới đều có đặc thù riêng.

Chiến lược dự kiến của lực lượng đặc nhiệm do Eurostat thành lập đặc trưng bởi ba yếu tố. Thứ nhất, bắt đầu bằng việc thử nghiệm ứng dụng cụ thể của các nguồn dữ liệu lớn để sản xuất số liệu thống

kê truyền thống trong phạm vi NSI. Các chương trình thử nghiệm chứng minh tiềm năng của dữ liệu lớn và cung cấp kinh nghiệm để thấy được ý nghĩa của dữ liệu lớn đối với số liệu thống kê chính thức. Thứ hai, việc thông qua lộ trình qua ba tầng nhận thức để tổ chức các kế hoạch hành động: ngắn hạn, trung hạn và dài hạn. Các chương trình thử nghiệm sau đó sẽ là một phần trong chiến lược ngắn hạn. Thứ ba, xem xét lộ trình dựa trên bài học kinh nghiệm từ chương trình thử nghiệm và sự phát triển về phương pháp luận, kỹ thuật trong dữ liệu lớn.

## 3. Kinh nghiệm cho đến nay, bao gồm số liệu thống kê chính thức

Việc sử dụng các dữ liệu hoạt động web để dự báo các chỉ tiêu kinh tế-xã hội được đề xuất từ đầu năm 2005 bởi [7] cho tỷ lệ thất nghiệp. Dựa trên ý tưởng phần lớn việc thu thập thông tin liên quan đến công việc được thực hiện thông qua Internet, tác giả nghiên cứu mối quan hệ ở Mỹ từ Báo cáo dữ liệu 500 từ khóa của WordTracker (truy cập http://www.top-keywords.com/longterm.html tháng 9/2014) và tỷ lệ thất nghiệp hàng tháng do Cục Thống kê lao động công bố. Kết quả nghiên cứu cho thấy có sự liên kết quan trọng tích cực giữa công cụ tìm kiếm sử dụng từ khóa dữ liệu và số liệu thất nghiệp chính thức. Tuy nhiên nghiên cứu này không dự báo được tỷ lệ thất nghiệp qua việc sử dụng dữ liệu tìm kiếm web mà mới đơn giản ở mức thiết lập mối tương quan giữa hai nguồn dữ liệu.

### *3.1. Google Trends*

Năm 2006, Google tung ra sản phẩm "Google Trends" (xem [16] ví dụ về thông báo trên phương tiện truyền thông trực tuyến), đây là dịch vụ cung cấp dữ liệu bằng cách nhập vào công cụ tìm kiếm điều kiện cụ thể trong thời gian nhất định. Công cụ ban đầu dùng để nhận biết xu hướng điều khoản, nghĩa là điều khoản theo đó luôn luôn gia tăng số tìm kiếm đã được theo dõi. Tuy nhiên, tính kịp thời cao của Google Trends đã thúc đẩy đáng kể số lượng các nghiên cứu dành riêng cho việc sử dụng nguồn để dự báo các chỉ số kinh tế-xã hội với mục đích đạt được kết quả nhanh hơn so với các công bố của cơ quan thống kê chính thức.

Google công bố năm 2009 trong nhật ký nghiên cứu trên mạng là một trong những nỗ lực đầu tiên dự báo các chỉ tiêu kinh tế-xã hội dựa trên dữ liệu Google Trends. Bài viết sử dụng dữ liệu tìm kiếm để dự báo một số chỉ tiêu ngắn hạn như: doanh số bán xe, doanh số bán lẻ, doanh số bán nhà và số lượng khách truy cập. Kết quả cho thấy đối với mô hình chuỗi thời gian tự động thoái lui đơn giản, giới thiệu các dữ liệu tìm kiếm như dự báo độ chính xác qua các dự báo ngắn hạn của họ. Ngoài các yếu tố dự báo trễ, dữ liệu tìm kiếm hiện tại được sử dụng để dự báo các chỉ tiêu. Khi dữ liệu tìm kiếm qua Google Trends được phát hành với tính kịp thời cao, sau khi tham khảo vài ngày, các mô hình như vậy sẽ cho phép đưa ra dự báo thực tế cho thời điểm hiện tại.

Một số nghiên cứu khác cũng đã sử dụng dữ liệu Google Trends để đưa ra dự báo của một số chỉ tiêu giống nhau và một số chỉ tiêu khác. Giữa các chỉ số khác, chúng ta có thể tìm thấy dịch bệnh cúm [14], thất nghiệp ([10], [9], [25]), và tiêu dùng cá nhân ([15], [13],[22]).

### *3.2. Những bài học từ nghiên cứu dịch cúm trên Google Trends*

Dựa trên các nghiên cứu tập trung vào việc sử dụng hoạt động web để giám sát dịch cúm, Google đưa ra năm 2008, xu hướng dịch cúm trên Google Trends, đã sử dụng bằng cách tổng hợp dữ liệu tìm kiếm của Google để dự đoán diễn biến dịch cúm ở Hoa Kỳ, đem lại tính kịp thời cao hơn so với các chỉ

số công bố từ Trung tâm Kiểm soát và Phòng chống dịch bệnh (CDC).

Kinh nghiệm từ dịch cúm trên Google Trends (GFT) cung cấp các bài học về việc sử dụng dữ liệu tìm kiếm để ước tính nhanh một lĩnh vực của thống kê chính thức. Giữa năm 2009 và giữa năm 2013 GFT hoạt động tốt. Tuy nhiên, năm 2009, GFT đã thất bại trong việc ước lượng chính xác các số liệu chính thức từ CDC bởi theo ước tính tỷ lệ mắc các bệnh cúm, do sự thay đổi trong hành vi tìm kiếm của người dân dẫn đến sự thay đổi thuật toán của GFT. Năm 2013, theo quy luật tự nhiên, dự báo trong mùa cúm cao điểm năm 2012/2013, GFT ước tính tăng gần gấp đôi con số CDC đưa ra. Nguyên nhân chính được chỉ ra là do tin tức từ phương tiện thông tin đại chúng về dịch cúm tàn khốc năm đó.

