HỌC VIỆN CÔNG NGHỆ 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

LUẬN ÁN TIẾN SĨ KỸ THUẬT

HÀ NỘI – 2018

HỌC VIỆN CÔNG NGHỆ 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

LUẬN ÁN TIẾN SĨ KỸ THUẬT

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. PGS.TS Nguyễn Hoàng Phương

2. TS. Hoàng Xuân Dậu

HÀ NỘI – 2018

LỜI CAM ĐOAN

Tôi cam đoan rằng nội dung của luận án này là kết quả nghiên cứu của bản thân. Tất cả những tham khảo từ các nghiên cứu liên quan đều được nêu rõ nguồn gốc một cách rõ ràng trong danh mục tài liệu tham khảo được đề cập ở phần sau của luận án. Những đóng góp trong luận án là kết quả nghiên cứu của tác giả đã được công bố trong các bài báo của tác giả ở phần sau của luận án và chưa được công bố trong bất kỳ công trình khoa học nào khác.

Tác giả luận án

Lê Thị Ngọc Anh

LỜI CẢM ƠN

Trong suốt quá trình học tập và hoàn thành luận án, tôi đã nhận được sự hướng dẫn, giúp đỡ quý báu của các thày, các anh, chị, em và các bạn bè đồng nghiệp. Với lòng kính trọng và biết ơn sâu sắc tôi xin được bày tỏ lời cảm ơn chân thành tới:

- Tập thể thày hướng dẫn PGS.TS Nguyễn Hoàng Phương và Tiến sĩ Hoàng Xuân Dậu, hai người thày kính mến đã hết lòng giúp đỡ, dạy bảo, động viên và tạo mọi điều kiện thuận lợi cho tôi trong suốt quá trình học tập và hoàn thành luận án.

- PGS.TS Hà Quang Thụy, PGS.TS Nguyễn Hải Châu- Trường Đại Học Công nghệ - Đại học Quốc Gia Hà Nội đã đóng góp những ý kiến vô cùng quý báu trong quá trình nghiên cứu và hoàn thiện luận án.

- Tập thế cán bộ Trung tâm nghiên cứu và đào tạo nguồn nhân lực y tế, tập thể cán bộ Trung tâm y tế dự phòng Hà nội, tập thể cán bộ Trung tâm Nghiên cứu khí tượng thủy văn Trung ương, Sở khoa học và công nghệ thành phố Hà nội đã tạo điều kiện cho tôi trong quá trình thu thập số liệu và tiến hành nghiên cứu.

- Xin gửi lời cảm ơn sâu sắc tới Ban giám đốc, Khoa quốc tế và đào tạo Sau đại học của Học viên Công nghệ Bưu chính Viễn thông đã giúp đỡ và tạo mọi điều kiện thuận lợi trong quá trình học tập và nghiên cứu.

- Xin gửi lời cảm ơn tới Ban Giám Hiệu, Ban quản lý dự án Việt Nam – Hà Lan, Phòng Công nghệ thông tin của Trường Đại học Y Hà Nội, các bạn bè, đồng nghiệp đã giúp đỡ, động viên những lúc tôi gặp khó khăn và tạo mọi điều kiện thuận lợi nhất cho tôi thực hiện nghiên cứu và hoàn thành luận án.

- Xin dành tất cả sự yêu thương và lời cảm ơn tới gia đình, bố mẹ, các anh chị em và người thân luôn bên cạnh động viên và giúp đỡ tôi học tập, làm việc và hoàn thành luận án.

Xin chân thành cảm ơn.

