BỘ THÔNG TIN VÀ TRUYỀN THÔNG

HỌC VIỆN CÔNG NGHỆ VÀ BƯU CHÍNH VIỄN THÔNG

LÊ THỊ NGỌC ANH

NGHIÊN CỨU MỘT SỐ MÔ HÌNH DỰ BÁO DỊCH TẢ DỰA TRÊN KHAI PHÁ DỮ LIỆU VÀ PHÂN TÍCH KHÔNG GIAN ỨNG DỤNG CÔNG NGHỆ GIS

Chuyên ngành : Hệ thống thông tin

Mã số : 9.48.01.04

TÓM TẮT LUẬN ÁN TIẾN SĨ

Hà Nội, 2018

Công trình được hoàn thành tại:
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học:

PGS.TS. Nguyễn Hoàng Phương

TS. Hoàng Xuân Dậu

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án được bảo vệ trước Hội đồng cấp học viên tại Học viên Công nghệ Bưu chính Viễn thông, 122 Hoàng Quốc Việt, Hà nội.

Vào lúc:

Có thể tìm hiểu luận án tại:

Thư viện Học viện Công nghệ Bưu chính Viễn thông.

MỞ ĐẦU

Dự báo là một hoạt động thường xuyên có tính tất yếu của các cá nhân và tổ chức nhằm đưa ra những thông tin chưa biết trên cơ sở các thông tin đã biết. Trong lĩnh vực y tế và chăm sóc sức khỏe, có một lớp lớn các bài toán dự báo với phạm vi ở nhiều cấp độ cần được giải quyết. Cùng với sự phát triển nhanh chóng của khoa học công nghệ, nhiều phương pháp và kỹ thuật mới đã được sử dụng cho dự báo. Trong đó, mô hình dự báo dựa trên các kỹ thuật khai phá dữ liệu, học máy là một nhóm trong các kỹ thuật đang có xu hướng được áp dụng rộng rãi. Trong những năm gần đây, sự sẵn có và ngày càng tăng các nguồn dữ liệu, đặc biệt là dữ liệu khí hậu - thời tiết thu thập từ các cảm biến từ xa và những dữ liệu phân tích lại, cũng như sự phát triển của các kỹ thuật dự báo đã mang lại cơ hội mới cho phân tích và dự báo dịch bệnh trong ngành y tế. Bên cạnh đó, việc lan truyền của dịch bệnh có liên hệ mật thiết với sự lân cận về không gian và thời gian. Do vậy, việc nghiên cứu các kỹ thuật xây dựng mô hình dự báo dịch bệnh có xem xét đến ảnh hưởng của các yếu tố không gian, thời gian và khí hậu tới sự xuất hiện và lan truyền dịch bệnh là rất cần thiết bằng các kỹ thuật học phân tích và khai phá dữ liệu là rất cần thiết.

MỤC TIÊU VÀ PHẠM VI NGHIÊN CỨU

Mục tiêu của luận án là nghiên cứu hệ thống hóa cơ sở khoa học trong dự báo, ứng dụng các kỹ thuật khai phá dữ liệu, học máy trong dự báo làm cơ sở xây dựng mô hình dự báo dịch bệnh tả có sự kết hợp dữ liệu không gian, thời gian và khí hậu.Phạm vi không gian nghiên cứu áp dụng mô hình là toàn bộ thành phố Hà Nội cùng với các giả thiết gồm:

- Bệnh dịch xảy ra trong một khoảng thời gian đủ ngắn để đảm bảo lượng dân số luôn ổn định.

- Chu kỳ ủ bệnh và các yếu tố xã hội, hành vi, thói quen trong khu vực nghiên cứu được coi là không đáng kể.
- Người nhiễm bệnh đã hết bệnh thì không còn khả năng tái nhiễm bệnh trong cùng một khoảng thời gian dự báo.

Ý NGHĨA VÀ ĐÓNG GÓP

Luận án đã nghiên cứu hệ thống hóa các phương pháp dự báo dịch bệnh, đánh giá mức độ phù hợp của từng nhóm phương pháp và đề xuất lựa chọn giải pháp thích hợp trong dự báo dịch tả với đặc thù của Hà nội. Các mô hình dự báo đề xuất trong luận án là nền tảng cung cấp thông tin y tế như một dịch vụ công để cộng đồng có những phản ứng tốt và tích cực hơn.

Những đóng góp chính của luận án:

- Đề xuất mô hình dự báo dịch tả dựa trên khai phá luật kết hợp và học máy hồi qui, phân lớp.
- Đề xuất mô hình dự báo dịch tả ngắn hạn có đánh giá mức độ ảnh hưởng của các yếu tố khí hậu và địa lý đến sự bùng phát dịch tả .
- Đề xuất mô hình dự báo dịch tả tổng quát dựa trên phân tích không gian ứng dụng công nghệ GIS.

BỐ CỤC CỦA LUẬN ÁN

Ngoài phần Mở đầu và Kết luận, luận án được bố cục gồm 4 chương. **Chương 1:** Tổng quan về các mô hình dự báo dịch bệnh. **Chương 2:** Đề xuất mô hình dự báo dịch tả dựa trên khai phá luật kết hợp và học máy hồi qui, phân lớp. **Chương 3**: Đề xuất mô hình dự báo ngắn hạn – đánh giá độ ảnh hưởng của các yếu tố khí hậu và địa lý tới dịch tả tại Hà Nội. **Chương 4:** Đề xuất mô hình dự báo dịch tả trên địa bàn Tp. Hà Nội có xem xét đến ảnh hưởng của biến đổi khí hậu trên cơ sở ứng dụng các kỹ thuật phân tích không gian dựa trên công nghệ GIS.

CHƯƠNG 1:TỔNG QUAN VỀ CÁC MÔ HÌNH DỰ BÁO DỊCH BỆNH

1.1 Khái niệm và thuật ngữ:

Dự báo là một khoa học và nghệ thuật tiên đoán những sự việc sẽ xảy ra trong tương lai, trên cơ sở phân tích khoa học về các dữ liệu đã thu thập được.

Mô hình là một biểu diễn các thành phần quan trọng của một hệ thống có sẵn (hoặc sắp được xây dựng) với mục đích biểu diễn tri thức của hệ thống đó dưới một dạng có thể sử dụng được.

Trong các tình huống chưa chắc chắn, dự báo (tiếng Anh "*predict*", "*forecast*", "*foresight*") được dùng để chỉ kiểu hoạt động của các cá nhân, các tổ chức và các quốc gia hướng tới mục tiêu nhận biết được giá trị chưa biết của các đại lượng nhằm hỗ trợ ra quyết định. Trong tiếng Việt, hai thuật ngữ "*dự báo*" và "*dự đoán*" được sử dụng trong hầu hết các trường hợp của dự báo. Tuy nhiên, trong một số trường hợp, hai thuật ngữ này được sử dụng theo hai nghĩa phân biệt, chẳng hạn, "*dự báo*" là dự báo về một giá trị chưa biết trong tương lai còn "*dự đoán*" là dự đoán về một giá trị chưa biết trong hiện tại (giá trị đó chắc chắn đã có), hoặc "*dự báo*" là dự báo xu hướng còn "*dự đoán*" là dự đoán giá trị.