Điều này tạo ra phản ứng dữ dội đối với dữ liệu lớn. Những phóng đại về tiềm năng về ứng dụng dựa trên dữ liệu rất lớn trở thành những thảo luận về hạn chế của dữ liệu lớn. Tuy nhiên, có những cải tiến tốt để mô hình dự báo GFT hạn chế sai sót xảy ra. Đây là một phần của quá trình xây dựng sản phẩm thống kê đáng tin cậy và GFT có lẽ vẫn chưa sẵn sàng "để sản xuất". Bài học rút ra là khi phát hành một sản phẩm trước khi trưởng thành có thể dẫn đến phá hủy danh tiếng của chính nó. Bài học khác là “tham vọng quá mức về dữ liệu lớn”, tin rằng dữ liệu lớn sẽ thay thế tất cả bộ sưu tập dữ liệu truyền thống. Chìa khóa để khai thác dữ liệu lớn cho số liệu thống kê chính thức được tích hợp trong các hệ thống sản xuất thống kê đa nguồn.

Một bài học khác được rút ra từ kinh nghiệm GFT là sự cần thiết về tính minh bạch và khả năng nhân rộng. Google không tung ra tất cả các ứng dụng của GFT. Ví dụ, không biết đến các thuật ngữ tìm kiếm được sử dụng. Tính minh bạch là một trong những nguyên tắc cơ bản của thống kê chính thức [23], đây là điều cần thiết để giải thích tính chính xác của số liệu thống kê chính thức bởi người sử dụng, bao gồm cả những nhà nghiên cứu muốn đánh giá các số liệu thống kê khi tiến hành nghiên cứu của mình. Khả năng nhân rộng cũng rất quan trọng trong giai đoạn này, nơi học hỏi kinh nghiệm của NSI.

GFT và các ví dụ khác về ứng dụng được đề cập trong phần trước dựa trên Google Trends (GT), chỉ số được tính toán từ các câu hỏi tìm kiếm cá nhân của người sử dụng. Google không cung cấp quyền truy cập vào dữ liệu các câu hỏi tìm kiếm cá nhân. Nhiều chỉ số được tính dựa trên mẫu các câu hỏi tìm kiếm thay đổi hàng ngày [19]. Như một hệ quả, GT trình bày kết quả hơi khác nhau tùy thuộc vào ngày dữ liệu được thu thập và đưa ra nguồn bổ sung không chắc chắn, sai số mẫu (những người khác là tỷ lệ phần trăm người sử dụng tìm kiếm web, tỷ lệ phần trăm người sử dụng dịch vụ của Google và mối quan hệ giữa hành vi tìm kiếm và phân tích các hiện tượng). Một đặc điểm không mong muốn của GT là phương pháp lấy mẫu không được Google tiết lộ, mà thực tế có thể tạo ra hộp đen.

### *3.3. Các nguồn khác về dữ liệu hoạt động web*

Web tìm kiếm dữ liệu, đặc biệt là Google Trends, không phải là nguồn duy nhất của hoạt động trực tuyến được sử dụng để dự báo các chỉ tiêu kinh tế-xã hội. Các lượt truy cập trên Twitter và Wikipedia cũng đã được sử dụng để dự báo các chỉ tiêu kinh tế-xã hội.

Số lượt truy cập trên Wikipedia được sử dụng trong [5] để dự đoán bệnh giống cúm ở Mỹ. So với GFT, mô hình dự báo phát triển tốt hơn trong một số tình huống. Mô hình dự báo dựa trên quan điểm của

Wikipedia xác định những tuần cao điểm của mùa cúm chính xác hơn so với GFT. Tuy nhiên, kết quả dự đoán 4 trong 6 mùa cúm của GFT sát thực tế hơn Wikipedia.

Ví dụ sử dụng Twitter để dự báo số liệu thống kê chính thức [7]. Trong nghiên cứu này, mô hình di cư quốc tế và nội địa được ước lượng từ dữ liệu định vị địa lý từ 500.000 người sử dụng Twitter. Kết quả cho thấy có thể sử dụng phương pháp này để dự báo bước ngoặt trong xu hướng di cư và tăng sự hiểu biết về mối quan hệ giữa di cư nội địa và quốc tế.

### *3.4. Kinh nghiệm trong số liệu thống kê chính thức*

NSI đã bắt đầu khám phá việc sử dụng các dấu vết hoạt động web để dự báo các chỉ tiêu kinh tế-xã hội.

CBS đã nghiên cứu mối quan hệ giữa niềm tin tiêu dùng hàng tháng và ý kiến về tin nhắn trên Facebook và Twitter [20]. Kết quả cho thấy, vì tính kịp thời của các phương tiện truyền thông xã hội và dữ liệu được xử lý nhanh gọn, dự báo về sự tự tin của người tiêu dùng chính thức có thể được công bố trước các số liệu chính thức và ở tần số cao hơn.

ISTAT sử dụng dữ liệu trên Google Trends để dự báo trước một tháng số lượng người tìm kiếm một công việc theo ước tính của Điều tra lực lượng lao động [8].

## 4. Một ví dụ rất đơn giản về ứng dụng với Google Trends

Trong phần này chúng tôi cố gắng hiển thị đơn giản nhất để có thể tích hợp Google Trends (GT) vào mô hình dự báo và vẫn nhận được những cải tiến đáng kể về độ chính xác của dự báo.

Chúng tôi trình bày ví dụ về việc áp dụng chuỗi thời gian GT nhằm cải thiện dự báo thống kê thất nghiệp ở Pháp và Italy. Ở đây dự báo đề cập đến hiện tại (dự báo tức thời) [11]. Thật vậy, các mô hình thảo luận trong việc này được dựa trên [12], [10] và [11], trong đó dữ liệu GT được sử dụng để cải thiện mô hình dự báo đơn giản.

### *4.1. Mô hình*

Chúng ta xem xét hai mô hình:

1) Cơ sở là mô hình tự hồi quy giản đơn, trong đó tỷ lệ thất nghiệp ở tháng t được dự báo bằng cách sử dụng số liệu tỷ lệ thất nghiệp tháng t-1:

$$y_t = a + b*\log y_{t-1} + e_t$$

Trong đó: $y_t$ là tỷ lệ thất nghiệp tháng t, a và b là các hệ số ước lượng, $e_t$ là tổng giá trị phần dư.