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

TỪ VIẾT TẮT	DIỄN GIẢI	
	TIẾNG ANH	TIẾNG VIỆT
AIC	Akaite's Information Criterion	Chuẩn số thông tin
CC	Correlation coefficient	Hệ số tương quan
DT	Decission Trees	Cây quyết định
GIS	Geographic Information System	Hệ thống thông tin đại lý
GWR	Geographically Weighted Regression	Hồi qui trọng số không gian
IDW	Inverse Distance Weight	Nội suy trong số không gian
kNN	K Nearest Neighbors	Thuật toán K láng giềng
LM	Linear Regression	Hồi qui tuyến tính
MAE	Mean Absolute error	Sai số tuyệt đối
MSE	Mean square error	Sai số quân phương
MDIMC	Multi Dimensional Inhomogeneous Makov Chain	Mô hình Makov đa chiều không đồng nhất
OLS	Ordinary Least square	Hồi qui ước lượng bình phương nhỏ nhất.
RF	Random Forest	Rừng ngẫu nhiên
RMSE	Root mean square error	Sai số quân phương
SOI	Southern Oscillation Index	Chỉ số dao động phía nam đo sự thay đổi cường độ ElNino và Lania
SIR	Susceptible – Infectious- Recoved	Mô hình lan truyền dịch bệnh
SVM	Support Vector Machine	Máy vector hỗ trợ
V.vibrios	Vibrio Cholera	Vi khuẩn tả

DANH MỤC CÁC KÝ HIỆU

KÝ HIỆU	DIỄN GIẢI
KPDL	Khai phá dữ liệu
CSDL	Cơ sở dữ liệu
β	Tốc độ truyền nhiễm
γ	Tỷ lệ hồi phục
R_0	Lượng tái nhiễm cơ bản
β	Tốc độ truyền nhiễm
$D_{example}$	Tập dữ liệu là tài nguyên cư bản cho xây dựng mô hình
D_{test}	Tập dữ liệu để kiểm thử đánh giá mô hình
DL_1	Tập dữ liệu theo ngày
DL_2	Tập dữ liệu theo tháng
KH_i	Giá trị khí hậu tại thời điểm i
QH_i	Quận/ huyện thứ i
$LCQH_i$	Quận/huyện lân cận của QH_i
$DT_{i,t}$	Giá trị dịch tả tại QH_i tại thời điểm t
$DTLC_{i,t}$	Giá trị dịch tả của quận/huyện lân cận với quận/huyện đang xem xét tại thời điểm t
DT_{t-2}	Giá trị dịch tả thời thời điểm trong quá khứ 2 tháng trước

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

DANH MỤC BIỂU ĐỒ

MỞ ĐẦU

Tính cấp thiết

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 từ địa phương, quốc gia, thế giới cần được giải quyết. Chính vì vậy, dự báo trong y tế nói chung và dự báo dịch bệnh nói riêng luôn nhận được sự quan tâm của cộng đồng nghiên cứu. Nhằm góp phần ngăn chặn sự bùng phát và lây lan của dịch bệnh, đã có nhiều công trình nghiên cứu được công bố và ứng dụng, trong đó dự báo sớm là một biện pháp góp phần đáng kể. Các kết quả nghiên cứu dự báo dịch bệnh trong thời gian qua là bằng chứng quan trọng cho việc lập kế hoạch và quản lý các hoạt động chăm sóc sức khỏe. Dự báo được coi là công cụ hữu ích cho các nhà quản lý và hoạch định chính sách. 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 bối cảnh việc thực hiện các nghiên cứu thường bị hạn chế về cả thời gian và nguồn lực, việc sử dụng mô hình khai phá dữ liệu, học máy trong dự báo dịch bệnh là một phương pháp thích hợp, có khả năng giải quyết được tính phức tạp của bài toán dự báo dịch bệnh với chi phí thấp. Ở Việt Nam, ứng dụng khai phá dữ liệu, học máy trong dự báo dịch bệnh vẫn là một lĩnh vực non trẻ. Số lượng các chuyên gia về lĩnh vực này cũng như các nghiên cứu ứng dụng các phương pháp dự báo dịch bệnh trong y tế còn hạn chế trong khi nhu cầu cần bằng chứng trong xây dựng các chương trình, chính sách y tế đang ngày càng gia tăng.