1.2 Tổng quan về dự báo dịch bệnh và các mô hình dự báo hiện có.

Mô hình dịch tễ học toán học: mô hình dịch bệnh dựa trên việc chia quần thể đang nghiên cứu thành một số lượng nhỏ các ngăn tương ứng với số lượng trạng thái liên quan tới bệnh dịch mà các cá nhân trong quần thể có thể rơi vào; Ba trạng thái điển hình nhất trong mô hình dịch tễ học toán học gồm:

- Dễ bị nhiễm (**S**:*Susceptible***)**: cá nhân không có khả năng miễn dịch với các tác nhân gây bệnh, và như vậy có thể bị lây nhiễm khi tiếp xúc với các cá nhân đang nhiễm bệnh,

- Nhiễm bệnh (**I**:*Infectious*): cá nhân hiện đang bị nhiễm bệnh và có thể truyền bệnh cho các cá nhân tiếp xúc với họ,
- Đã hồi phục (**R**:*Recovered*): Các cá nhân miễn dịch với dịch bệnh, và do đó không ảnh hưởng đến động lực học truyền bệnh theo bất kỳ cách nào khi họ tiếp xúc với các cá nhân khác.

Mô hình dịch tễ học toán học xem xét các phương trình biến đổi các giá trị *S(t), I(t), R(t)* theo thời gian *t*. Dựa trên các giá trị đã biết, các tham số trong các phương trình này được xác định. Mô hình kết quả được sử dụng để dự báo các giá trị *S(t), I(t), R(t)* tại một thời điểm *t* trong tương lai.

Mô hình khai phá dữ liệu: Mô hình tiếp cận theo hướng sử dụng luật kết hợp (association rule), học máy hồi qui, phân lớp để dự báo. Những mô hình này cơ bản dựa trên lý thuyết các quá trình ngẫu nhiên nhằm lượng hóa tốc độ lan truyền giữa các cá thể thuộc các tầng lớp xã hội đa dạng, có cư trú địa lý khác nhau trong một dân số ổn định. Bên cạnh đó việc khai phá ngữ nghĩa trong các mạng xã hội cũng được áp dụng để giải quyết bài toán dự báo sớm từ thông tin ở các mạng xã hội thông qua các kỹ thuật hồi qui và phân lớp như cây quyết đinh, Bayes, máy vector hỗ trợ,...

Mô hình không gian: Trong y tế, hệ thống thông tin địa lý – Geographic Information System (GIS) cung cấp các công cụ phân tích thống kê, mô hình hóa không gian, hỗ trợ cho việc nghiên cứu các mối quan hệ giữa các yếu tố điều kiện tự nhiên, môi trường và tình hình sức khỏe, bệnh tật của người dân, theo dõi và dự báo diễn biến dịch bệnh, từ đó hỗ trợ ra quyết định phù hợp ở từng thời điểm và ở các cấp quản lý khác nhau. Các kỹ thuật phân tích không gian điển hình bao gồm nội suy không gian, phân tích điểm nóng, hồi qui không gian ước lượng bình phương nhỏ nhất và hồi qui trọng số không gian...Các kỹ

thuật này đã đóng góp hoặc bổ trợ cùng các kỹ thuật dự báo khác để đưa công việc dự báo dịch bệnh ngày càng hiện đại và hiệu quả hơn. Các mô hình dự báo dịch bệnh đề cập trên đều có những ưu điểm và nhược điểm riêng nhìn theo góc độ của kết quả nghiên cứu đạt được. Các mô hình dự báo dịch được công bố có thể được chia thành ba lớp chính như Bảng 1.1.

Bảng 1.1 Đánh giá ưu nhược điểm của các lớp mô hình dự báo dịch bệnh

Nhóm mô hình	Ưu điểm	Nhược điểm
Mô hình dịch tễ học toán học và các biến thể	Lược bỏ được các thành phần phức tạp, chỉ tập trung vào bản chất của mô hình	- Khó khăn xác định được các tham số chủ yếu - Cần nhiều dữ liệu quan sát - Khó khăn trong triển khai đối với mô hình động khi giữa các lớp của mô hình có tương tác với nhau.
Các mô hình dựa trên học máy, khai phá dữ liệu	- Giải quyết được các bài toán dự báo với dữ liệu lớn. - Thu thập dữ liệu nhanh - Phong phú về kỹ thuật/ thuật toán và công cụ - Hỗ trợ mô phỏng	- Kết hợp nhiều kiến thức các chuyên ngành khác, đòi hỏi nhiều nỗ lực và nhân lực các chuyên ngành phối hợp. - Phụ thuộc vào dữ liệu
Các mô hình khác (bao gồm mô hình dựa trên tác tử)	Mã hóa dễ dàng bởi các ký hiệu biểu diễn tri thức	- Khó khăn để chuyển thế giới thực thành những mô tả hình tượng một cách chính xác và đầy đủ. - Đòi hỏi nhiều thời gian để có được kết quả

1.3 Dịch tả và nhu cầu dự báo dịch tả

Theo Tổ chức Y tế Thế giới bệnh tả là bệnh truyền nhiễm nguy hiểm - hầu hết lan truyền qua đường nước - là nguyên nhân đứng hàng thứ năm gây tử vong trên toàn cầu, và đứng hàng thứ hai gây tử vong đối với trẻ em dưới năm tuổi. Dịch tả là một trong những bệnh dịch nhạy cảm với các yếu tố biến đổi thời tiết - khí hậu và được coi như một hình mẫu về tác động của biến đổi khí hậu tới các bệnh dịch. Nhiều công trình nghiên cứu về mối liên quan của biến đổi khí hậu với dịch tả đã được công bố. Các kết quả nghiên cứu cho thấy nguyên nhân bùng phát dịch tả phụ thuộc vào các nhóm yếu tố như: Vị trí địa lý, các biến đổi đa dạng khí hậu, các yếu tố kinh tế-xã hội, nhân khẩu học, vệ sinh môi trường của con người. Mỗi nhóm tác động lan truyền dịch tả trên lại bao gồm rất nhiều yếu tố có thể mà mỗi một khu vực cụ thể tác động của mỗi yếu tố như vậy lại lớn/nhỏ khác nhau. Điều đó có nghĩa là mỗi mô hình dự báo cho một khu vực địa lý cụ thể cần xác định các yếu tố liên quan nhất tới hình thành và lan truyền dịch tả cũng như giá trị cụ thể của các tham số mô hình kết hợp với các yếu tố đó. Ở Việt Nam, trước năm 2005 chỉ có một vài trường hợp bệnh tả đã được báo cáo ở miền Bắc. Tuy nhiên, vào cuối năm 2007, bùng phát dịch tả đã xảy ra tại khu vực này, diễn ra phức tạp. Vì vậy công tác theo dõi, giám sát và dự báo dịch tả để chuẩn bị sẵn sàng các biện pháp ứng phó, phòng chống dịch là vô cùng quan trọng và cần thiết.

1.4 Định hướng nghiên cứu của luận án

Trên cơ sở nghiên cứu lý thuyết và thực tiễn, xây dựng mô hình và lựa chọn kỹ thuật phù hợp để giải quyết từng nội dung của bài toán dự báo, luận án tập trung: (i) Nghiên cứu bài toán dự báo và lựa chọn thuật toán phù hợp để xác định các yếu tố trong mô hình. (ii) Đánh giá tính lân cận không gian địa lý trong mô hình dự báo (đáp ứng đặc thù Việt

Nam). (iii)Tích hợp mô hình với yếu tố lân cận không gian để giải quyết bài toán dự báo dịch bệnh.

1.5 Dữ liệu sử dụng trong nghiên cứu: Để tiến hành nghiên cứu lựa chọn được kỹ thuật phù hợp cho việc thiết lập mô hình dự báo dịch tả, luận án đã tiến hành thu thập dữ liệu nghiên cứu trong giai đoạn 2001-2012 bao gồm các số liệu về số ca dịch tả, về khí hậu và thủy văn khu vực Hà Nội từ các đơn vị như Trung tâm Y học Dự phòng Hà Nội, Trung tâm Nghiên cứu Khí Tượng Thủy Văn Trung Ương, Trung Tâm Nghiên Cứu Môi Trường thuộc Bộ Tài Nguyên Môi Trường. Và dữ liệu về chỉ số dao động phía Nam (Southern Oscillation Index- SOI)- đo sự tiến triển và cường độ của El Nino và La Nina . Tập dữ liệu này được lấy từ nguồn của chính quyền bang Queensland, Úc.

