



NÂNG CAO HIỆU QUẢ DỰ BÁO NHU CẦU DU LỊCH VỚI GOOGLE TRENDS

Ngô Văn Sơn¹, Lê Văn Hoài², Thái Thị Phương³, Hoàng Trọng Lợi⁴, Võ Viết Minh Nhật^{5*}

¹ Trường Đại học Khoa học, Đại học Huế, Việt Nam

² Trường Du lịch – Đại học Huế, Việt Nam

³ Ban Đào tạo và Công tác Sinh viên, Đại học Huế, Việt Nam

⁴ Khoa Kỹ thuật và công nghệ - Đại học Huế, Việt Nam

⁵ Viện Khảo thí và Bảo đảm chất lượng giáo dục - Đại học Huế, Việt Nam

Tóm tắt. Sự phát triển nhanh chóng của Internet tạo ra một lượng dữ liệu khổng lồ. Phân tích dữ liệu lớn để thu được thông tin hữu ích đang thu hút nhiều sự quan tâm. Một hướng nghiên cứu thú vị gần đây là sử dụng "dữ liệu công cụ tìm kiếm", ở đó kết quả phân tích dữ liệu tìm kiếm đã đem lại nhiều thông tin kinh doanh có giá trị. Google Trends là một mục tiêu phổ biến trong việc nghiên cứu dữ liệu công cụ tìm kiếm, vì nó có sẵn và dễ truy cập. Bài báo phân tích phương pháp dự báo dựa trên dữ liệu công cụ tìm kiếm, bao gồm thu thập dữ liệu từ Google Trends, khám phá cách sử dụng dữ liệu và mô tả làm thế nào đưa ra kết quả dự báo. Bài báo sử dụng dữ liệu du khách quốc tế đến Việt Nam để đánh giá hiệu quả của phương pháp tiếp cận kết hợp dựa trên dữ liệu lịch sử và Google Trends.

Từ khóa: dự báo nhu cầu du lịch; dữ liệu công cụ tìm kiếm; Google Trends; du lịch Việt Nam

Enhancing tourism demand forecasting effectiveness with Google Trends

Ngo Van Son¹, Le Van Hoai², Thai Thi Phuong³, Hoang Trong Loi⁴, Vo Viet Minh Nhat^{5*}

¹ University of Sciences, Hue University, Vietnam

² School of Hospitality & Tourism - Hue University, Vietnam

³ Department of Training and Student Affairs, Hue University, Vietnam

⁴ Faculty of Engineering and Technology – Hue University, Vietnam

⁵ Institute of Testing and Education Quality Assurance – Hue University, Vietnam

Abstract. The rapid development of the Internet has generated an enormous amount of data. Analyzing big data to obtain helpful information has attracted much attention. An interesting recent research direction is “search engine query data”, where the results of analyzing search data have provided valuable business information. Google Trends is a popular target in search engine data research because it is available and easily accessible. This paper analyzes the forecasting method based on search engine data, including collecting data from Google Trends, exploring how to use the data, and describing how

* Liên hệ: vvmnhat@hueuni.edu.vn

to make forecasts. The paper uses international tourist arrival data from Vietnam to evaluate the effectiveness of a combined approach based on historical data and Google Trends.

Keywords: Tourism demand forecasting, search engine query data, Google Trends, Vietnam tourism

1 Giới thiệu

Internet đã trở nên phổ biến hiện nay. Nhiều hoạt động kinh doanh đã chuyển từ các giao dịch vật lý sang môi trường trực tuyến với nền tảng web [1]. Sự phát triển của công nghệ thông tin đã tạo ra một lượng lớn dữ liệu lớn từ người dùng, như các bình luận, bình chọn, các truy vấn tìm kiếm hay các bài phân tích. Đối với các doanh nghiệp, dữ liệu lớn có thể có ý nghĩa quan trọng trong việc hỗ trợ người quản lý ra quyết định. Các nhà quản lý đã bắt đầu nhận thức được tầm quan trọng và giá trị của "*dữ liệu lớn*" và một hướng nghiên cứu liên quan đến "*dữ liệu công cụ tìm kiếm*" cũng đã xuất hiện. Thực tế, đã có khá nhiều nghiên cứu về khai thác giá trị của dữ liệu công cụ tìm kiếm xuất hiện trong những năm trở lại đây [2][3][4].

Google và Baidu là hai công cụ tìm kiếm cung cấp dữ liệu tìm kiếm lớn nhất thế giới. Baidu chủ yếu phổ biến tại Trung Quốc, trong khi Google được sử dụng khắp toàn cầu. Với khả năng cung cấp dữ liệu có sẵn và dễ truy cập, Google Trends trở thành mục tiêu để nghiên cứu dữ liệu công cụ tìm kiếm. Trong bài báo này, chúng tôi thu thập dữ liệu nhu cầu du lịch (khách du lịch quốc tế đến) Việt Nam từ Google Trends [5].

Du lịch là một ngành công nghiệp không khói, đóng vai trò quan trọng trong nền kinh tế của nhiều địa phương và quốc gia cả trực tiếp và gián tiếp bằng cách cung cấp việc làm và nguồn thu ngoại tệ. Điều này giải thích tại sao rất nhiều quốc gia nhỏ phụ thuộc nhiều vào du lịch. Dự báo chính xác nhu cầu du lịch do đó là một nhu cầu cấp thiết đối với các quốc gia và địa phương nhằm có thể xây dựng chính sách, hoạch định kế hoạch, chuẩn bị nguồn lực nhằm cung cấp tốt nhất các sản phẩm và dịch vụ cho khách du lịch. Các mô hình dự báo truyền thống dựa trên dữ liệu lịch sử khách du lịch đến không còn đảm bảo độ chính xác và xu hướng kết hợp với những loại dữ liệu khác, chẳng hạn dữ liệu công cụ tìm kiếm, trở thành xu hướng mới cho dự báo du lịch trong kỷ nguyên dữ liệu lớn [6][7].

Nghiên cứu trong bài báo này có mục đích khám phá việc kết hợp dữ liệu công cụ tìm kiếm với dữ liệu lịch sử khách du lịch đến tại một điểm đến nhằm nâng cao hiệu quả dự báo. Google Trends được sử dụng để thu thập dữ liệu công cụ tìm kiếm, kết hợp với dữ liệu lịch sử khách du lịch quốc tế đến ở hai điểm đến Việt Nam và Vương quốc Anh. Kết quả dự báo sau đó so sánh với dữ liệu du lịch từ Cục Du lịch Quốc gia Việt Nam [8] và Văn phòng thống kê Quốc gia của Vương quốc Anh [9].