Ngày nay, các bệnh truyền nhiễm đang có xu hướng giảm trong cộng đồng, nhưng dưới sự tác động của nhiều yếu tố như biến đổi khí hậu, môi trường và ý thức con người, nhiều bệnh dịch truyền nhiễm đã được thanh toán trước đây, nay tái xuất hiện và cùng với đó, nhiều bệnh dịch mới nổi lên, đặc biệt ở các vùng chịu ảnh hưởng của biến đổi khí hậu và đời sống kinh tế khó khăn. Chính vì vậy việc tìm hiểu nguyên

nhân dịch bệnh đã không còn gói gọn trong việc phát hiện căn nguyên vi sinh vật, mà mở rộng ra cho nhiều loại yếu tố tự nhiên, xã hội và sinh học có các mức độ liên quan với số ca mắc bệnh trong cộng đồng. Ngoài việc phát hiện ra căn nguyên và các yếu tố ảnh hưởng, cần xây dựng các mô hình dự báo sử dụng các kỹ thuật khác nhau dựa vào các thông số về tự nhiên, như khí hậu, môi trường, và hành vi, thói quen trong cộng đồng..., nhằm cảnh báo sớm dịch bệnh, giúp giảm thiểu nguy cơ, tổn thất có thể xảy ra cho con người. 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.

Tình hình nghiên cứu

Hiện nay đã có nhiều mô hình được xây dựng nhằm cảnh báo dịch bệnh sớm giúp giảm thiểu nguy cơ, tổn thất xảy ra cho con người dựa vào các thông số về thời tiết [20],[33],[46], [52], [62], [82] [86] ,[94],[95], [100]. Các phương pháp dự báo dịch bệnh ban đầu đều dựa trên mô hình lan truyền dịch bệnh, điển hình là mô hình dịch tễ học toán học SIR (Susceptible – Infectious – Recovered) [24], [35]. Mô hình lan truyền dịch bệnh này chia quần thể nghiên cứu thành ba lớp, bao gồm lớp chứa các thành phần dễ bị nhiễm bệnh (*Susceptible*), lớp nhiễm bệnh chứa các thành phần bị nhiễm bệnh và có khả năng truyền bệnh cho người khác (*Infectious*) và lớp hết bệnh chứa các thành phần đã hồi phục hoặc tử vong do nhiễm bệnh (*Recovered*). 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 đầu vào 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 thời điểm *t* trong tương lai. Mô hình dịch tễ học toán học đã được áp dụng thành công với các hệ thống không quá phức tạp hoặc đã có nhiều kết quả quan sát về hệ thống.

Tuy nhiên, trong trường hợp các quan sát thu nhận được quá phức tạp hoặc không rõ ràng thì việc xây dựng các phương trình theo tiếp cận của mô hình dịch tễ học toán học gặp rất nhiều khó khăn.

Trong trường hợp các quan sát thu nhận được quá phức tạp hoặc không rõ ràng, tiếp cận theo mô hình học máy thống kê có nhiều ưu thế trong giải quyết bài toán dự báo dịch bệnh. Một mô hình thống kê thường là một tập các phương trình với các tham số điều khiển mà giá trị của tham số này nhận được nhờ một quá trình "*học*" từ dữ liệu quan sát. Cấu trúc các phương trình này là một kết hợp của các tham số điều khiển và các đặc trưng hệ thống, có thể ở dạng đơn giản (tuyến tính), hoặc ở dạng phức tạp (phi tuyến). Mô hình thống kê được chia làm hai loại là mô hình hồi qui và mô hình phân lớp, trong đó mô hình hồi qui tương ứng với miền giá trị của biến đầu ra liên tục còn mô hình phân lớp tương ứng với miền giá trị đầu ra rời rạc. Ở những năm 1990, phương pháp phân tích hồi quy tuyến tính được sử dụng thường xuyên trong việc thiết lập các mô hình cảnh báo dịch bệnh [10], [65],[67],[77],[79].