1.6 Kết luận: Chương này giới thiệu tổng quan về một số mô hình dự báo dịch tả trên thế giới. Nội dung chương cũng đã phân tích các ưu điểm và những tồn tại chưa được giải quyết trong các mô hình hiện tại giúp định hướng cho việc nghiên cứu mô hình dự báo với đặc thù Việt Nam. Chương này cũng mô tả các tập dữ liệu phục vụ cho nghiên cứu của luận án.

CHƯƠNG 2 DỰ BÁO DỊCH TẢ DỰA TRÊN KHAI PHÁ LUẬT KẾT HỢP VÀ HỒI QUI, PHÂN LỚP

2.1 Dự báo dịch tả dựa trên khai phá luật kết hợp

Trên cơ sở sử dụng ngôn ngữ R để tạo ra một bảng dữ liệu các ca mắc tả của từng quận, huyện trong thành phố Hà Nội (DL1), tiến hành xây dựng bộ dữ liệu bệnh tả thứ cấp từ tập dữ liệu DL1 dưới dạng danh sách các giao dịch (transaction). Bộ dữ liệu này được lưu trữ ở dạng tệp văn bản gồm nhiều dòng, mỗi dòng là một giao dịch theo ngày. Mỗi giao dịch có các trường dữ liệu: Ngày tháng và danh sách các quận, huyện có ít nhất một ca mắc bệnh tả trong ngày đó. Luận án sử dụng phương pháp dự đoán khả năng xuất hiện bệnh tả bằng việc

sinh các luật kết hợp từ bộ dữ liệu các ca bệnh tả tại các quận huyện ở Hà Nội từ năm 2001 đến năm 2012. Quy trình sinh hay khai phá luật kết hợp bao gồm hai giai đoạn: (1) Tạo ra các tập phổ biến sử dụng thuật toán Apriori [17] và (2) Sinh ra các luật kết hợp sử dụng thuật toán sinh luật.

Mỗi luật có LHS là vế trái của luật, RHS là vế phải của luật; Support, Confidence và Lift tương ứng là các độ đo: độ hỗ trợ, độ tin cậy và độ chắc chắn thống kê. Các tham số thực hiện thuật toán Apriori sinh luật kết hợp được lựa chọn gồm: độ hỗ trợ tối thiểu là 30%, độ tin cậy tối thiểu là 70% và độ dài vế trái (LHS) tối thiểu là 1. Sử dụng bộ dữ liệu DL1, tiến hành khai phá dữ liệu các ca mắc tả theo ngày (từ 1/1/2001 đến 31/12/2012), nghiên cứu đã thu được 50 luật như mô tả trên Bảng 2.1.

Bảng 2.1.Trích một số luật trong số 50 luật kết hợp sinh từ bộ dữ liệu

Rule #	LHS	RHS	Support	Confidence	Lift
R1	{Đống Đa, Hai Bà Trưng, Hoàng Mai}	{ThanhXuan}	0.3027027	0.8615385	2.097166
R2	{Đống Đa, Hoàng Mai}	{Cầu Giấy}	0.3081081	0.7307692	2.048368
R3	{Hai Bà Trưng, Hoàng Mai}	{ThanhXuan}	0.3081081	0.8260870	2.010870
……					
R9	{Từ Liêm}	{ThanhXuan}	0.3027027	0.7272727	1.770335
R10	{ThanhXuan}	{Từ Liêm}	0.3027027	0.7368421	1.770335
……					
R49	{Hà Đông}	{Hoàng Mai}	0.3027027	0.7466667	1.354248
R50	{Hai Bà Trưng}	{Hoàng Mai}	0.3729730	0.7113402	1.290176

Từ kết quả nghiên cứu có thể rút ra một số nhận định:

- Các ca mắc tả có xu hướng cùng xuất hiện tại các quận/huyện có các con sông ô nhiễm của thành phố Hà Nội là Tô Lịch, Kim Ngưu, Nhuệ chảy qua địa bàn với độ chắc chắn cao (trên 70%);
- Các ca mắc tả tại các quận có các sông ô nhiễm chảy qua địa bàn và các ca mắc tả tại các quận tiếp giáp, như Hoàn Kiếm có xu hướng cùng xảy ra với độ chắc chắn cao (trên 70%).

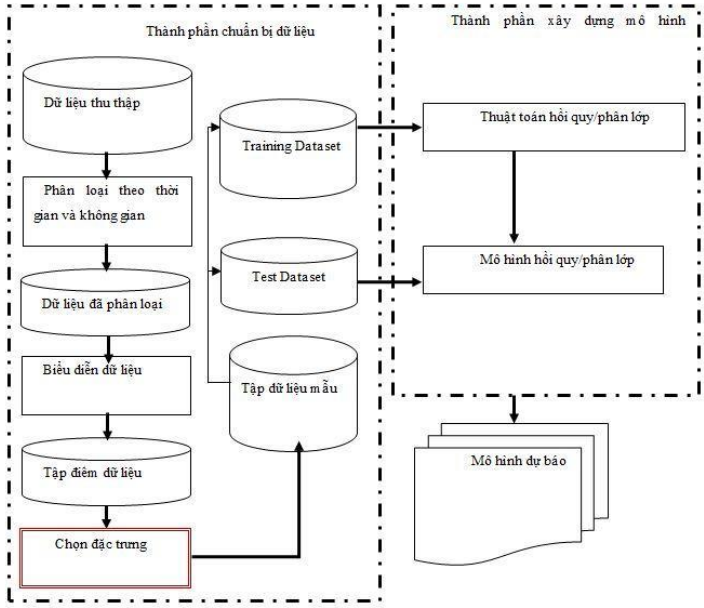
Kết quả giúp khẳng định khai phá luật kết hợp phù hợp với mô hình dự báo dịch tả trong điều kiện không có sự khác biệt nhiều về điều kiện tự nhiên khí hậu giữa các vùng miền. Đây là một bằng chứng khoa học có giá trị thể hiện tính lân cận không gian giữa các quận huyện có ảnh hưởng đến mô hình dự báo.

2.2 Dự báo dịch tả dựa trên học máy hồi qui, phân lớp.

Ý tưởng trong thực nghiệm này là thiết lập mô hình dự báo phân vùng phù hợp với yêu cầu dự báo theo phạm vi quận/ huyện tại Hà nội. Mô hình dự báo sẽ xem xét hai trường hợp biến cục bộ (giá trị từng quận/huyện) và mô hình biến toàn cục (giá trị trong toàn bộ khu vực bao gồm nhiều quận/ huyện). Tại mô hình cục bộ, các yếu tố trong mô hình bao gồm (i) trạng thái dịch tả trong quá khứ và các giá trị khí hậu trong quá khứ ở quận-huyện đang được xem xét và (ii) trạng thái dịch tả trong quá khứ ở các quận – huyện lân cận với quận-huyện đang được xem xét. Giá trị các yếu tố khí hậu tương ứng với một quận-huyện được lấy từ giá trị đo được tại trạm đo gần nhất tới quận - huyện đó. Tại mô hình dự báo toàn cục sẽ xét biến mục tiêu là một vector tình trạng dịch tả cho toàn bộ khu vực (bao gồm các quận – huyện), còn các biến điều kiện bao gồm mọi giá trị quá khứ trạng thái tả và giá trị quá khứ khí hậu trong toàn Hà Nội.