Các đóng góp của bài báo gồm: Phân tích cách phát hiện từ khóa dựa trên công cụ Google Suggest để thu thập được dữ liệu tốt từ việc sử dụng đúng từ khóa; Xây dựng giải thuật dự báo dựa trên dữ liệu thu thập từ Google Trends với từ khóa xác định được ở bước trên; Đề xuất một cải tiến giải thuật dự báo dựa trên dữ liệu Google Trends với việc kết hợp thêm dữ liệu cùng kỳ; Triển khai các giải thuật với dữ liệu khách du lịch đến Việt nam để đánh giá hiệu quả của phương pháp dự báo dựa trên kết hợp của dữ liệu công cụ tìm kiếm và dữ liệu lịch sử khách du lịch đến Việt Nam.

Nội dung tiếp theo của bài báo gồm: phần 2 phân tích các nghiên cứu liên quan đến Google Trends và dự báo du lịch. Phần 3 mô tả việc thu thập dữ liệu, xử lý dữ liệu và đề xuất các thuật toán dự báo. Phần 4 so sánh các kết quả dự báo thực nghiệm khác nhau của các thuật toán được đề xuất. Phần 5 kết luận bài báo này.

2 Các nghiên cứu liên quan

2.1 Google Trends

Google là công cụ tìm kiếm internet lớn nhất thế giới, chiếm hơn 90% hoạt động tìm kiếm. bắt đầu công bố sẵn sàng từ tháng 01 năm 2004, dữ liệu Google Trends là tổng hợp các truy vấn tìm kiếm riêng lẻ trên Google theo thuật ngữ, thời gian, danh mục và vị trí dựa trên địa chỉ IP (Internet Protocol) mà tìm kiếm được thực hiện. Google sử dụng quy trình lấy mẫu và đưa lỗi đo lường vào chuỗi thời gian. Dữ liệu tải xuống cho cùng một truy vấn tìm kiếm trong những ngày khác nhau là dựa trên các mẫu khác nhau và do đó sẽ dẫn đến các chuỗi thời gian có khác nhau một ít. Để giảm thiểu lỗi đo lường này, theo cách tiếp cận mà Carrieré-Swallow và Labbé (2013) sử dụng [10], dữ liệu Google Trends được thu thập mỗi ngày trong khoảng thời gian 30 ngày và giá trị trung bình của mỗi quan sát được tính toán trong khoảng thời gian này. Các chuỗi hàng tháng này được chuẩn hóa theo tổng số truy vấn tìm kiếm của Google ở một địa phương hay quốc gia nhất định và nằm trong khoảng giá trị $[0, 100]$. Chi tiết về việc dữ liệu Google Trends được mô tả tại www.google.com/trends.

Nghiên cứu trong bài báo này xem xét dữ liệu nhu cầu du lịch tại Việt Nam. Do đó các từ khóa tìm kiếm liên quan đến du lịch Việt Nam sẽ được phân tích ở Mục 3.1. Trong số các từ khóa tìm kiếm liên quan đến du lịch Việt Nam, "*visit to Vietnam*" là thuật ngữ tìm kiếm trực tuyến có liên quan nhất đối với phân tích thực nghiệm về lượng khách du lịch đến Việt Nam.

2.2 Dự báo du lịch với dữ liệu chuỗi thời gian

Dự báo nhu cầu du lịch là một chủ đề nghiên cứu rất phổ biến. Đã có nhiều nghiên cứu khám phá vấn đề nhu cầu du lịch ở các địa phương và quốc gia khác nhau, vì điều đó là rất

quan trọng đối với nền kinh tế của địa phương và quốc gia này. Trước đây, các mô hình chuỗi thời gian được sử dụng để phân tích và dự báo nhu cầu du lịch bằng cách xem xét dữ liệu lịch sử trong quá khứ [11][12][13]. Gần đây, việc kết hợp các phương pháp khác nhau trong dự báo là một hướng tiếp cận đã được nghiên cứu và thử nghiệm nhằm cải thiện độ chính xác của dự báo [14][15][16]. Bài báo này chủ yếu tập trung vào các mô hình dự báo du lịch dựa trên kết hợp các nguồn dữ liệu chuỗi thời gian khác nhau, trong đó đặc biệt tập trung vào dữ liệu truy vấn tìm kiếm.

Thực tế, có mối quan hệ nhân quả giữa việc tìm kiếm thông tin và quyết định đi du lịch của khách du lịch. Việc nắm bắt mối quan hệ giữa khách du lịch, với tư cách là người tìm kiếm thông tin và công cụ tìm kiếm sẽ phản ánh tốt hơn nhu cầu của khách du lịch đến tại một điểm đến. Trong [17], K. Volchek và cộng sự (CS) đã nghiên cứu cách thức cải thiện độ chính xác của dự báo nhu cầu du lịch ở cấp độ vi mô. Số lượng lượt du khách vào 05 bảo tàng London được dự báo và sức mạnh dự đoán của Naive, Naive theo mùa, trung bình động tự hồi quy theo mùa, trung bình động tự hồi quy theo mùa với các biến giải thích, lấy mẫu dữ liệu tần suất hỗn hợp SARMAX và các mô hình mạng nơ-ron nhân tạo được so sánh. Kết quả thực nghiệm giúp mở rộng thêm sự hiểu biết về các loại dữ liệu và thuật toán dự báo khác nhau đến cấp độ các điểm tham quan cụ thể. Việc đưa thêm chỉ số Google Trends vào các mô hình chuỗi thời gian đã giúp nâng cao hiệu quả dự báo về lượng khách đến các điểm tham quan. Tuy nhiên, cũng dựa trên kết quả thực nghiệm, không có mô hình nào vượt trội hơn trong hầu hết các kịch bản. Độ chính xác dự báo của các mô hình khác nhau thay đổi đối với các dự đoán nhu cầu ngắn hạn và dài hạn. Việc áp dụng dữ liệu truy vấn tìm kiếm có tần suất cao hơn cho phép tạo ra các dự đoán hàng tuần, điều này rất cần thiết cho việc lập kế hoạch ở cấp độ điểm tham quan và điểm đến.