2) Mô hình thay thế là mô hình cơ sở điều chỉnh câu hỏi điều kiện $q_i$:

$$y_t = a + b_0*y_{t-1} + \text{sum}_i\ (b_i * q_{i,t})$$

Trong đó: a và $b_i$ là hệ số; $q_{i,t}$ là số lượng tìm kiếm câu hỏi $q_i$ tại thời điểm t.

Tiếp theo, chúng ta lựa chọn thuật ngữ truy vấn mà người sử dụng tìm kiếm trên Google khi thất nghiệp.

Đối với Pháp, chúng tôi đưa ra 3 câu hỏi điều kiện sau đây:

- “pole employ” là cơ quan chính phủ Pháp để người thất nghiệp đăng kí, giúp họ tìm việc làm và đề nghị viện trợ tài chính;

- “Indemnité” đề cập đến việc phân bổ;

- “etre au chomage” là một câu hỏi, chúng tôi tin rằng những người thất nghiệp truy cập nhằm tìm nguồn thông tin hữu ích để cải thiện tình trạng này.

Đối với Italy, chúng tôi đưa ra 4 câu hỏi điều kiện:

- "Impiego" là công việc;

- "Offerte Lavoro" là tuyển dụng việc làm;

- "Curriculum" là thời hạn cho người tìm kiếm việc làm nhằm tìm ra những gợi ý hữu ích để cải thiện cơ hội nhưng vẫn giữ được sự chú ý đối với nhà tuyển dụng;

- "Infojobs" là trang web phổ biến để tham khảo tìm kiếm công việc ở Italy.

### *4.2. Dữ liệu*

Thời gian tải về ngày 16/7/2014. Số liệu chính thức trong điều chỉnh dữ liệu thất nghiệp hàng tháng không theo mùa vụ từ cơ sở dữ liệu Eurostat.

Ở Pháp, dữ liệu GT cho ba thuật ngữ này được tải về từ đường dẫn:

*www.google.fr/trends/explore#q=pole%20emploi&geo=FR&cmpt=q*

*www.google.fr/trends/explore#q=%27indemnit%C3%A9%20chomage%27&geo=FR&cmpt=q*

*www.google.fr/trends/explore#q=%27etre%20au%20chomage%27&geo=FR&cmpt=q*

Dữ liệu hàng tuần với điều kiện "*pole emploi*" và "indemnité" được tổng hợp trên cơ sở hàng tháng. Chỉ sau vài tháng, dữ liệu đã có sẵn trong toàn bộ bốn bộ dữ liệu được lưu giữ để phân tích thêm, gồm 63 tháng kể từ tháng 3/2009 đến tháng 5/2014.

Ở Italy, dữ liệu cho bốn thuật ngữ được tải về từ đường dẫn:

*www.google.fr/trends/explore#cat=0-958-60&q=impiego&geo=IT&cmpt=q*

*www.google.fr/trends/explore#cat=0-958-60&q=%27offerte%20lavoro%27&geo=IT&cmpt=q*

*www.google.fr/trends/explore#cat=0-958-60&q=curriculum&geo=IT&cmpt=q*

*www.google.fr/trends/explore#cat=0-958-60&q=infojobs&geo=IT&cmpt=q*

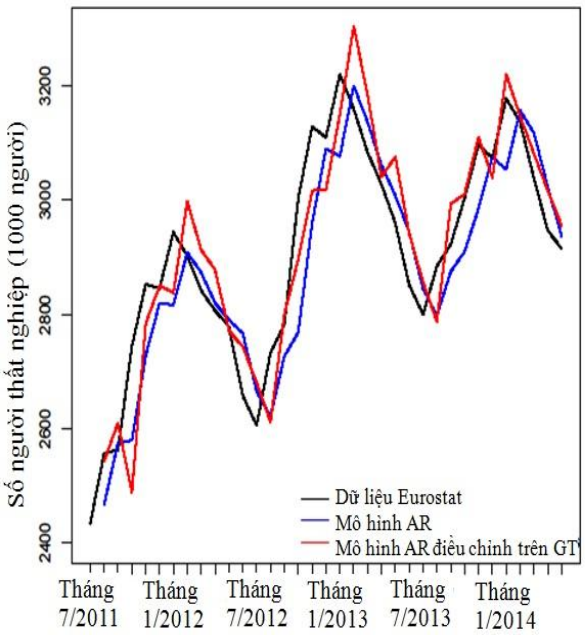
Sau vài tháng, dữ liệu đã có sẵn trong toàn bộ bốn bộ dữ liệu được lưu giữ để phân tích thêm, gồm 77 tháng kể từ tháng 1/2008 đến tháng 5/2014.

### *4.3. Kết quả ở Pháp*

Ở các kết quả tiếp theo, tất cả tính toán được thực hiện trên phần mềm R.

Ở mỗi tháng t sau tháng 8/2011 chúng tôi gán hai mô hình trên tất cả các tháng trước đó (tức là từ tháng 8/2011 là t-1) và dự đoán tỷ lệ thất nghiệp ở tháng t.

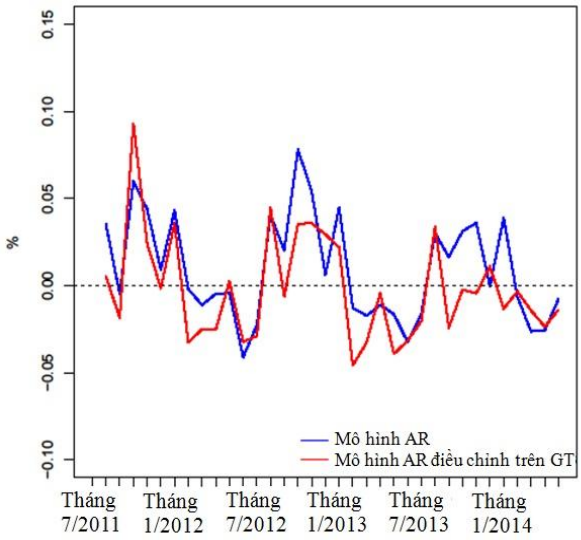
Hình 1: Giá trị dự báo tháng t ở hai mô hình dựa trên dữ liệu các tháng trước đây




Hình 1 cho thấy mô hình điều chỉnh phù hợp với dữ liệu thực tế hơn so với mô hình AR đơn giản, thể hiện bằng mức trung bình của giá trị tuyệt đối

của sai số dự đoán tương đối (còn gọi là sai số bình quân): $MAE_{AR}$ = 2.5% và $MAE_{AR\ điều\ chỉnh}$ = 2.4%. Hệ số tương quan Pearson $r_{AR}$ = 0.88 và $r_{AR\ điều\ chỉnh}$ = 0.9.