Dữ liệu thực nghiệm được lựa chọn từ tập dữ liệu đã thu thập của luận án tập trung vào giai đoạn các năm 2007-2010. Dữ liệu được chia thành 2 tập: Tập dữ liệu được dùng để học mô hình là tập dữ liệu từ tháng 01/2007 đến tháng 06/2010 và tập dữ liệu kiểm tra mô hình là

tập dữ liệu từ tháng 07/2010 đến tháng 12/2010. Thông qua giải pháp lựa chọn đặc trưng, mối tương quan giữa yếu tố khí hậu với trạng thái dịch tả cũng được xem xét.Nghiên cứu này sử dụng bộ công cụ STATISTICA để khảo sát độ tương quan giữa biến mục tiêu (trạng thái dịch tả trong tương lai) với các biến điều kiện (trạng thái dịch tả, yếu tố khí hậu hiện thời và trong quá khứ) và chỉ các biến điều kiện có tương quan thực sự với biến mục tiêu mới được giữ lại trong biểu diễn dữ liệu cho mô hình dự báo. Bài toán giải quyết trong phần này là xây dựng mô hình dự báo tình trạng dịch tả trong tháng tiếp theo dựa vào dữ liệu về dịch tả và khí hậu của thời điểm hiện tại và các thời điểm trong quá khứ của một tháng trước và hai tháng trước. Phương pháp mô hình hóa được lựa chọn là các phương pháp hồi qui (Linear Regression) và phân lớp (RandomForest, SVM, Bayes). Hình 2.4 thể hiện lưu đồ nghiên cứu xây dựng mô hình dự báo dịch tả tại khu vực Hà Nội.



Hình 2.4. Lưu đồ xây dựng mô hình dự báo dịch tả dựa trên hồi qui, phân lớp

Sử dụng một số độ đo đánh giá mô hình dự báo như sai số tuyệt đổi trung bình (Mean absolute error: MAE), sai số trung bình quân phương (Root mean squared error: RMSE), hệ số tương quan (Correlation coefficient: CC), độ hồi tưởng (Recall), độ chính xác (Precision) và độ đo F (F-Measure).

Tiếp cận cục bộ: Kết quả dự báo thực nghiệm cho 29 quận-huyện có tính phân tán, trong đó độ đo đánh giá mô hình kết quả cho các quận-huyện có rất ít ca dịch tả là khá cao, trong khi, độ đo đánh giá mô hình kết quả cho các quận-huyện nằm trong vùng dịch tả là khá thấp. Lý giải về các quận-huyện có ít ca dịch tả, thậm chí không có ca dịch tả nào trong nhiều thời điểm là độ biến động giá trị biến mục tiêu rất nhỏ sẽ tương ứng với việc lựa chọn các tham số mô hình rất nhỏ (gần giá trị 0) và cho kết quả là sai số nhỏ. Hệ số tương quan (CC) của biến mục tiêu đối với các biến điều kiện đối với hầu hết các quận-huyện rất thấp ngoài trừ tại một số quận-huyện, hệ số tương quan có giá trị được chú ý như Gia Lâm (0.4345), Hoàng Mai (0.5317), Phúc Thọ (0.8624), Tây Hồ (-0.6170), Thạch Thất (0.4328). Đối với các quận-huyện có nhiều ca dịch tả , các độ đánh giá mô hình cho các giá trị thấp. Dựa vào kết quả thực nghiệm cho thấy mô hình hồi quy chưa thực sự thuyết phục khi dự báo tại các quận huyện. Riêng đối với mô hình phân lớp RandomForest cho kết quả các độ đo tương tự như LibSVM và là giải pháp tốt hơn so với những Bayes.*(Xem phụ lục 2)*

Tiếp cận toàn cục: Kết quả thực nghiệm thực hiện theo lựa chọn: (i) các biến điều kiện kết hợp bao gồm cả yếu tố dịch tả và yếu tố khí hậu, (ii) biến điều kiện chỉ là các yếu tố dịch tả và (iii) biến điều kiện chỉ bao gồm yếu tố khí hậu. Tham số độ dài nhịp thời gian quá khứ được chọn là 2 tháng(t-12) và 1 tháng (t-1). Kết quả thực nghiệm là cơ sở để so sánh tác động của biểu diễn cục bộ và biểu diễn toàn cục cũng như lựa chọn được kỹ thuật xây dựng mô hình phù hợp cho từng trường

hợp dự báo. Qua phân tích các kết quả thực nghiệm, so sánh tác động của biểu diễn cục bộ và biểu diễn toàn cục có thể rút ra một số nhận xét:

- Tồn tại sự tương quan giữa các biến điều kiện khí hậu với biến mục tiêu trạng thái dịch tả trong nhiều trường hợp.
- Với biểu diễn dữ liệu chứa các biến điều kiện kết hợp (dịch tả và khí hậu) hoặc chỉ có các biến điều kiện trạng thái dịch tả, thuật toán phân lớp Random Forest cho kết quả tốt hơn hai thuật toán Naïve Bayes và SVM; ngược lại, với biểu diễn dữ liệu chỉ chứa các biến điều kiện khí hậu, thuật toán RandomForest tỏ ra kém hiệu quả hơn.
- Độ đo F1 trong trường hợp tốt nhất của các thuật toán phân lớp đều từ 0.8 trở lên cho thấy có khả năng triển khai một bộ phân lớp kết hợp cho mô hình dự báo dịch tả tại Hà Nội.

Hiệu chỉnh mô hình dự báo với dữ liệu không cân bằng: Để giải quyết vấn đề dữ liệu không cân bằng trong bài toán dự báo dịch tả tại Hà Nội, nghiên cứu sử dụng phương pháp thay đổi phân bố dữ liệu để gia tăng thêm mẫu của lớp tối thiểu. Dữ liệu đầu vào sử dụng cho mô hình dự báo là chuỗi dữ liệu thời gian, gồm các giá trị liên tục của các biến số thời tiết nhiệt độ, độ ẩm, lượng mưa, số giờ nắng... theo ngày của khu vực Hà nội. Chuỗi dữ liệu đầu vào này được biến đổi thành đặc trưng trước khi áp dụng kỹ thuật học máy. Để xác định khoảng thời gian nào có khả năng xảy ra dịch, dữ liệu đầu vào được phân chia thành các đoạn dữ liệu, sử dụng phương pháp cửa sổ trượt với kích cỡ *w* ngày. Các đoạn dữ liệu có thể tách rời hoặc chồng lấn. Thuật toán Random Forest được sử dụng để huấn luyện xây dựng mô hình, sau đó sử dụng kết quả này làm cơ sở so sánh với một số thuật toán phân lớp phổ biến khác nhằm tìm kiếm được thuật toán tối ưu cho bài toán dự báo. Kết quả so sánh độ đo F1 của mô hình dự báo sử dụng các bộ phân lớp khác với nhau được thể hiện ở bảng 2.13.

Bảng 2.13. Bảng so sánh khả năng phân lớp của các bộ phân lớp phổ biến

		Trễ (tuần)						
		0	2	4	6	8	10	12
F1	Random Forest	0.979	0.980	0.978	**0.981**	0.979	0.980	0.976
	NaiveBayes	0.545	0.631	0.641	0.640	0.636	0.655	0.633
	Random Tree	0.943	0.930	0.938	0.962	0.936	0.958	0.951
	SVM	0.773	0.851	0.870	0.859	0.864	0.870	0.853
	J48	0.947	0.957	0.947	0.943	0.947	0.950	0.955
	I-NN	0.979	0.978	0.978	0.976	0.975	0.981	0.974

Kết quả độ đo F1 của mô hình dự báo dựa trên các bộ phân lớp cho trên Bảng 2.13 khẳng định rằng việc sử dụng kỹ thuật phân đoạn dữ liệu là phù hợp và thuật toán RandomForest cho kết quả tốt nhất trong các thuật toán phân lớp sử dụng với độ trễ thời gian là 6 tuần.