Nghiên cứu trong [18] cho rằng việc phân tích dữ liệu công cụ tìm kiếm có giá trị vì thông tin kinh doanh được tạo ra từ phân tích cung cấp thông tin chi tiết về các cơ hội kinh doanh và Google Trends là một mục tiêu phổ biến để nghiên cứu dữ liệu công cụ tìm kiếm, vì nó dễ dàng có sẵn và dễ truy cập. Vì vậy Y. Li và CS đã đề xuất một phương pháp tiếp cận có hệ thống để có được dữ liệu công cụ tìm kiếm Google Trends, để khám phá cách sử dụng dữ liệu và sau đó đưa ra dự báo. Các tác giả đã sử dụng nhu cầu du lịch Đài Loan làm mục tiêu nghiên cứu, trong đó cả ước tính và dự báo đều được thực hiện. Kết quả dự báo sau đó được so sánh với dữ liệu thực từ Cục Du lịch Đài Loan. Nghiên cứu cũng đã chỉ ra rằng, ba vấn đề cần phải giải quyết để cải thiện hiệu quả của việc dự báo dựa trên dữ liệu truy vấn tìm kiếm: Đầu tiên, công cụ tìm kiếm mà hầu hết khách du lịch sử dụng là gì? Thứ hai, ngôn ngữ mà khách du lịch sử dụng khi họ duyệt Internet là gì? Thứ ba, từ khóa nào phù hợp cho Google Trends để thu thập dữ liệu có liên quan cho nhu cầu du lịch là gì?

Trong [19], T. Havranek và CS xem xét tính hữu ích của dữ liệu Google Trends trong việc dự đoán lượng khách du lịch đến hàng tháng và số đêm lưu trú tại Praha trong giai đoạn từ tháng 01 năm 2010 đến tháng 12 năm 2016. Các tác giả đã phân tích liệu Google Trends có cung cấp những cải tiến dự báo đáng kể so với các mô hình không có dữ liệu tìm kiếm hay không và đánh giá liệu một biến tần suất cao (Google Trends hàng tuần) có hữu ích hơn cho việc dự báo chính xác so với một biến tần suất thấp (lượng khách du lịch đến hàng tháng) bằng cách sử dụng lấy mẫu dữ liệu hỗn hợp hay không. Kết quả thực nghiệm cho thấy được tiềm năng của Google Trends trong việc đưa ra những dự đoán chính xác hơn trong bối cảnh du lịch. Cụ thể, thông tin Google Trends, trước 2 tháng và 1 tuần so với lượng khách đến, đều hữu ích để dự đoán số lượng khách du lịch thực tế đến. Mô hình dự báo sử dụng dữ liệu Google Trends hàng tuần hoạt động tốt hơn các mô hình sử dụng dữ liệu Google Trends hàng tháng và các mô hình không có dữ liệu Google Trends.

Tóm lại, dữ liệu Google Trends có một đóng góp đáng kể đối với kết quả dự báo trong các mô hình kết hợp. Tuy nhiên, làm thế nào để xác định được dữ liệu đúng từ Google Trends và cách kết hợp với dữ liệu lịch sử? Phần tiếp theo sẽ mô tả và phân tích chi tiết phương pháp kết hợp dữ liệu lịch sử và Google Trends trong mô hình dự báo nhu cầu du lịch, với trường hợp nghiên cứu là khách du lịch quốc tế Việt Nam và Vương quốc Anh.

3 Dự báo dựa trên kết hợp dữ liệu lịch sử và Google Trends

3.1 Thu thập dữ liệu

Trường hợp nghiên cứu chính đối với mô hình dự báo kết hợp dữ liệu lịch sử và Google Trends là du khách quốc tế đến Việt Nam. Tuy nhiên, để có sự đối sánh, nghiên cứu này cũng xem xét thêm dữ liệu khách quốc tế đến Anh. Do vậy, nguồn thu thập dữ liệu sẽ từ Cục Du lịch Quốc gia Việt Nam [8] và Văn phòng Thống kê Quốc gia của Vương quốc Anh [9].

Dữ liệu thống kê được tổ chức theo tháng. Bài báo xem xét dữ liệu khách quốc tế đến hai điểm đến này từ tháng 01 năm 2008 đến tháng 12 năm 2023. Dựa trên khối lượng dữ liệu thu thập được, dữ liệu từ 01/2008 đến 12/2021 được sử dụng làm dữ liệu mẫu để huấn luyện mô hình và dữ liệu từ tháng 01/2022 đến 12/2023 được dùng để kiểm tra mô hình và dự báo ngoài mẫu.

Đối với dữ liệu Google Trends, một vấn đề là làm thế nào để đảm bảo dữ liệu này phản ánh chính xác sự quan tâm của du khách quốc tế. Nói một cách khác, việc lựa chọn từ khóa phù hợp là rất quan trọng để xác định được dữ liệu này. Quy trình lựa chọn từ khóa đã được thực hiện qua nhiều bước, bao gồm sử dụng các công cụ hỗ trợ (Google Suggest) và phân tích tương quan thống kê (Pearson). Đối với dữ liệu du khách đến Anh, việc phân tích từ khóa cũng thực

hiện tương tự như dữ liệu du khách đến Việt Nam nhưng không có bước thực hiện khảo sát du khách.

a. Lựa chọn từ khóa tiềm năng:

- Google Suggest: Chúng tôi sử dụng công cụ Google Suggest để tìm kiếm các từ khóa tiềm năng liên quan đến du lịch Việt Nam. Google Suggest là công cụ tự động đề xuất các từ khóa phổ biến mà người dùng thường tìm kiếm khi nhập từ khóa vào ô tìm kiếm của Google. Đây là một nguồn dữ liệu phong phú giúp tìm được những từ khóa mà người dùng thực sự quan tâm khi tìm hiểu về một điểm đến du lịch như Việt Nam. Ví dụ: khi nhập từ *"visit to"*, Google Suggest có thể đề xuất các từ khóa như *"visit to Vietnam"*, *"Vietnam travel"*, *"Vietnam attractions"*, *"Vietnam tourism"*, v.v. Những từ khóa này được xem là gợi ý tiềm năng cho việc xác định dữ liệu Google Trends sau này.

- Phỏng vấn du khách quốc tế: Đồng thời với việc dựa trên các từ khóa gợi ý của Google, nghiên cứu này cũng tiến hành phỏng vấn trực tiếp du khách quốc tế đã hoặc có ý định đến Việt Nam. Các câu hỏi phỏng vấn được thiết kế để tìm hiểu xem họ đã sử dụng những từ khóa nào khi tìm kiếm thông tin về Việt Nam trước chuyến đi. Kết quả từ phỏng vấn cung cấp cái nhìn thực tế về hành vi tìm kiếm của du khách quốc tế, giúp lọc ra những từ khóa có mức độ sử dụng cao và liên quan nhất.

b. Lọc và sàng lọc từ khóa

Sau khi thu thập danh sách từ khóa tiềm năng từ Google Suggest và phỏng vấn du khách, nghiên cứu này tiếp tục tham khảo ý kiến của các chuyên gia du lịch để sàng lọc các từ khóa phù hợp với xu hướng du lịch quốc tế và thị trường du lịch Việt Nam. Các chuyên gia này có kinh nghiệm trong việc dự đoán nhu cầu du lịch và hiểu rõ về hành vi của khách quốc tế. Kết quả danh sách các từ khóa liên quan đến du lịch Việt Nam gồm: *"visit to Vietnam"*, *"Vietnam travel"*, *"Vietnam tourist attractions"*, *"best time to visit Vietnam"*, *"Vietnam trip"*.