Hình 2: Tỷ lệ sai số theo % (thực tế giá trị - giá trị dự đoán) / giá trị thực tế; mô hình xây dựng dựa trên dữ liệu các tháng trước đây



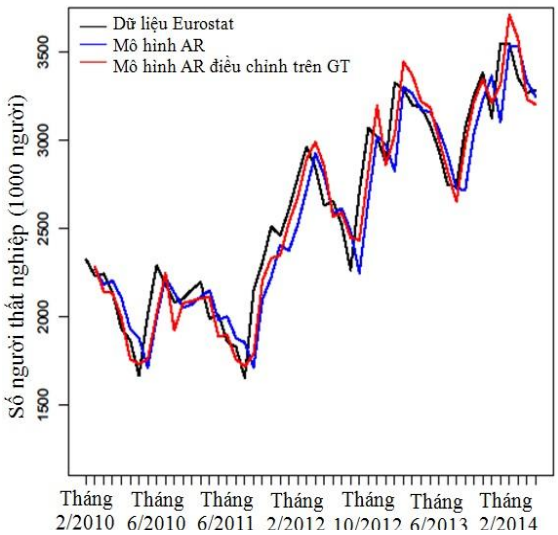

Hình 2 cho thấy sai số tương đối hai mô hình: rõ ràng mô hình điều chỉnh tốt hơn mô hình đơn giản sau vài tháng. Điều này có thể do trên thực tế, mô hình điều chỉnh có nhiều hệ số ước lượng và có nhiều quan sát hơn (tháng). Sai số theo mùa vụ (con số không hiển thị), cho biết cả hai mô hình cần cải tiến mạnh mẽ.

### *4.4. Kết quả ở Italy*

Trong trường hợp của Italy, khả năng giải thích câu hỏi điều kiện nhằm nâng cao hiệu quả các mô hình dự báo cơ sở.

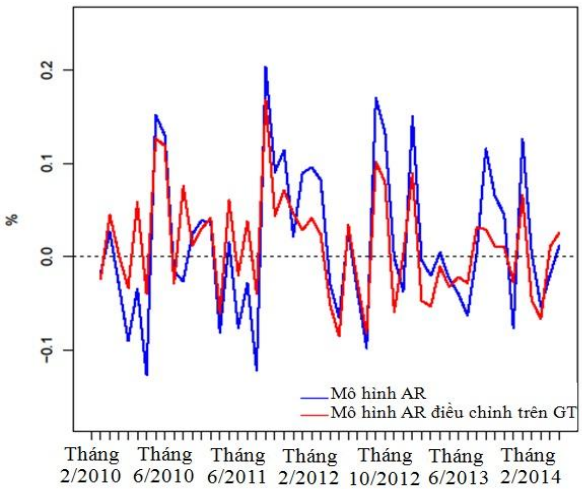
Sai số bình quân $MAE_{AR}$ = 6.3% ($r_{AR}$ = 0.93) và $MAE_{AR\ điều\ chỉnh}$ = 4.7% ($r_{AR\ điều\ chỉnh}$ = 0.97), xem Hình 3.

Hình 3: Giá trị dự báo tháng t ở hai mô hình dựa trên dữ liệu các tháng trước đây




Điều này được xác nhận bởi sai số tương đối thể hiện trong Hình 4

Hình 4: Tỷ lệ sai số theo % (thực tế giá trị - giá trị dự đoán) / giá trị thực tế; mô hình xây dựng dựa trên dữ liệu các tháng trước đây




## 5. Kinh nghiệm về các chỉ tiêu nhanh trong Eurostat

### *5.1. Ước tính nhanh HICP khu vực đồng Euro là gì?*

Ước tính nhanh HICP khu vực đồng Euro (hài hòa chỉ số giá tiêu dùng) từ trên xuống cho các phần chính là sản phẩm thống kê được sản xuất hàng tháng và là một trong những chỉ số đáng chú ý nhất

do Eurostat tạo ra. Vào ngày cuối tháng (nếu ngày cuối tháng rơi vào ngày cuối tuần, thì được công bố vào ngày làm việc tiếp theo), giá trị lạm phát ước tính của tháng được công bố. Kể từ tháng 9/2012 Eurostat đã thường xuyên công bố các ước tính nhanh không chỉ cho tất cả các mặt hàng, mà còn cho các thành phần chính. Tháng 9/2014, 3 thành phần chính được bổ sung vào giỏ hàng hiện tại cho các ước tính nhanh, giỏ hàng bao gồm: "tất cả mặt hàng", "thức ăn", "thực phẩm đã qua chế biến", "thực phẩm chưa qua chế biến", "hàng hóa công nghiệp phi năng lượng", "năng lượng", "dịch vụ", "tất cả mặt hàng trừ năng lượng", "tất cả mặt hàng trừ năng lượng và thực phẩm" và "tất cả mặt hàng trừ năng lượng và thực phẩm chưa qua chế biến".

Ước tính nhanh giá trị lạm phát là chỉ số quan trọng cho công chúng, thị trường tài chính nói chung nhưng quan trọng nhất đối với Ngân hàng Trung ương châu Âu (ECB). Trên thực tế, các ước tính nhanh là một yêu cầu từ ECB, cập nhật giá trị lạm phát mới nhất trong cuộc họp Hội đồng quản trị ECB, chịu trách nhiệm xây dựng chính sách tiền tệ khu vực đồng Euro.

Khi tính toán một chỉ số quan trọng như vậy, chú ý thêm về chất lượng một cách bao quát hơn là hết sức cần thiết. Độ chính xác chỉ là một phần phản ánh chất lượng nhưng tính kịp thời cũng có liên quan. Ngoài việc có thể công bố các ước tính trong ngày định trước, nó cũng quan trọng để không bị nhỡ ấn phẩm. Khi sản xuất đã bắt đầu thì không thể dừng lại.