2.3 Kết luận: Thực nghiệm khai phá luật kết hợp trong mô hình dự báo với bộ dữ liệu phân bố phi tuyến tính và không có sự khác biệt nhiều về điều kiện tự nhiên đã thu được các luật kết hợp với độ tin cậy và chắc chắn thống kê khá cao, có thể sử dụng như là các yếu tố hỗ trợ ra quyết định trong công tác phòng chống dịch tại thành phố Hà nội.

Với mô hình dự báo dự trên các kỹ thuật học máy hồi qui và phân lớp, các kết quả thực nghiệm cho thấy trong mô hình cục bộ, hồi qui tuyến tính không phù hợp với dữ liệu không phân bố chuẩn (phi tuyến). Đối với biểu diễn toàn cục, các mô hình phân lớp dựa trên LibSVM và Random Forest cho kết quả các độ đo như nhau và phù hợp với mô hình dự báo phi tuyến. Khi áp dụng phương pháp cửa sổ trượt và phân bố dữ liệu theo ngày thì Random Forest cho kết quả ưu việt hơn các kỹ thuật phân lớp phổ biến khác.

CHƯƠNG 3 ẢNH HƯỞNG CỦA YẾU TỐ KHÍ HẬU VÀ ĐỊA LÝ TRONG DỰ BÁO DỊCH TẢ NGẮN HẠN

3.1 Xây dựng mô hình dự báo dịch tả ngắn hạn: Thực nghiệp sử dụng phương pháp tổng hợp số liệu theo ngày cho mô hình dự báo (ngoài trừ dữ liệu địa lý). Điều này giúp tăng số điểm dữ liệu trong giai đoạn nghiên cứu và thuận lợi hơn trong xây dựng mô hình dự báo ngắn hạn. Các tập dữ liệu thời tiết, SOI và số ca bệnh được tổng hợp theo ngày và trộn thành một tập dữ liệu duy nhất, gọi là FS. Tập dữ liệu FS có 35 biến và 4383 quan sát. Trong số 35 biến, có 6 biến thời tiết bao gồm: nhiệt độ không khí, độ ẩm, lượng mưa, số giờ nắng, tốc độ gió và SOI. Các biến còn lại là số ca mắc tả cho 29 quận/huyện của Hà Nội.

3.2 Thực nghiệm và đánh giá mô hình: Tiến hành xây dựng 29 mô hình dự báo cho 29 quận/huyện của thành phố Hà Nội. Giả sử *d* là độ trễ thời gian khởi động của mô hình. Các biến vào và ra của mô hình được mô tả như sau:

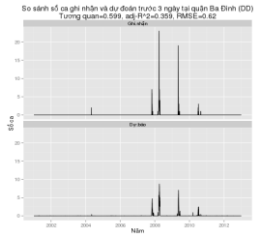
Các biến vào bao gồm:

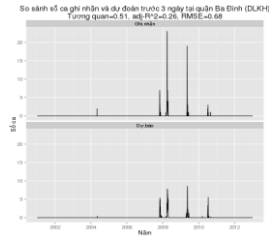
Nhóm biến khí hậu :- Độ ẩm trung bình ngày, nhiệt độ trung bình ngày, lượng mưa ngày, số giờ nắng ngày, tốc độ gió theo ngày, chỉ số dao động phía Nam SOI (theo ngày)

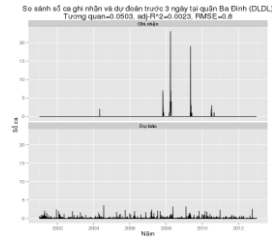
Nhóm biến lân cận: Các biến liên quan số ca mắc tả của các quận/huyện lân cận. Số ca mắc tả của các quận/huyện lân cận trong 0, 1, 2, …, *d* ngày trước đó. Quận/huyện *i* được gọi là lân cận với quận/huyện *j* nếu *i* và *j* có chung đường ranh giới hành chính. Việc xác định toàn bộ các quận/huyện lân cận của một quận/huyện được thực hiện bằng truy vấn không gian trên CSDL không gian được xây dựng từ dữ liệu địa lý của Hà Nội.

Biến ra: Số ca mắc tả trong 0, 1, 2, …, n ngày tiếp theo ở một quận/huyện.

Các tham số có thể thay đổi được của các mô hình là *d* (độ trễ thời gian) và *n* (số ngày dự báo). Với mỗi quận/huyện của Hà Nội, xây dựng 3 mô hình dự báo: (1) mô hình dự báo đầy đủ (DD) bao gồm cả dữ liệu khí hậu và dữ liệu địa lý lân cận, (2) mô hình độc lập khí hậu (DLKH) không sử dụng dữ liệu khí hậu và (3) mô hình độc lập địa lý lân cận (DLDL) không sử dụng dữ liệu địa lý lân cận. Mục đích của việc thiết lập này là để lựa chọn được mô hình dự báo tốt nhất cho Hà Nội và đánh giá được mức độ ảnh hưởng của dữ liệu không gian địa lý lân cận và khí hậu đến độ chính xác của mô hình dự báo. Tất cả các mô hình đều có đầu ra là số ca bệnh tả. Mỗi mô hình có một tham số độ trễ *l* tính theo ngày. Tham số này có nghĩa là sẽ sử dụng số lượng ca bệnh tả tại thời điểm hiện tại và *l-1* ngày trước đó trong quận đang xem xét như là một biến dự báo cho mô hình. Mô hình sẽ dự báo số ca bệnh tả của quận hiện tại trong *l* ngày tiếp theo. Nghiên cứu sử dụng kỹ thuật hồi qui Random Forest (RF) để xử lý tập dữ liệu chuỗi thời gian theo phương pháp cửa sổ trượt. Sử dụng các độ đo thông dụng như sai số trung bình quân phương (Root mean square error – RMSE) và hệ số xác định điều chỉnh (Adjusted detemination coefficient –R^2). Các giá trị RMSE và R^2 được tính toán cho tất cả các mô hình. Để so sánh ảnh hưởng của các yếu tố khí hậu và địa lý đến độ chính xác dự báo, nghiên cứu sử dụng phương pháp đánh giá Tukey với 4 khoảng dự báo 3, 7, 14 và 30 ngày.







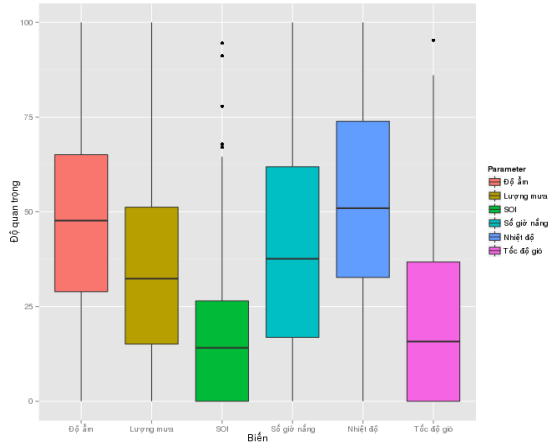
Hình 3.1 Minh họa so sánh giữa ca ghi nhận với mô hình dự báo trước 3 ngày ở quận Ba đình

Xét khoảng cách của độ tin cậy và giá trị trung bình của các cặp mô hình DLDL-DD và DLKH-DD có thể thấy các mô hình đầy đủ (DD) có độ đo R^2 cao nhất cũng là mô hình tốt nhất. Các mô hình độc lập địa lý (DLDL) có độ đo R^2 thấp nhất. Như vậy, có thể kết luận số ca mắc tả ở một quận/huyện có liên kết chặt chẽ với số ca mắc tả ở các quận/huyện lân cận.