c. Tính toán hệ số tương quan Pearson

Để đánh giá mức độ liên quan của từng từ khóa với lượng khách quốc tế đến Việt Nam, chúng tôi áp dụng hệ số tương quan Pearson giữa chuỗi dữ liệu Google Trends của từng từ khóa và số lượng khách quốc tế đến Việt Nam, Vương quốc Anh. Hệ số tương quan Pearson là một chỉ số đo lường mức độ liên hệ tuyến tính giữa hai biến. Hệ số này có giá trị từ -1 đến 1, với: 1 là biểu thị mối tương quan hoàn hảo dương (tức là khi một biến tăng thì biến kia cũng tăng), 0 là biểu thị không có mối tương quan, và -1 là biểu thị mối tương quan hoàn hảo âm (tức là khi một biến tăng thì biến kia giảm).

Công thức tính hệ số tương quan Pearson:

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} \quad (1)$$

trong đó, X là chuỗi số lượng khách du lịch quốc tế hàng tháng, Y là giá trị Google Trends cho một từ khóa cụ thể, n là số lượng điểm dữ liệu được sử dụng để tính toán và r là hệ số tương quan Pearson.

d. Chọn từ khóa có tương quan cao nhất

Sau khi tính toán hệ số tương quan Pearson cho từng từ khóa, các hệ số này được so sánh giá trị để tìm ra từ khóa có mức tương quan cao nhất với chuỗi dữ liệu về lượng khách du lịch quốc tế. Từ khóa có hệ số tương quan càng gần 1, tức là càng có mối liên hệ mạnh với sự thay đổi trong lượng khách du lịch, sẽ được ưu tiên lựa chọn. Kết quả từ khóa "*visit to Viet Nam*" có hệ số tương quan cao nhất với chuỗi dữ liệu số lượng khách du lịch quốc tế đến Việt Nam, cho thấy sự thay đổi trong lượng tìm kiếm từ khóa này có mối quan hệ mạnh mẽ với sự thay đổi về số lượng khách. Do đó, từ khóa "*visit to Viet Nam*" đã được chọn làm từ khóa chính để sử dụng trong mô hình dự báo.

Thông qua quy trình lựa chọn từ khóa, bao gồm việc sử dụng Google Suggest, phỏng vấn du khách quốc tế và tính toán hệ số tương quan Pearson, từ khóa "*visit to Viet Nam*" được chọn. Từ khóa này thể hiện mối tương quan mạnh với lượng khách du lịch quốc tế đến Việt Nam, và do đó, nó được sử dụng làm cơ sở dữ liệu Google Trends cho mô hình dự báo. Việc sử dụng dữ liệu từ khóa phù hợp sẽ giúp tăng độ chính xác của mô hình dự báo và phản ánh chính xác hơn xu hướng thực tế của du khách quốc tế.

3.2 Giải thuật dự báo với Google Trends

Để dự báo lượng khách quốc tế đến Việt Nam, chúng tôi sử dụng mô hình dự báo dựa trên kết hợp dữ liệu lịch sử và chỉ số Google Trends. Các bước cụ thể của giải thuật dự báo với Google Trends (Forecasting with Google Trends – FGT algorithm) gồm các bước như sau:

Bước 1 - Tính trung bình số lượng khách du lịch: Số lượng khách du lịch cho mỗi tháng từ dữ liệu mẫu (từ năm 2008 đến năm 2021) được tính trung bình

$$A_{tour,m} = \frac{1}{n} \sum_{y=2008}^{2021} tour_{y,m} \quad (2)$$

trong đó: $A_{tour,m}$ là số lượng khách du lịch trung bình trong tháng m và $tour_{y,m}$ là số lượng khách du lịch thực tế trong tháng m của năm y .

Bước 2 - Tính trung bình Google Trends: Giá trị trung bình Google Trends của từ khóa "visit to Viet Nam" (với dữ liệu Việt Nam) cho mỗi tháng từ dữ liệu mẫu được xác định như sau:

$$A_{google,m} = \frac{1}{n} \sum_{y=2008}^{2021} google_{y,m} \quad (3)$$

trong đó: $google_{y,m}$ là số lượng tìm kiếm thực tế qua Google trong tháng m của năm y .

Bước 3 - Tính tỷ lệ giữa số lượng khách du lịch và Google Trends: Dựa trên số lượng khách du lịch thực tế và giá trị Google Trends trung bình của từng tháng, tỷ lệ giữa số lượng khách du lịch và Google Trends được tính như sau:

$$R_m = \frac{A_{tour,m}}{A_{google,m-1}} \quad (4)$$

trong đó, R_m là tỷ lệ giữa số lượng khách du lịch và giá trị Google Trends của tháng m .

Bước 4 - Dự báo cho tháng tiếp theo bằng cách sử dụng tỷ lệ R_m và giá trị $google_{y,m}$ của tháng hiện tại m . Số lượng khách du lịch cho tháng tiếp theo được dự báo theo công thức sau:

$$F_{y,m+1} = google_{y,m} \times R_m \quad (5)$$

3.3 Giải thuật dự báo với Google Trends cải tiến

Có thể thấy rằng, lượng du khách đến hàng tháng có sự liên quan với nhau, luôn thể hiện một xu hướng nào đó về tăng hay giảm. Để nâng cao độ chính xác của dự báo, phương pháp dự báo ở trên được cải tiến nhằm điều chỉnh kết quả dự báo dựa trên xu hướng của các tháng trước đó. Các bước của giải thuật dự báo với Google Trends cải tiến (Improved Forecasting with Google Trends – IFGT algorithm) thực hiện như sau:

Bước 1 - Tính sự thay đổi giữa các tháng: sự khác biệt giữa số lượng khách du lịch trong hai tháng liên tiếp được tính toán để xác định xu hướng tăng hoặc giảm:

$$d_{y,m} = tour_{y,m} - tour_{y,m-1} \quad (6)$$

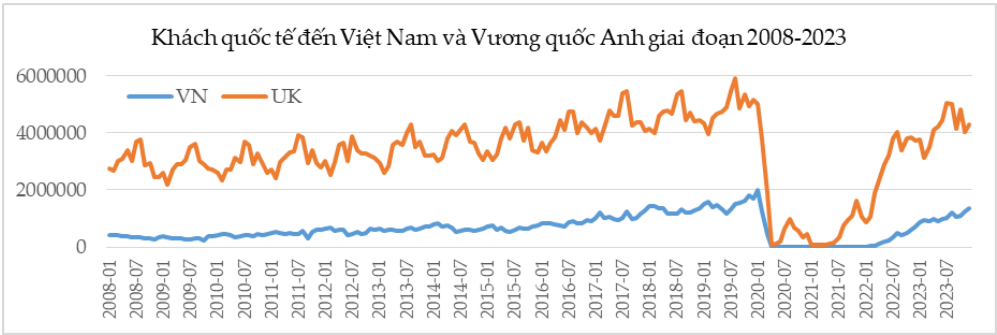
Bước 2 - Xác định xu hướng: Nếu sự thay đổi dy,m dương trong ba tháng liên tiếp, có thể nhận định rằng xu hướng đang tăng; ngược lại, nếu dy,m âm trong ba tháng liên tiếp, có thể xác định xu hướng đang giảm.