### *5.2. Làm thế nào để tính toán*

Ước tính nhanh HICP khu vực đồng Euro kết hợp thông tin ban đầu gửi từ một số nước thành viên cùng với dữ liệu dự báo từ những nước còn lại. Trong hầu hết trường hợp, "thông tin ban đầu" là ước tính sơ bộ dựa trên giá thu thập sẽ là một phần bộ dữ liệu HICP cuối cùng nhưng đưa vào quá trình sản xuất từ sớm, ví dụ: dữ liệu không hoàn toàn hợp lệ, không điều chỉnh chất lượng thực hiện, vv… Từ khi dữ liệu sơ bộ căn cứ trên cùng một giá thu thập hơn chỉ số HICP cuối cùng, đó không phải là điều bất ngờ, nó rất chính xác. Thực tế đã chứng minh dữ liệu sơ bộ chính xác hơn nhiều so với bất kỳ mô hình dự báo cơ sở nào. Do đó, dữ liệu sơ bộ luôn luôn được ưa thích hơn.

Dữ liệu sơ bộ có thể được chú trọng hơn do các thủ tục ước tính nhanh phát triển tại Eurostat sửa chữa bất cứ khi nào có thể với một quy trình định cỡ phát triển cho mục đích cụ thể.

Thật không may, không phải tất cả các nước đều có thể cung cấp dữ liệu sơ bộ đúng thời gian: các quốc gia cần phải dự báo dữ liệu bị mất.

Các thành phần chính khác nhau của sự lạm phát có những hành vi ngẫu nhiên rất khác biệt, một số trong đó không ổn định và khó dự đoán. Như vậy, mỗi thành phần được xử lý riêng và bất kỳ dữ liệu phụ nào đều có thể cải thiện các dự báo được tính đến. Các dữ liệu phụ sử dụng bởi ước tính nhanh là giá năng lượng mục Bản tin Dầu hàng tuần, sản xuất bởi Tổng cục Năng lượng của Ủy ban Châu Âu (DG ENER), nguồn dữ liệu hành chính.

Do thời gian ngắn, những ước tính nhanh thường không quá 3 giờ, công cụ dự báo tự động do Eurostat phát triển.

### *5.3. Các ước tính nhanh là ví dụ toàn diện về việc sử dụng dữ liệu lớn trong thống kê chính thức*

Ước tính nhanh HICP khu vực đồng Euro không sử dụng dữ liệu lớn. Tuy nhiên, cần phải sử dụng nguồn dữ liệu hành chính để khắc phục vấn

đề tin tức (khu vực đồng Euro hoàn toàn không được bao phủ bởi dữ liệu sơ bộ), có thể cung cấp như là một ví dụ cho thấy khả năng sử dụng dữ liệu lớn trong sản xuất số liệu thống kê chính thức thường xuyên.

Dữ liệu phụ sử dụng trong ước tính nhanh là rất hữu ích do một số yếu tố:

- Chi phí rẻ: đây không phải là các nguồn dữ liệu Eurostat cần thu thập, biên soạn, vv… Eurostat chỉ cần lấy về;

- Thường xuyên: hàng tuần DG ENER công bố bản cập nhật về giá năng lượng;

- Dễ dàng có sẵn: dữ liệu có sẵn miễn phí trên web cho bất cứ ai muốn sử dụng.

Một khía cạnh quan trọng khác là Bản tin Dầu hàng tuần dự định có mục đích khác so với ước tính nhanh HICP khu vực đồng Euro: Mục đích chính để cải thiện tính minh bạch của giá dầu và củng cố thị trường nội địa. Tuy nhiên, dữ liệu hiện đang sử dụng cũng như để cải tiến tính toán chỉ số lạm phát, ứng dụng không được lường trước khi DG ENER triển khai tập hợp dữ liệu này.

Tuy nhiên, việc sử dụng nguồn dữ liệu hành chính này có thể do hai sự kiện rất quan trọng:

- Dữ liệu có sẵn một cách thường xuyên, không bị gián đoạn. Đây là khía cạnh rất quan trọng vì các ước tính nhanh "một khi bắt đầu không thể dừng lại". Eurostat không thể đủ khả năng cung cấp người sử dụng chỉ số quan trọng như vậy trong khoảng thời gian dài vì không có sẵn nguồn dữ liệu thay thế. Hơn nữa, ngay cả nếu có tính hiệu lực công nhận DG ENER (có hành động pháp lý bắt buộc các nước thành viên phải báo cáo giá năng lượng: Hội đồng quyết định ngày 22/4/1999) bởi một số lý do không có sẵn, kế hoạch dự phòng, ví dụ: Giá dầu thô Brent châu Âu;

- Điều đó là chắc chắn, có sự tương quan ổn định quan trọng hơn giữa một số thành phần chính HICP và nguồn dữ liệu hành chính. Đây cũng là khía cạnh rất quan trọng vì Eurostat không thể đủ khả năng để sản xuất số liệu thống kê với độ chính xác đáng tin cậy, sau một vài tháng tính chính xác giảm đi gây nguy hiểm nếu được phát hành.

Một khía cạnh khác về nguồn dữ liệu hành chính rất quan trọng, đó là việc sử dụng thành công trong ước tính nhanh. Bản tin Dầu hàng tuần là giá tham khảo các sản phẩm năng lượng, trong đó liên quan nhiều đến mức giá trung bình mà người tiêu dùng trả. Vì vậy, khi sử dụng nguồn dữ liệu hành chính hầu như không có nguy cơ gây nhầm lẫn với dữ liệu gây nhiễu khác, dường như có thể liên quan với các chỉ số HICP. Đây có thể không phải là trường hợp khi chúng ta nói về nguồn dữ liệu lớn và/hoặc sự kết nối giữa hai nguồn dữ liệu không quá rõ ràng.