3.3 Mối quan hệ giữa độ chính xác và khoảng thời gian dự báo: nghiên cứu sử dụng mô hình đầy đủ để dự báo với khoảng dự báo là 3,7,14 và 30 ngày để xem xét mối quan hệ giữa độ chính xác và khoảng thời gian dự báo. Cụ thể, kết quả số ca mắc tả dự báo của từng mô hình sẽ được so sánh với số ca mắc tả thực tế để xem xét sự thay đổi của độ đo R^2 với độ dài của khoảng thời gian dự báo. Nghiên cứu tiến hành thực hiện xây dựng mô hình hồi qui tuyến tính với hai tập biến vào/ra như sau: **Các biến vào**: số ngày dự báo, quận/huyện, **biến ra**: độ chính xác dự báo, sử dụng độ đo R^2

Kết quả thực nghiệm mô hình hồi qui tuyến tính đã xây dựng cho thấy khi độ dài dự báo tăng lên 1 ngày, thì độ đo R^2 giảm đi 0.0076 với khoảng tin cậy 95% là [-.0095, -0.0057]. Chi tiết kết quả mô hình hồi qui này được trình bày trong Phụ lục 4 của luận án.

3.4 Mức độ quan trọng của các biến khí hậu: sử dụng biểu đồ boxplot để thể hiện giá trị các biến trong tất cả các mô hình như trình bày trên Hình 3.6.

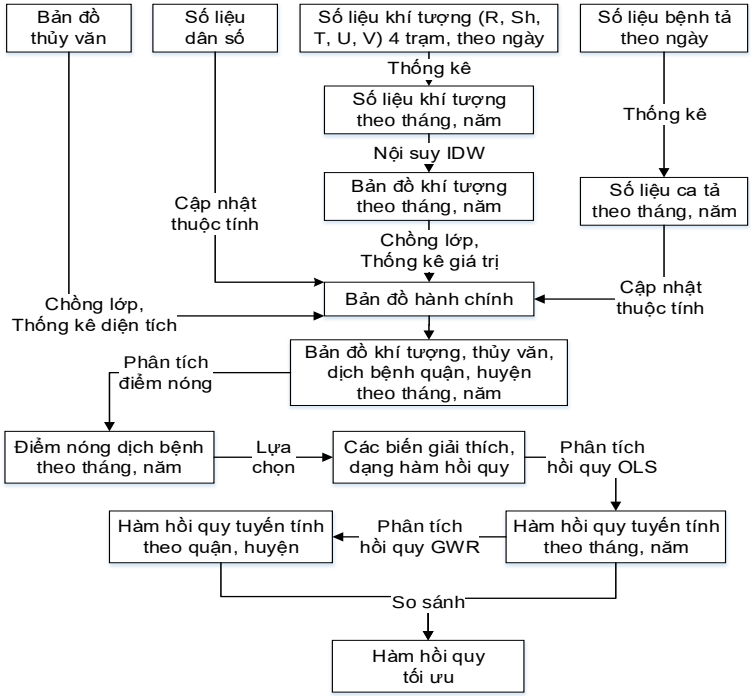


Hình 3.6. Mức độ quan trọng của các biến khí hậu trong các mô hình hồi qui RF

3.5 Kết luận: Các kết quả so sánh, phân tích cũng khẳng định rằng sự lân cận về địa lý và số ca bệnh ở các quận/huyện lân cận có mối liên hệ chặt chẽ. Nếu loại trừ yếu tố lân cận về địa lý trong xây dựng mô hình, hệ số xác đinh **R^2 của mô hình tăng lên đáng kể: 0.237 với dự báo trước 3 ngày, 0.115 với dự báo trước 7 ngày.** Các yếu tố khí hậu cũng có ảnh hưởng theo mức độ khác nhau đến số ca bệnh. Kết quả nghiên cứu cũng chỉ ra rằng, độ chính xác của mô hình dự báo giảm nếu tăng khoảng dự báo, với hệ số R^2 giảm trung bình 0,0076 nếu khoảng dự báo tăng 1 ngày.

CHƯƠNG 4 DỰ BÁO DỊCH TẢ DỰA TRÊN PHÂN TÍCH KHÔNG GIAN VỚI CÔNG NGHỆ GIS

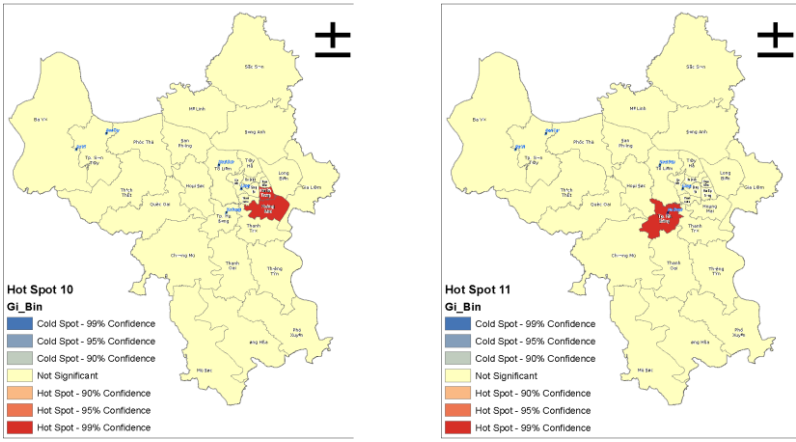
4.1 Mô hình dự báo đề xuất dựa trên phân tích không gian: Chương này nghiên cứu đề xuất mô hình dự báo dịch tả trên địa bàn Tp. Hà Nội với các yếu tố ảnh hưởng của biến đổi khí hậu trên cơ sở ứng dụng các kỹ thuật phân tích không gian của công nghệ GIS - Geographic Information System. Mô hình dự báo đề xuất dựa trên phân tích không gian mô tả trên Hình 4.1



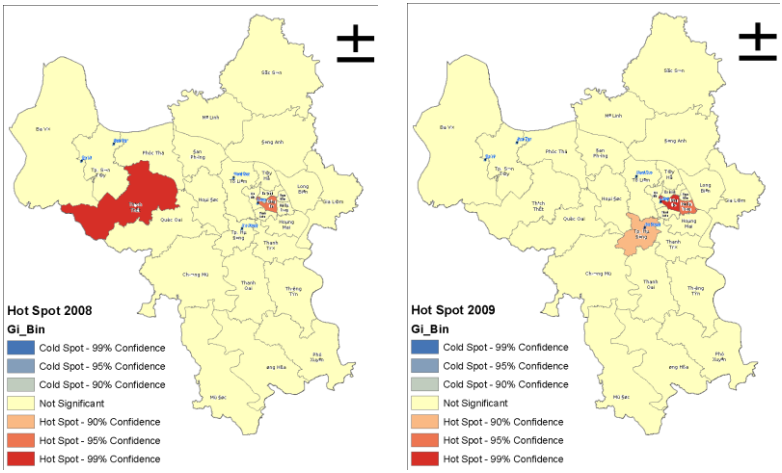
Hình 4.1. Mô hình dự báo đề xuất dựa trên phân tích không gian

4.2 Kết quả thực nghiệm: nghiên cứu tiến hành phân tích điểm nóng theo từng tháng, từng năm .Theo đó, có thể thấy các điểm nóng về số ca bệnh tả thay đổi theo từng tháng, từng năm tuy nhiên thường tập

trung quanh khu vực nội đô bao gồm các quận Ba Đình, Hoàn Kiếm, Hai Bà Trưng, Thanh Xuân, Đống Đa, Cầu Giấy. Đây là vùng tập trung dân cư đông đúc, tiếp giáp với một số con sông ô nhiễm chảy qua địa bàn.