Bước 3 - Điều chỉnh dự báo: Nếu có xu hướng tăng, dự báo sẽ được điều chỉnh tăng thêm một lượng bằng độ lệch chuẩn của số lượng du khách trong các tháng trước đó. Ngược lại, nếu có xu hướng giảm, dự báo sẽ được điều chỉnh giảm theo tương tự.

4 Cài đặt và phân tích kết quả

4.1 Chuẩn hóa dữ liệu

Trường hợp nghiên cứu cho mô hình dự báo nêu trên là du khách quốc tế đến tại điểm đến Việt Nam và Anh trong giai đoạn 2008 đến 2023. Dữ liệu lịch sử được thu thập từ Cục Du lịch Quốc gia Việt Nam [8] và Văn phòng Thống kê Quốc gia của Vương quốc Anh [9]. Hình 1 mô tả xu hướng tăng giảm khách du lịch đến Việt Nam và Vương quốc Anh trong giai đoạn 2008-2023.



Hình 1. Biểu đồ du khách quốc tế đến Việt Nam và Vương quốc Anh giai đoạn 2008-2023

Chỉ số Google Trends đối với từ khóa “visit to Vietnam” và “visit to UK” cũng được thu thập trong giai đoạn 2008-2023 này.

Bảng 1. Mô tả dữ liệu thu thập

Việt Nam			Vương quốc Anh	
TimeIndex	Du khách đến (tour)	Chỉ số Google Trends (google)	Du khách đến (tour)	Chỉ số Google Trends (google)
2008-01	399.556	20	2.349.000	50
2008-02	411.032	15	2.249.000	46
2008-03	414.332	13	2.597.000	50
...	...	...	...	...
2023-10	1.112.526	73	3.715.000	50
2023-11	1.233.371	75	2.795.000	44
2023-12	1.371.135	85	2.931.000	46

4.2 Thực hiện dự báo và phân tích kết quả

Để tường minh, quá trình thực hiện dự báo được trình bày như sau:

Dự báo với giải thuật FGT

Phần này trình bày các kết quả từ quá trình phân tích thực nghiệm, bao gồm các bước để dự báo số lượng khách du lịch theo mô hình và sau đó áp dụng phương pháp cải tiến để cải thiện độ chính xác của dự báo. Các bảng dữ liệu về lượng khách du lịch, điểm Google Trends và tỷ lệ giữa hai biến này được sử dụng để phân tích.

Bảng 2. So sánh lượng khách du lịch và xu hướng tìm kiếm trung bình giữa Việt Nam và Vương quốc Anh theo tháng trong năm

Tháng	Việt Nam (VN)			Vương quốc Anh (UK)		
	A _{tour}	A _{google}	R	A _{tour}	A _{google}	R
01	799.963	38	27.585	2.388.000	61	49.751
02	777.558	29	20.462	2.054.000	61	33.664
03	655.984	29	22.620	2.381.000	55	39.025
04	639.016	28	22.035	2.566.000	59	46.662
05	581.746	28	20.777	2.789.000	59	47.274
06	538.371	27	19.228	2.704.000	59	45.834
07	582.284	28	21.566	3.205.000	72	54.328
08	649.621	30	23.201	3.246.000	78	45.084
09	580.437	29	19.348	2.644.000	56	33.899
10	606.147	28	20.902	2.808.000	61	50.143
11	689.804	30	24.636	2.453.000	51	40.211
12	699.549	29	23.318	2.333.000	48	45.735

trong đó, A_{tour} là lượng khách du lịch trung bình đến trong các tháng từ tháng 01 đến tháng 12; . A_{google} là giá trị trung bình của xu hướng tìm kiếm từ khóa trên Google từ tháng 01 đến tháng 12; R là tỷ lệ giữa lượng khách du lịch trung bình và xu hướng tìm kiếm trung bình trên Google.

Trình tự các bước được thực hiện như sau:

Bước 1: Trung bình lượng khách du lịch theo tháng. Trong cột A_{tour} của Bảng 2 cho thấy lượng khách du lịch trung bình đến trong các tháng từ tháng 01 đến tháng 12. Với dữ liệu Việt Nam, lượng khách du lịch cao nhất được ghi nhận vào tháng 01 với mức trung bình 799.962 người và thấp nhất vào tháng 06 với 538.370 người. Thông qua dữ liệu, có thể nhận thấy sự biến động rõ rệt về lượng khách du lịch theo mùa, trong đó các tháng mùa hè (tháng 06, 07, 08) có mức tăng đáng kể về số lượng khách, đặc biệt là vào tháng 07 và tháng 08.

Bước 2: Xu hướng tìm kiếm trên Google theo tháng. Trong cột A_{google} của Bảng 2 cho thấy xu hướng tìm kiếm về du lịch trên Google từ tháng 01 đến tháng 12. Số liệu Việt Nam cho thấy mức độ quan tâm của người dùng giảm dần từ tháng 01 (38) xuống mức thấp nhất vào tháng 05 (28), sau đó có xu hướng tăng trở lại trong những tháng cuối năm. Điều này cho thấy người dùng có xu hướng tìm kiếm thông tin du lịch vào đầu năm, sau đó giảm dần trước khi tăng trở lại vào mùa du lịch cao điểm.

Bước 3: Tỷ lệ giữa lượng khách du lịch và xu hướng tìm kiếm. Trong cột R của Bảng 2 cung cấp tỷ lệ giữa lượng khách du lịch và xu hướng tìm kiếm trên Google. Tháng 01 trong dữ liệu Việt Nam có tỷ lệ cao nhất với 27.653, cho thấy mối quan hệ chặt chẽ giữa số lượng tìm kiếm và lượng khách du lịch trong tháng này. Trong khi đó, tháng 09 có tỷ lệ thấp nhất với 19.120, có thể phản ánh rằng không phải tất cả các tìm kiếm du lịch đều dẫn đến hành động thực tế.