Thực hiện song song giữa việc sử dụng dữ liệu hành chính trong sản xuất số liệu thống kê chính thức và khả năng sử dụng dữ liệu lớn, có thể kết luận như sau:

- Có thể có rất nhiều dữ liệu được tạo ra cho nhiều mục đích khác số liệu thống kê chính thức, nhưng thực chất lại là bộ phận quan trọng của quy trình sản xuất số liệu thống kê chính thức. Chúng tôi, các nhà thống kê chính thức, chỉ có thể khuyến khích và thúc đẩy để tìm chúng;

- Đồng thời chúng tôi, như các nhà thống kê chính thức phải biết chọn lọc khi kết hợp nguồn dữ liệu không theo quy ước trong sản xuất số liệu thống kê chính thức. Trước khi kết hợp thêm nguồn dữ liệu, cần phải trả lời hai câu hỏi quan trọng:

+ Nguồn dữ liệu lớn sẽ có sẵn trong tương lai nên tôi đảm bảo rằng tôi có thể công bố số liệu thống kê chính thức mà không bị buộc phải dừng lại sau một vài lần phát hành?

+ Những gì tôi đang chiết xuất từ các dữ liệu khổng lồ có sẵn thực sự là một dấu hiệu hay nó chỉ là dữ liệu vô nghĩa? Và nếu đó là một dấu hiệu, liệu có thể đo lường các hiện tượng mà tôi muốn?

**6. Chương trình giới thiệu dữ liệu hoạt động web trong tính toán các ước tính nhanh**

Sử dụng dữ liệu hoạt động web được nêu trong bài viết này (Google Trends) không phải là khó khăn để cải thiện các dự báo của mô hình chuỗi thời gian đơn giản. Bài viết cũng chỉ ra có rất nhiều tài liệu cho thấy trường hợp mô hình cơ sở được cải thiện bằng cách sử dụng nguồn dữ liệu lớn này, ngay cả khi các tài liệu là không thống nhất.

Tuy nhiên, việc sử dụng các nguồn như Google Trends để tính toán thường xuyên các ước tính nhanh của số liệu thống kê chính thức đặt ra những thách thức mà chúng ta cần phải giải quyết. Tiếp đó, bước tiếp theo chúng ta cần làm gì để tích hợp các nguồn dữ liệu hoạt động web trong tính toán ước tính nhanh chính thức?

***6.1. Nghiên cứu cân bằng về sử dụng nguồn dữ liệu hoạt động web cho dự báo***

Như đã chỉ ra bởi [3], kết quả nghiên cứu thường được trình bày khi việc sử dụng các dữ liệu hoạt động web cải thiện thành công quá trình ước tính giá trị của một biến ở thời điểm hiện tại, dự báo tức thời, nhưng khi kết quả nghiên cứu thành công thì không được phổ biến. Vì vậy, đọc các tài liệu về chủ đề này cung cấp một cái nhìn tổng quan không cân bằng về tiềm năng nói chung của loại dữ liệu để dự báo các chỉ tiêu kinh tế - xã hội.

Tìm kiếm các kết quả khả quan là một mình thông tin liên quan mời nghiên cứu hơn về vấn đề này. Tuy nhiên, để có ý tưởng chính xác hơn về tiềm năng của loại dữ liệu này, có thể tác động đầu tư hơn vào Viện Thống kê Quốc gia, nghiên cứu cân bằng là bắt buộc. Một số nghiên cứu giống như giới thiệu trong [3] và [19] cung cấp một cái nhìn tổng quan cân bằng bằng cách bao gồm nhiều quốc gia và nhiều chỉ tiêu. Bước tiếp theo sẽ được chỉ nghiên cứu cân bằng quy mô lớn hơn bao gồm một số chỉ tiêu kinh tế-xã hội và một số quốc gia theo cách tiếp cận tương tự, trong đó sẽ đưa ra các kết quả cả tích cực và tiêu cực, như vậy có thể đánh giá một cách tổng thể.

***6.2. Sự đa dạng hoá và sự đánh giá các nguồn dữ liệu về hoạt động web***

Các nguồn dữ liệu lớn, đặc biệt là nguồn dữ liệu hoạt động web, đưa ra nhiều thách thức đối với một số nguyên tắc hướng dẫn số liệu thống kê chính thức (ở đây chúng tôi làm theo Luật Thống kê châu Âu về thực hành - CoP). Như các nguồn thứ cấp bên ngoài, chúng được thoát khỏi sự kiểm soát của NSI. Trong trường hợp các nguồn truyền thống, NSI hoặc có kiểm soát đầy đủ trong trường hợp khảo sát hoặc có mức độ ảnh hưởng nhất định tùy từng quốc gia, vì đó là trường hợp hồ sơ hành chính. Sự thiếu kiểm soát đó đưa ra một số rủi ro.

Thứ nhất, đó là nguy cơ nguồn dữ liệu là hộp đen. NSI cố gắng làm cho tài liệu càng đầy đủ càng tốt cho quy trình sản xuất các số liệu thống kê chính thức. Sự minh bạch này là cần thiết để giữ mức độ tin tưởng của xã hội và các bên liên quan đến chính trị về các số liệu thống kê chính thức. Tuy nhiên, trong trường hợp các nguồn dữ liệu lớn do các công ty tư nhân nắm giữ thì có thể không đảm bảo cùng một mức độ minh bạch. Đây là yêu cầu trong một số

trường hợp việc tiết lộ xử lý dữ liệu dịch vụ web có thể đẩy nhà cung cấp dữ liệu đến bất lợi cạnh tranh trên thị trường này.

Thứ hai, trừ khi NSI kiểm toán triệt để việc xử lý dữ liệu dịch vụ web, nó không thể đảm bảo rằng các nguồn không phải là đối tượng để thao tác, bất kể các thao tác đó diễn ra hay không. Một kiểm toán quá kỹ càng có thể không thực hiện được (nếu nhà cung cấp dữ liệu nằm ngoài thẩm quyền của cơ quan quản lý thống kê) hoặc rất tốn kém.

Thứ ba, nguồn dữ liệu có thể thường xuyên bịu ngắt trong chuỗi series. Thủ tục xử lý dữ liệu của các dịch vụ web được thiết kế theo nhu cầu của doanh nghiệp và thay đổi theo thời gian. Như đã chỉ ra trong [4], đây là trường hợp của Google, kể từ khi tung ra Google Trends năm 2006, Google đã điều chỉnh một số thuật toán làm ảnh hưởng đến dữ liệu đã có sẵn thông qua Google Trends.