Hình 4.6. Minh họa phân tích điểm nóng số ca bệnh tả tháng 10, 11



Hình 4.8. Minh họa hân tích điểm nóng số ca bệnh tả năm 2008, 2009

Các kết quả phân tích điểm nóng về ca bệnh tả theo tháng, năm, đều cho thấy các điểm nóng thường tập trung tại những khu vực dân

cư đông đúc và nằm gần các con sông. Từ nhận định trên kết hợp với các nghiên cứu đi trước về phân tích bệnh tả, nghiên cứu lựa chọn các biến giải thích phát sinh dịch tả , trên địa bàn Tp. Hà Nội như sau:

Theo tháng: các biến R, Sh, T, U, V lấy trung bình tháng; diện tích mặt nước (km^2).

Theo năm: các biến R, Sh lấy tổng theo năm; các biến T, U, V lấy trung bình năm; diện tích mặt nước (km^2), dân số (nghìn người).

Do số ca mắc bệnh tả phân bố rất không đều theo tháng và theo năm nên nghiên cứu lựa chọn hàm hồi qui logarit để giải thích số ca bệnh tả (y) với dạng như sau:

Theo tháng: Logarit $(y + 1) = \alpha + \beta_1 * R + \beta_2 * Sh + \beta_3 * T + \beta_4 * U + \beta_5 * V + \beta_6 *$ diện tích mặt nước $+ \varepsilon$ (sai số ngẫu nhiên)

Theo năm: Logarit $(y + 1) = \alpha + \beta_1 * R + \beta_2 * Sh + \beta_3 * T + \beta_4 * U + \beta_5 * V + \beta_6 *$ diện tích mặt nước $+ \beta_7 *$ dân số $+ \varepsilon$ (sai số ngẫu nhiên)

Trong đó: α là hệ số chặn, β_i là hệ số hồi qui.

Mô hình hồi qui OLS: Áp dụng hồi qui OLS theo tháng, theo năm để giải thích sự xuất hiện ca bệnh.

Bảng 4.2. Kết quả phân tích hồi qui OLS theo tháng khu vực Hà Nội

Tháng	Biến giải thích	R^2	*p_value
3	Hằng số, T,U,V	0.548761	p< 0,01
4	Hằng số, Mặt nước	0.250669	p< 0,01
5	Hằng số, Mặt nước, V	0.719093	p< 0,01
6	Hằng số, Mặt nước, R	0.414949	p< 0,01
7	Hằng số, Mặt nước, R, Sh,V	0.569390	p< 0,01
10	Hằng số, Mặt nước, Sh,T,V	0.549334	p< 0,01
11	Hằng số, R, Sh	0.380233	p< 0,01
12	Hằng số, Sh	0.324019	p< 0,01

**p_value: giá trị thống kê*

Bảng 4.3 Tổng hợp kết quả phân tích hồi qui OLS theo năm

Năm	Biến giải thích	R^2	*p_value
2007	Hằng số, Mặt nước,V	0.258771	< 0,01
2008	Hằng số, mặt nước, Dân số	0.424545	< 0,01
2009	Hằng số, mặt nước, V,Dân số	0.704000	< 0,01
2010	Hằng số, mặt nước, V,Dân số	0.637462	< 0,01
**p_value: giá trị thống kê*			

Mô hình hồi qui GWR: Phương pháp thống kê cục bộ hồi qui trọng số không gian (Geographically Weighted Regression- GWR) xem xét tính không đồng nhất của các mối quan hệ theo không gian. Dựa trên kết quả phân tích hồi qui OLS theo năm cho toàn khu vực, luận án xây dựng mô hình hồi qui trọng số không gian GWR tương ứng nhằm cải thiện khả năng giải thích của mô hình OLS, cũng như thiết lập hàm tuyến tính phù hợp cho từng quận huyện. Nghiên cứu sử dụng phương pháp chuẩn số thông tin AIC (Akaite's Information Criterion) để so sánh hai mô hình.

Bảng 4.4. So sánh hiệu quả giữa hai mô hình OLS và GWR theo năm

Chỉ số	2007		2008		2009		2010	
	OLS	GWR	OLS	GWR	OLS	GWR	OLS	GWR
AIC	101,10	101,10	105,38	104,65	81,83	73,51	81,81	78,94
R^2 hiệu chỉnh	0,26	0,26	0,42	0,46	0,70	0,84	0,64	0,69

4.3 Nhận xét: Qua phân tích các mô hình dự báo dịch tả dựa trên hồi qui OLS và GWR, luận án rút ra một số nhận xét như sau:

- Xét theo tháng, yếu tố khí hậu và mặt nước có ảnh hưởng đến dịch tả trên địa bàn Hà Nội trong giai đoạn 2001 - 2012. Đối với khí hậu, tác động này có thể quan sát được vào các tháng 3, 5, 6, 7, 10, 11, 12. Trong khi với mặt nước, là các tháng 4, 5, 6, 7, 10.
- Xét theo năm, tác động của yếu tố khí hậu đến số ca bệnh biểu hiện trong các năm 2007, 2009, 2010 là đáng kể, ngược lại trong năm 2008 tác động này không đáng kể. Đối với mặt nước, tác động của yếu tố này đến số ca bệnh thể hiện liên tục từ năm 2007 đến 2010.

Yếu tố dân số có ảnh hưởng đến số ca bệnh trong hai năm 2008 và 2010.

- Xét về không gian, số ca bệnh dự báo tại các khu vực nội đô thường nhỏ hơn số ca bệnh thực tế. Ngược lại, tại các khu vực phía Bắc và Nam, số ca bệnh dự báo thường lớn hơn số ca bệnh thực tế.
- Xét về mô hình, cả hai mô hình OLS và GWR đều có thể giải thích được số ca bệnh. Tuy nhiên, mô hình GWR cho kết quả tốt hơn mô hình OLS theo năm nhờ khả năng ước lượng các hệ số của mô hình thay đổi theo không gian.Một ưu điểm khác của mô hình GWR đó là khả năng hiển thị trực quan các hệ số ước lượng của mỗi biến giải thích theo từng đơn vị không gian, ở đây là các quận huyện. Điều này giúp cho việc khám phá các mối quan hệ phức tạp trở nên dễ dàng hơn.

4.4. Kết luận: Các kết quả đạt được trong thực nghiệm của chương khẳng định khả năng của GIS trong phân tích dự báo dịch tả trên địa bàn nghiên cứu khi chỉ ra được những điểm nóng, cũng như lý giải mối liên hệ giữa các biến khí hậu, mặt nước phân bố theo không gian với số ca bệnh phân bố theo thời gian. Đồng thời, các kết quả nghiên cứu cũng tạo tiền đề quan trọng cho quá trình mô phỏng, dự báo dịch tả trên địa bàn Tp. Hà Nội.

KẾT LUẬN & HƯỚNG PHÁT TRIỂN:

Luận án này tập trung xây dựng lớp các mô hình dự báo cho các kịch bản phòng chống dịch tả trên địa bàn thành phố Hà Nội, trong đó tập trung giải quyết ba vấn đề còn tồn tại trong công tác dự báo dịch tả , bao gồm (1) vấn đề lựa chọn kỹ thuật phù hợp xây dựng mô hình dự báo dịch tả với đặc thù dữ liệu thiếu và không cân bằng trên địa bàn thành phố Hà Nội, (2) vấn đề dự báo sự bùng phát dịch tả trong ngắn hạn, có xem xét toàn diện ảnh hưởng của các yếu tố khí hậu và địa lý và (3) xây dựng mô hình dự báo dịch tả tổng quát cho thành phố Hà Nội.