Bước 4: Bước này sử dụng Google Trends ngoài mẫu để tính toán dự báo cho lượng khách du lịch trong tương lai. Ví dụ: dự báo cho tháng 01/2023 (Việt Nam) dựa trên chỉ số Google Trends của tháng 12/2022 và hệ số tỷ lệ đã tính toán. Công thức dự báo cho lượng khách tháng 01/2023 là:

$$F_{2023,01} = google_{2022,12} - R_{2023,01} \tag{7}$$

trong đó, $F_{2023,01}$ là dự báo cho tháng 01/2023; $google_{2022,12}$ là chỉ số Google Trends của tháng 12/2022; $R_{2023,01}$ là hệ số tương quan giữa chỉ số Google Trends và lượng khách du lịch thực tế đã được tính toán dựa trên dữ liệu trước đó.

Từ Bảng 3 so sánh lượng khách du lịch thực tế và dự báo của Việt Nam cho năm 2022 và 2023 cho thấy kết quả dự báo (Việt Nam) cho tháng 01/2023 là 1.572.118 khách, trong khi số lượng khách thực tế là 871.162.

Bảng 3. Lượng du khách đến thực tế và kết quả dự báo năm 2022 và 2023 tại điểm đến Việt Nam và Vương quốc Anh

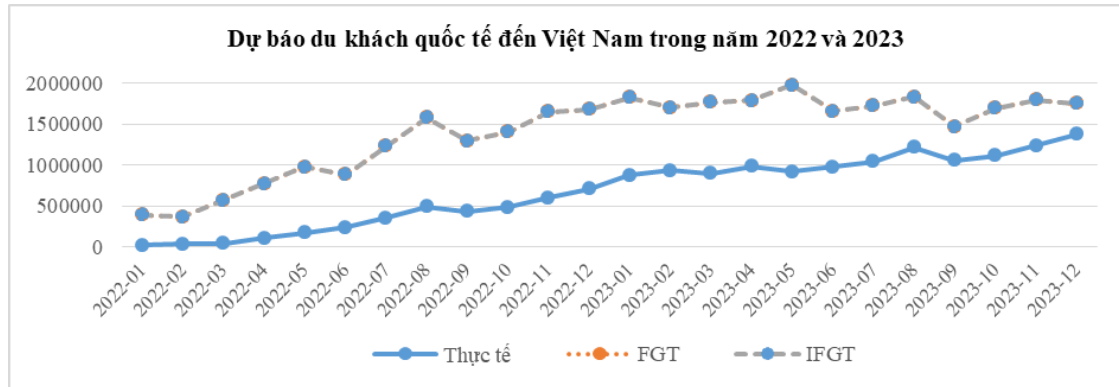
Thời gian	Việt Nam			Vương quốc Anh		
	Thực tế	Kết quả dự báo với FGT	Kết quả dự báo với IFGT	Thực tế	Kết quả dự báo với FGT	Kết quả dự báo với IFGT
2022-01	19.727	386.190	386.190	839.000	1.840.787	1.840.787
2022-02	29.517	368.316	368.316	1.045.000	1.649.536	1.649.536
2022-03	41.740	565.500	565.500	1.860.000	2.302.475	2.302.475
2022-04	101.373	771.225	771.225	2.252.000	2.333.100	2.333.100
2022-05	172.948	976.519	976.519	2.728.000	2.883.714	2.883.714

Thời gian	Việt Nam			Vương quốc Anh		
	Thực tế	Kết quả dự báo với FGT	Kết quả dự báo với IFGT	Thực tế	Kết quả dự báo với FGT	Kết quả dự báo với IFGT
2022-06	236.677	884.488	884.488	2.977.000	2.566.704	2.566.704
2022-07	352.579	1.229.262	1.229.262	3.445.000	3.042.368	3.042.368
2022-08	486.407	1.577.668	1.577.668	3.545.000	2.840.292	2.840.292
2022-09	431.909	1.296.316	1.296.316	2.956.000	2.169.536	2.169.536
2022-10	484.355	1.400.434	1.400.434	3.322.000	2.457.007	2.457.007
2022-11	596.919	1.650.612	1.650.612	3.238.000	2.171.394	2.171.394
2022-12	707.071	1.678.896	1.678.896	3.038.000	2.149.545	2.149.545
2023-01	871.162	1.820.610	1.820.610	2.898.000	2.288.546	2.288.546
2023-02	932.969	1.698.346	1.698.346	2.192.000	1.952.512	1.952.512
2023-03	895.425	1.764.360	1.764.360	2.596.000	2.224.425	2.224.425
2023-04	984.146	1.784.835	1.784.835	3.126.000	2.566.410	2.566.410
2023-05	916.257	1.973.815	1.973.815	3.300.000	2.789.166	2.789.166
2023-06	975.010	1.653.608	1.653.608	3.479.000	2.475.036	3.614.692
2023-07	1.038.563	1.725.280	1.725.280	4.012.000	3.151.024	4.250.076
2023-08	1.217.421	1.832.879	1.832.879	3.810.000	2.795.208	3.959.500
2023-09	1.054.449	1.470.448	1.470.448	3.106.000	2.101.738	2.101.738
2023-10	1.112.526	1.693.062	1.693.062	3.715.000	2.356.721	2.356.721
2023-11	1.233.371	1.798.428	1.798.428	2.795.000	2.010.550	2.010.550
2023-12	1.371.135	1.748.850	1.748.850	2.931.000	2.012.340	2.012.340