Thứ tư, nguy cơ thiếu tính liên tục do NSI không thể đảm bảo nguồn sẽ có sẵn lâu dài khi cần thiết. Tính hữu ích của các dữ liệu từ các dịch vụ web cụ thể, chẳng hạn như một công cụ tìm kiếm phụ thuộc trực tiếp vào sự nổi tiếng của nó mà thay đổi theo thời gian. Sự sẵn có của nguồn cũng có thể bị phá vỡ bởi những thay đổi công nghệ không nằm dưới sự kiểm soát của NSI.

Một số rủi ro có thể được giảm bớt bằng cách sử dụng kết hợp nhiều nguồn dữ liệu hoạt động web trong các mô hình dự báo. Điều này làm giảm ảnh hưởng của các nguồn dữ liệu cá nhân, NSI không kiểm soát, trong các giá trị dự báo và cung cấp một sự đảm bảo rằng ước tính nhanh chính thức không bị can thiệp vào. Sự đa dạng của các nguồn cũng cho phép xây dựng các kế hoạch dự phòng cho sự thiếu liên tục của một số nguồn. Ví dụ, trong trường hợp ước tính nhanh tỷ lệ việc làm, một nguồn có thể thực hiện, chưa kể những nguồn đã được đề cập trong bài viết này, có thể truy cập các trang web liên quan đến việc làm.

Chúng ta cũng cần đánh giá lại một cách thường xuyên các mô hình dự báo để thích ứng với sự gián đoạn trong chuỗi series.

Cuối cùng, thành lập các thủ tục cho việc kiểm định và chứng nhận các nguồn dữ liệu lớn cho các số liệu thống kê chính thức [2], nên được thiết lập để đảm bảo tính minh bạch và chất lượng của các nguồn.

### *6.3. Tích hợp dữ liệu hoạt động web với các nguồn dữ liệu thống kê chính thức truyền thống*

Một số ví dụ về dự báo các chỉ tiêu kinh tế-xã hội nêu trong bài viết này và hầu hết trong số đó không do các văn phòng chính thức thống kê (Viện thống kê quốc gia và các cơ quan thống kê châu Âu và quốc tế) thực hiện. Câu hỏi chính đáng đặt ra là lý do tại sao các cơ quan thống kê chính thức không tự làm nếu người khác có thể làm được điều đó.

Trong bài viết này, chúng tôi không cố gắng trả lời câu hỏi này. Những gì chúng ta lập luận là nếu thống kê chính thức cung cấp ước tính nhanh các chỉ tiêu kinh tế-xã hội bằng cách sử dụng mô hình dự báo dựa trên dữ liệu hoạt động web, thì mô hình này không nên chỉ sử dụng đơn giản như là tái sử dụng những thứ mà các mô hình khác có thể làm, thay vào đó có thể tận dụng những ưu điểm tương đối đặc trưng của nó.

Ưu thế tương đối rõ nhất của các cơ quan thống kê chính thức thể hiện ở chỗ họ là những người sản xuất ra các chỉ tiêu thống kê chính thức, là vị trí tốt nhất để biết đặc trưng riêng của các chỉ tiêu, trong một số trường hợp có dữ liệu tạm thời (như trường hợp ước tính nhanh giá trị lạm phát) mà cũng

có thể được đưa ra trong mô hình. Ưu thế tương đối nữa là kinh nghiệm qua các cuộc khảo sát, và trong trường hợp cụ thể của NSIs, trên thực tế họ có hệ thống thu thập dữ liệu lớn.

Do đó, các cơ quan thống kê chính thức nên tích hợp tính toán các ước tính nhanh trong hệ thống sản xuất thống kê thường xuyên của mình. Nghĩa là có thể sử dụng thông tin chi tiết hơn về các chỉ tiêu so với công bố. Các cuộc khảo sát cũng có thể được điều chỉnh để họ cung cấp những thông tin giúp việc sử dụng bằng chứng hoạt động web hay chính xác hơn trong nguồn dữ liệu lớn.

### *6.4. Nghiên cứu về mối quan hệ giữa hoạt động web và các hiện tượng được dự báo*

Tính thiết thực của giá trị dự báo dựa trên các mô hình dự báo từ dữ liệu hoạt động web chỉ thực sự được đảm bảo nếu có sự hiểu biết tốt về mối quan hệ giữa hiện tượng được dự báo với hoạt động web cá nhân. Vì vậy, chương trình về việc đưa ra loại nguồn trong tính toán các các ước tính nhanh cần phải đi kèm với các nghiên cứu về chủ đề này.

### *6.5. Nỗ lực chung về sự phát triển của các mô hình dự báo thích hợp*

Mặc dù trong bài viết này, chúng tôi tập trung vào những thách thức của việc sử dụng dữ liệu hoạt động web trong dự báo các chỉ tiêu kinh tế-xã hội, sự phát triển của các mô hình dự báo thích hợp cũng rất quan trọng. Những mô hình chúng tôi trình bày trong bài viết này rất đơn giản và chỉ phục vụ cho mục đích minh họa, có thể dùng để cải thiện tính chính xác của các giá trị dự báo với dữ liệu tìm kiếm web từ Google Trends. Việc sử dụng này trong các ước tính nhanh sẽ đòi hỏi các mô hình phức tạp hơn, có thể bao gồm nhiều biến khác.

Để đảm bảo tính minh bạch, các "mô hình sản xuất" cần thảo luận một cách cởi mở giữa các bên liên quan, như các nhà hoạch định chính sách trong Ủy ban châu Âu và Ngân hàng Trung ương châu Âu trong trường hợp của châu Âu, và giữa các cơ quan thống kê với các nhà nghiên cứu, học gi, nhằm thống nhất mô hình chung để đánh giá và so sánh quốc tế từ các bài học kinh nghiệm.