Đối với vấn đề lựa chọn kỹ thuật phù hợp xây dựng mô hình dự

báo dịch tả với đặc thù dữ liệu thiếu và không cân bằng trên địa bàn thành phố Hà Nội, luận án đề xuất sử dụng phương pháp cửa sổ trượt nhằm tăng số điểm dữ liệu và khảo sát một lớp các kỹ thuật học máy thống kê và hồi quy cho xây dựng mô hình dự báo để nhằm thay thế cho mô hình dịch tễ học toán học. Các kỹ thuật xây dựng mô hình bao gồm ba bộ phân lớp (RandomForest, Naïve Bayes, SVM,) và hồi qui tuyến tính. Các kết quả quả thực nghiệm khẳng định phương pháp cửa sổ trượt là phù hợp và kỹ thuật hồi qui không phù hợp, phân lớp Random Forest cho kết quả dự báo tốt nhất trong số các kỹ thuật được sử dụng để xây dựng mô hình phân 3 lớp.

Đối với vấn đề dự báo sự bùng phát dịch tả trong ngắn hạn, có xem xét toàn diện ảnh hưởng của các yếu tố khí hậu và địa lý, luận án đề xuất sử dụng kỹ thuật hồi qui Random Forest để xây dựng các mô hình dự báo trong ngắn hạn, có xem xét mức độ ảnh hưởng của các yếu tố khí hậu và lân cận địa lý. Các mô hình đầy đủ (DD), độc lập khí hậu (DLKH) và độc lập địa lý (DLDL) đã được xây dựng cho từng quận/huyện của Hà Nội để lựa chọn mô hình tốt nhất và khảo sát mức độ ảnh hưởng của các yếu tố khí hậu và lân cận địa lý lên độ chính xác dự báo. Kết quả cho thấy mô hình đầy đủ cho kết quả dự báo tốt nhất và độ chính xác của mô hình dự báo giảm nếu tăng khoảng dự báo, với hệ số R^2 giảm trung bình 0,0076 nếu khoảng dự báo tăng 1 ngày. Các kết quả so sánh, phân tích mức độ ảnh hưởng của các yếu tố địa lý và khí hậu khẳng định rằng sự lân cận về địa lý và số ca bệnh ở các quận/huyện lân cận có mối liên hệ chặt chẽ. Các yếu tố khí hậu cũng có ảnh hưởng theo các mức khác nhau đến số ca bệnh, trong đó nhiệt độ và độ ẩm trung bình ngày có mức ảnh hưởng lớn nhất, trong khi đó tốc độ gió và SOI có mức ảnh hưởng thấp nhất.

Đối với vấn đề xây dựng mô hình dự báo dịch tả tổng quát cho thành phố Hà Nội, luận án đề xuất xây dựng mô hình dự báo dịch tả tổng quát cho thành phố Hà Nội dựa trên các kỹ thuật phân tích không gian sử dụng công nghệ GIS. Các tập dữ liệu Bản đồ hành chính, Bản

đồ thủy văn, Số liệu dân số, Số liệu khí tượng và Số liệu bệnh tả được tích hợp, chồng lớp trên bản đồ hành chính sử dụng công nghệ GIS làm đầu vào cho quá trình xây dựng và thử nghiệm mô hình dự báo. Các kỹ thuật phân tích điểm nóng bùng phát dịch tả, các kỹ thuật hồi quy tuyến tính OLS và hồi quy trọng số không gian GWR được sử dụng để lựa chọn mô hình dự báo tối ưu. Các kết quả đạt được khẳng định khả năng sử dụng GIS hiệu quả trong phân tích dự báo dịch tả khi chỉ ra được những điểm nóng bùng phát dịch, cũng như lý giải mối liên hệ giữa các biến khí hậu, mặt nước phân bố theo không gian với số ca bệnh phân bố theo thời gian. Kết quả thực nghiệm cũng khẳng định hồi quy trọng số không gian GWR cho kết quả dự báo chính xác nhất trong hầu hết các trường hợp.

Tổng hợp những đóng góp chính của luận án bao gồm:

- Đề xuất mô hình dự báo dịch tả dựa trên khai phá luật kết hợp và học máy hồi qui, phân lớp.
- Đề xuất mô hình dự báo dịch tả ngắn hạn có đánh giá mức độ ảnh hưởng của các yếu tố khí hậu và địa lý đến sự bùng phát dịch tả .
- Đề xuất mô hình dự báo dịch tả tổng quát dựa trên phân tích không gian ứng dụng công nghệ GIS.

Luận án có thể được tiếp tục phát triển theo các hướng sau:

- **Vấn đề thứ nhất:** Nghiên cứu nâng cấp các mô hình thành hệ hỗ trợ ra quyết định hoàn chỉnh phục vụ cho dự báo dịch bệnh trong ngành y tế.
- **Vấn đề thứ hai:** Tiếp tục bổ sung dữ liệu với khoảng thời gian lớn hơn và tích hợp các mô hình để giải thích thêm các yếu tố không gian, địa lý, sự lây truyền bệnh từ người sang người và có tích hợp sử dụng các mô hình dịch tễ học.Nghiên cứu thiết lập một bộ phân lớp kết hợp để có được kết quả tốt hơn.

DANH MỤC CÁC BÀI BÁO CÔNG BỐ

[1] Le Thi Ngoc Anh, Hoang Xuan Dau and Nguyen Hoang Phuong (2015), "Cholera forecast based on mining association rules", *2015 International Conference on Communications, Management and Telecommunications (ComManTel)*, DaNang, 2015, pp. 133-137. DOI: 10.1109/ComManTel.2015.7394274

[2] Lê Thị Ngọc Anh, Hoàng Xuân Dậu(2015), "Dự báo dịch tả dựa trên mô hình học máy phân lớp", *Kỷ yếu hội thảo quốc gia 2015 về điện tử, truyền thông và công nghệ thông tin (ECIT2015*).ISBN:978-604-67-0635-9, tr:348-352.

[3] Lê Thị Ngọc Anh, Nguyễn Thị Thanh Xuân, Hoàng Xuân Dậu, Bùi Trung Dũng (2016), "Kỹ thuật học máy phân lớp với dự báo dịch tả ". *Tạp chí khoa học công nghệ Đại học Đà Nẵng*, Vol3(100), ISSN 1859-1531, tr:1- 4.

[4] Ngoc-Anh Thi Le, Thi-Oanh Ngo, Huyen-Trang Thi Lai, Hoang-Quynh Le, Hai-Chau Nguyen, Quang-Thuy Ha (2016)."An Experimental Study on Cholera Modeling in Hanoi". *Intelligent Information and Database Systems - 8th Asian Conference, ACIIDS 2016, March 14-16, 2016, Da Nang, Vietnam, Volume: Proceedings, Part II*, pp:230-240

[5] Nguyen Hai Chau, Le Thi Ngoc Anh (2016),"Using Local Weather and Geographical Information to Predict Cholera Outbreaks in Hanoi, Vietnam", *Proceeding of the 4th International Conference on Computer Science, Applied Mathematics and Applications*, (ICCSAMA 2016)Advanced Computational Methods for Knowledge Engineering, pp.195-212.

[6] Lê Thị Ngọc Anh, Hoàng Xuân Dậu (2016), "Ứng dụng GIS trong dự báo dịch tả ", *Tạp chí Khoa học Công nghệ thông tin và truyền thông*, Vol1(CS1), ISSN:2525-2224, tr:69-78.

[7] Lê Thị Ngọc Anh, Hoàng Xuân Dậu, Nguyễn Hoàng Phương (2017), Thiết lập công cụ mô phỏng dự báo dịch tả bằng công nghệ GIS. *Tạp chí Khoa học và Công nghệ Đại học Thái Nguyên*, Vol6(166), ISSN 1859-2171,tr:21-26.