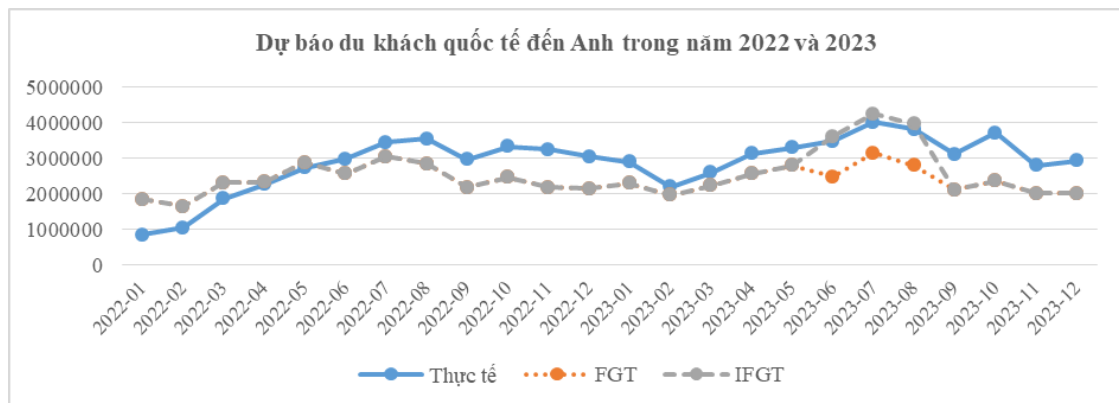
Dựa vào Bảng 3 và Hình 2, chúng ta thấy rằng có một độ lệch nhất định giữa dự báo và thực tế. Đối với Việt Nam, kết quả dự báo cho các tháng trong năm 2022 và 2023 có một sự khác biệt lớn giữa số lượng khách dự báo và thực tế. Điều này có thể là do mô hình chỉ dựa vào dữ liệu từ Google Trends mà không tính đến các yếu tố khác như chính sách du lịch, dịch bệnh (COVID-19) hay các sự kiện bất ngờ khác. Chẳng hạn, với dữ liệu Việt Nam vào tháng 01/2022, lượng du khách đến thực tế chỉ đạt 19.727, trong khi dự báo là 361.969. Sự chênh lệch này cho thấy mô hình dự báo đề xuất có thể gặp khó khăn khi dự báo trong bối cảnh có sự biến động bất thường lớn, như do COVID-19 vào thời điểm đầu năm 2022.

Đối với Vương quốc Anh, kết quả dự báo cho kết quả gần với thực tế hơn so với Việt Nam. Bắt đầu dự báo tháng 01 năm 2022, kết quả dự báo quá cao so với thực tế nhưng dần đến các

tháng sau đó kết quả dự báo càng chính xác hơn. Chẳng hạn, tháng 03 năm 2022, lượng du khách đến thực tế đạt 2.252.000 khách, trong khi dự báo là 2.333.100 khách.



Hình 2. So sánh dự báo và thực tế lượt khách quốc tế đến Việt Nam trong 2022 và 2023



Hình 3. So sánh dự báo và thực tế lượt khách quốc tế đến UK trong 2022 và 2023

Mô hình dự báo cần được điều chỉnh để tích hợp thêm các yếu tố như chính sách thị thực, sự kiện quốc tế, hoặc các yếu tố kinh tế để cải thiện độ chính xác của dự báo. Việc sử dụng Google Trends để dự báo lượng khách du lịch có tiềm năng lớn nhưng cần được kết hợp với các yếu tố khác để đảm bảo độ chính xác. Dữ liệu cho năm 2022 và 2023 cho thấy, trong bối cảnh biến động bất thường như đại dịch COVID-19, mô hình cần điều chỉnh để giảm thiểu sự chênh lệch giữa dự báo và thực tế.

Dự báo với giải thuật IFGT

Phương pháp cải tiến được sử dụng để cải thiện độ chính xác của dự báo bằng cách điều chỉnh dự báo dựa trên xu hướng biến động giữa các tháng liên tiếp. Trong phần này, chúng tôi thực hiện lặp lại các bước từ 1 đến 4 trong Giải thuật FGT. Sau đó, thực hiện Bước 5 như sau:

Tính sự khác biệt giữa hai lượng khách du lịch liên tiếp. Nghiên cứu này đặt ngưỡng thay đổi là 10.000. Để minh họa quá trình, chúng tôi có ví dụ cụ thể với tháng 07 và tháng 06/2022:

$$d_{2022,7} = tour_{2022,7} - tour_{2022,6} = 352.579 - 236.677 = 115.902 \quad (8)$$

Vì sự khác biệt lớn hơn ngưỡng 10.000, chúng ta coi đây là sự thay đổi “dương” (+). Tiếp tục với tháng 08 và tháng 07/2022:

$$d_{2022,8} = tour_{2022,8} - tour_{2022,7} = 486.407 - 352.579 = 133.828 \quad (9)$$

Sự khác biệt cũng lớn hơn ngưỡng 10.000, vì vậy đây cũng là sự thay đổi “dương” (+). Đối với tháng 09 và tháng 08 năm 2022:

$$d_{2022,9} = tour_{2022,9} - tour_{2022,8} = 431.909 - 486.407 = -54.498 \quad (10)$$

Vì sự khác biệt nhỏ hơn ngưỡng -10.000, đây là sự thay đổi “âm” (-). Cuối cùng, xét sự khác biệt giữa tháng 10 và tháng 09/2022:

$$d_{2022,10} = tour_{2022,10} - tour_{2022,9} = 484.355 - 431.909 = 52.446 \quad (11)$$

Sự khác biệt lớn hơn ngưỡng 10.000, vì vậy đây là sự thay đổi “dương” (+). Khi sự khác biệt vượt qua ngưỡng này, chúng ta có thể xác định hướng thay đổi (tăng hoặc giảm) để điều chỉnh dự báo phù hợp. Trong trường hợp không có sự thay đổi lớn, dự báo sẽ không được điều chỉnh nhiều.

Giải thuật IFGT được sử dụng để cải thiện độ chính xác của dự báo bằng cách điều chỉnh dự báo dựa trên xu hướng biến động giữa các tháng liên tiếp. Cụ thể, chúng tôi thực hiện các bước sau:

1. Xác định sự thay đổi giữa các tháng liên tiếp: Chúng tôi tính toán sự thay đổi trong lượng khách du lịch giữa các tháng liên tiếp trong năm 2022 và 2023.

2. Xác định xu hướng: Nếu có ba tháng liên tiếp tăng hoặc giảm, chúng tôi xác định xu hướng là tăng hoặc giảm, từ đó điều chỉnh dự báo. Ví dụ: với dữ liệu thực tế UK trong năm 2023, tháng 04 (3.126.000 lượt khách) tăng so với tháng 03 (2.596.000) và tháng 05 (3.300.000 lượt khách) cũng tăng so với tháng 04. Do đó, xu hướng cho tháng 06 được xác định là tăng.

Dựa trên xu hướng này, chúng tôi điều chỉnh dự báo cho tháng 06. Dự báo ban đầu là 2.475.036 lượt khách, và với xu hướng “Tăng”, chúng tôi tăng dự báo thêm một độ lệch chuẩn của tháng 06 ($\sigma = 1.139.656$). Dự báo mới cho tháng 06 là:

$Dự\ báo\ mới = Dự\ báo\ ban\ đầu + \sigma = 2.475.036 + 1.139.656 = 3.614.692$

Kết quả cho thấy phương pháp cải tiến giúp điều chỉnh dự báo tốt hơn so với phương pháp ban đầu đối với dự báo với dữ liệu Vương quốc Anh, giảm thiểu sự chênh lệch giữa dự báo và thực tế, đặc biệt là trong các tháng có sự biến động lớn. Kết quả dự báo được đánh giá bằng các chỉ số Root Mean Squared Error (RMSE) và Mean Absolute Error (MAE). Chúng tôi so sánh kết quả của giải thuật FGT và cải tiến IFGT của nó. Kết quả trong Bảng 4, cho thấy IFGT có độ chính xác dự báo cao hơn FGT trong trường hợp Vương quốc Anh nhưng trong trường hợp Việt Nam IFGT không thay đổi so với FGT.