### *Tài liệu tham khảo:*

[1] D. Butler, When Google got flu wrong., Nature Vol. 494 N. 7436 (2013), 155, http://www.nature.com/news/when-google-got-flu-wrong-1.12413, last accessed on 30 September 2014;

[2] D. Florescu and M. Karlberg and F. Reis and P.R. Del Castillo and M. Skaliotis and A. Wirthmann, Will 'big data'transform official statistics? (2014),

http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf, last accessed on 30 September 2014;

[3] D. Gayo-Avello, I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"--A Balanced Survey on Election Prediction using Twitter Data", arXiv preprint arXiv:1204.6441 (2012), http://arxiv.org/abs/1204.6441, last accessed on 30 September 2014;

[4] D. Lazer and R. Kennedy and G. King and A. Vespignani, The Parable of Google Flu: Traps in Big Data Analysis, Science Vol. 343 N. 41712 (2014),

http://dash.harvard.edu/bitstream/handle/1/12016836/The%20Parable%20of%20Google%20Flu%20%28WP-Final%29.pdf, last accessed on 30 September 2014;

[5] D.J. McIver and J.S. Brownstein, Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time, PLoS computational biology Vol. 10 N. 4 (2014), e1003581, http://www.ploscompbiol.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pcbi.1003581&representation=PDF, last accessed on 30 September 2014

[6] D.R. Olson and K.J. Konty and M. Paladini and C. Viboud and L. Simonsen, Reassessing google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales, PLoS computational biology Vol. 9 N. 10 (2013), e1003256, http://www.ploscompbiol.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pcbi.1003256&representation=PDF, last accessed on 30 September 2014

[7] E. Zagheni and V.R.K. Garimella and I. Weber and B. State, Inferring international and internal migration patterns from Twitter data (2014), 439--444, http://ingmarweber.de/wp-content/uploads/2014/02/Inferring-International-and-Internal-Migration-Patterns-from-Twitter-Data.pdf, last accessed on 30 September 2014;

[8] F. Bacchini and M. D'Alò and S. Falorsi and A. Fasulo and C. Pappalardo, Does Google index improve the forecast of Italian labour market? (2014),

http://www.sis2014.it/proceedings/allpapers/3019.pdf, last accessed on 30 September 2014;

[9] F. D'Amuri and J. Marcucci, Google it!" Forecasting the US unemployment rate with a Google job search index", ISER Working Paper Series (2009),

http://www.luiss.edu/dptea/files/Paper_Juri_Marcucci.pdf, last accessed on 30 September 2014;

[10] H. Choi and H.R. Varian, Predicting initial claims for unemployment benefits, Google Inc (2009), http://static.googleusercontent.com/media/research.google.com/fr//archive/papers/initialclaimsUS.pdf, last accessed on 30 September 2014

[11] H. Choi and H.R. Varian, Predicting the present with google trends, Economic Record Vol. 88 N. s1 (2012), 2--9, http://onlinelibrary.wiley.com/doi/10.1111/j.1475-4932.2012.00809.x/pdf, last accessed on 30 September 2014

[12] H. Choi and H.R. Varian, Predicting the present with Google Trends, Google Research Blog (2009), http://static.googleusercontent.com/media/www.google.com/fr//googleblogs/pdfs/google_predicting_the_present.pdf, last accessed on 30 September 2014

[13] I.J. Toth and M. Hajdu, Google as a tool for nowcasting household consumption: estimations on Hungarian data Vol. 7 (2013),

http://m.gvi.hu/data/research/ciret_2012_tij_hm_paper_120415.pdf, last accessed on 30 September 2014

[14] J. Ginsberg and M.H. Mohebbi and R.S. Patel and L. Brammer and M.S. Smolinski and L. Brilliant, Detecting influenza epidemics using search engine query data, Nature Vol. 457 N. 7232 (2009), 1012--1014, http://www.nature.com/nature/journal/v457/n7232/pdf/nature07634.pdf, last accessed on 30 September 2014;

[15] K.A. Kholodilin and M. Podstawski and B. Siliverstovs, Do Google searches help in nowcasting private consumption? A real-time evidence for the US (2010),

https://www.econstor.eu/dspace/bitstream/10419/36734/1/625127439.pdf, last accessed on 30 September 2014;

[16] M. Arrington, Google Trends Launches (2006), http://techcrunch.com/2006/05/10/google-trends-launches/, last accessed on 30 September 2014;

[17] M. Ettredge and J. Gerdes and G. Karuga, Using web-based search data to predict macroeconomic statistics, Communications of the ACM Vol. 48 N. 11 (2005), 87-92,

http://www.dsi.unive.it/~orlando/Topic-WSE-Queries/p87-ettredge.pdf, last accessed on 30 September 2014;

[18] M. Ojala, Searching for Business Trends and Trending Topics, Online Vol. Vol. 33 N. No. 6 (2009), http://www.questia.com/magazine/1G1-211794596/searching-for-business-trends-and-trending-topics, last accessed on 30 September 2014;

[19] N. Barreira and P. Godinho and P. Melo, Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends, NETNOMICS: Economic Research and Electronic Networking Vol. 14 N. 3 (2013), 129-165,

http://link.springer.com/content/pdf/10.1007%2Fs11066-013-9082-8.pdf, last accessed on 30 September 2014

[20] P.J.H. Daas and M.J.H. Puts, Social media sentiment and consumer confidence, ECB Statistics Paper Series (2014),

http://www.pietdaas.nl/beta/pubs/pubs/Daas_Puts_Sociale_media_cons_conf_Stat_Neth.pdf, last accessed on 30 September 2014

[21] S. Cook and C. Conrad and A.L. Fowlkes and M.H. Mohebbi, Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic, PloS one Vol. 6 N. 8 (2011), e23610,

http://www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0023610&representation=PDF, last accessed on 30 September 2014

[22] S. Vosen and T. Schmidt, Forecasting private consumption: survey-based indicators vs. Google trends, Journal of Forecasting Vol. 30 N. 6 (2011), 565--578,

http://www.econstor.eu/bitstream/10419/29900/1/614061253.pdf, last accessed on 30 September 2014;

[23] Statistical Commission of the United Nations, Fundamental Principles of Official Statistics (2013), http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf, last accessed on 30 September 2014;

[24] T. Harford, Big Data: are We Making a Big Mistake, Financial Times Magazine (2014), http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2xlNF6IjV, last accessed on 30 September 2014;

[25] Y. Fondeur and F. Karamé, Can Google data help predict French youth unemployment?, Economic Modelling Vol. 30 (2013), 117-125, http://site.univ-evry.fr/modules/resources/download/default/Recherche/Les%20laboratoires/epee/wp/12-03.pdf, last accessed on 30 September 2014;