Bảng 4. So sánh chỉ số RMSE, MAE giữa FGT và IFGT

Điểm đến	FGT		IFGT	
	RMSE	MAE	RMSE	MAE
Việt Nam	761.653	728.595	761.653	728.595
Vương quốc Anh	762.376	693.587	762.376	595.401

5 **Kết luận**

Nghiên cứu này đã áp dụng dữ liệu Google Trends để dự báo lượng khách du lịch quốc tế đến Việt Nam. Thông qua việc phân tích và lựa chọn từ khóa phù hợp, chúng tôi đã xây dựng một mô hình dự báo hiệu quả, kết hợp giữa dữ liệu lịch sử và xu hướng tìm kiếm. Phương pháp cải tiến đã cải thiện đáng kể độ chính xác của dự báo, đặc biệt trong các tình huống có sự biến động lớn. Kết quả nghiên cứu cho thấy tiềm năng của Google Trends trong việc dự báo xu hướng du lịch, và có thể được mở rộng áp dụng cho các quốc gia khác.

Lời cảm ơn

Bài báo được hoàn thành dưới sự hỗ trợ từ đề tài cấp Đại học Huế: DHH2024-10-38 của Trường Du lịch - Đại học Huế.

Tài liệu tham khảo

1. J. Winter and R. Ono, "Introduction to the Future Internet: Alternative Visions," *Public Adm. Inf. Technol.*, vol. 17, pp. 1–16, 2015, doi: 10.1007/978-3-319-22994-2_1.

2. L. Tang, J. Li, H. Du, L. Li, J. Wu, and S. Wang, "Big Data in Forecasting Research: A Literature Review," *Big Data Res.*, vol. 27, p. 100289, 2022, doi: 10.1016/j.bdr.2021.100289.

3. M. Fathi, M. Haghi Kashani, S. M. Jameii, and E. Mahdipour, "Big Data Analytics in Weather Forecasting: A Systematic Review," *Arch. Comput. Methods Eng.*, vol. 29, no. 2, pp. 1247–1275, 2022, doi: 10.1007/s11831-021-09616-4.
4. H. Li, M. Hu, and G. Li, "Forecasting tourism demand with multisource big data," *Ann. Tour. Res.*, vol. 83, no. January, p. 102912, 2020, doi: 10.1016/j.annals.2020.102912.
5. S. P. Jun, H. S. Yoo, and S. Choi, "Ten years of research change using Google Trends: From the perspective of big data utilizations and applications," *Technol. Forecast. Soc. Change*, vol. 130, no. February 2017, pp. 69–87, 2018, doi: 10.1016/j.techfore.2017.11.009.
6. Y. Li, Z. Lin, and S. Xiao, "Using social media big data for tourist demand forecasting: A new machine learning analytical approach," *J. Digit. Econ.*, vol. 1, no. 1, pp. 32–43, 2022, doi: 10.1016/j.jdec.2022.08.006.
7. M. Á. R. Reina, "Big Data: Forecasting and Control for Tourism Demand," no. January 2018, pp. 273–286, 2020, doi: 10.1007/978-3-030-56219-9_18.
8. "Số liệu thống kê Khách quốc tế." <https://vietnamtourism.gov.vn/statistic/international> (accessed Aug. 15, 2024).
9. "Overseas travel and tourism time series - Office for National Statistics." <https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/datasets/internationalpassengersurveytimeseriesspreadsheet> (accessed Aug. 21, 2024).
10. Y. Carrière-Swallow and F. Labbé, "Nowcasting with Google trends in an emerging market," *J. Forecast.*, vol. 32, no. 4, pp. 289–298, 2013, doi: 10.1002/for.1252.
11. C. Lim and M. McAleer, "Forecasting tourist arrivals," *Ann. Tour. Res.*, vol. 28, no. 4, pp. 965–977, 2001, doi: 10.1016/S0160-7383(01)00006-8.
12. V. Cho, "A comparison of three different approaches to tourist arrival forecasting," *Tour. Manag.*, vol. 24, no. 3, pp. 323–330, 2003, doi: 10.1016/S0261-5177(02)00068-7.
13. H. Hassani, E. S. Silva, N. Antonakakis, G. Filis, and R. Gupta, "Forecasting accuracy evaluation of tourist arrivals," *Ann. Tour. Res.*, vol. 63, pp. 112–127, 2017, doi: 10.1016/j.annals.2017.01.008.
14. M. E. Nor, A. I. M. Nurul, and M. S. Rusiman, "A Hybrid Approach on Tourism Demand Forecasting," *J. Phys. Conf. Ser.*, vol. 995, no. 1, pp. 1–11, Apr. 2018, doi: 10.1088/1742-6596/995/1/012034.
15. M. L. Shen, H. H. Liu, Y. H. Lien, C. F. Lee, and C. H. Yang, "Hybrid approach for forecasting tourist arrivals," *ACM Int. Conf. Proceeding Ser.*, vol. Part F1479, pp. 392–396, 2019, doi: 10.1145/3316615.3316628.
16. Purwanto, Sunardi, F. T. Julfia, and A. Paramananda, "Hybrid model of ARIMA-linear trend model for tourist arrivals prediction model in Surakarta City, Indonesia," *AIP Conf. Proc.*, vol. 2114, no. June, 2019, doi: 10.1063/1.5112481.
17. K. Volchek, A. Liu, H. Song, and D. Buhalis, "Forecasting tourist arrivals at attractions: Search engine empowered methodologies," *Tour. Econ.*, vol. 25, no. 3, pp. 425–447, 2019, doi: 10.1177/1354816618811558.
18. K. H. Huarng and T. H. K. Yu, "Application of Google trends to forecast tourism demand," *J. Internet Technol.*, vol. 20, no. 4, pp. 1273–1280, 2019, doi: 10.3966/160792642019072004025.
19. T. Havranek and A. Zeynalov, "Forecasting tourist arrivals: Google Trends meets mixed-frequency data," *Tour. Econ.*, vol. 27, no. 1, pp. 129–148, 2021, doi: 10.1177/1354816619879584.