Trong thời gian gần đây, mô hình phân tích chuỗi thời gian (*time-series*) đã được sử dụng rộng rãi trong nghiên cứu ảnh hưởng của khí hậu và số lượng ca mắc các bệnh truyền nhiễm ở những cộng đồng cụ thể và dự báo quy mô dịch bệnh trong tương lai[1],[58], [61]. Việc sử dụng mô hình phân tích chuỗi thời gian góp phần khắc phục nhược điểm của các mô hình hồi qui luận lý (*logistic*) hoặc hồi qui đa biến trước đó, do không có khả năng xem xét đến tính tự tương quan (*auto-correlation*) đối với những dữ liệu mang tính chuỗi thời gian, làm giảm khả năng tiên đoán.

Nhằm cải thiện độ chính xác trong thiết lập mô hình cảnh báo dịch bệnh, một số nhà nghiên cứu đã tiến hành lồng ghép mô hình phân tích chuỗi thời gian và mô hình GIS, nhằm xác định cụ thể ảnh hưởng của sự kết hợp giữa điều kiện địa lý và điều kiện khí hậu tới số ca mắc một bệnh truyền nhiễm nào đó. Sự kết hợp thống nhất giữa dữ liệu thuộc tính với dữ liệu không gian trong công nghệ GIS cho phép người sử dụng, ngoài các dữ liệu thuộc tính, thông tin định lượng, còn có khả năng quan sát trên không gian bản đồ, có tầm nhìn bao quát hơn trong quá trình phân tích số liệu, hoàn cảnh tình huống, đưa ra các dự báo và lựa chọn quyết định đúng đắn hơn [43].Vì

những lý do đó, công nghệ GIS đang ngày càng được ứng dụng rộng rãi trong nghiên cứu kiểm soát và dự báo dịch bệnh [43],[70].

Từ các phân tích nêu trên, luận án thực hiện nghiên cứu kết hợp mô hình GIS và mô hình chuỗi thời gian để thiết lập mô hình dự báo thống nhất, trong đó xem xét ảnh hưởng của các yếu tố khí hậu, không gian và thời gian đến độ chính xác của mô hình dự báo. Tại Việt Nam, các nghiên cứu về dự báo dịch bệnh còn rất thiếu, do đó cần phải có những nghiên cứu chuyên sâu về mô hình dự báo các dịch bệnh truyền nhiễm để đáp ứng các yêu cầu của việc bảo vệ, chăm sóc và nâng cao sức khỏe cho nhân dân một cách chủ động và toàn diện.

Lý do chọn đề tài

Trong những năm gần đây, các chương trình trọng điểm giám sát bệnh truyền nhiễm của ngành y tế Việt Nam đã được thực hiện và các dữ liệu thu thập đã được lưu trữ một cách có hệ thống. Từ đó, các kho dữ liệu về quá trình bùng phát dịch bệnh và dữ liệu về khí hậu, thủy văn cũng được hình thành và ngày càng đầy đủ hơn. Đây là một thuận lợi lớn cho việc xây dựng các mô hình dự báo bệnh dịch dựa trên khai phá dữ liệu. Tuy nhiên, theo khảo sát của tác giả, Việt Nam còn thiếu các mô hình dự báo dịch bệnh, đặc biệt là các mô hình dự báo kết hợp dựa trên các dữ liệu đa ngành, trong đó có xem xét đầy đủ các yếu tố như khí hậu, không gian, thời gian,... Từ phân tích trên, luận án tập trung nghiên cứu thiết lập mô hình dự báo dịch tả dựa trên các kỹ thuật khai phá dữ liệu và học máy thống kê, trong đó có xem xét ảnh hưởng của các yếu tố như khí hậu, không gian, thời gian. Đây sẽ là một công cụ thực sự hữu ích cho những người làm công tác y tế dự phòng và quản lý y tế.

Mục tiêu tổng quát:

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 có sự kết hợp dữ liệu không gian, thời gian và khí hậu.

Mục tiêu cụ thể:

Nghiên cứu tổng quan, lựa chọn phương pháp thích hợp trong dự báo dịch tả;

Mô hình hóa các yếu tố khí hậu ảnh hưởng đến dịch tả;

Xây dựng mô hình tích hợp dữ liệu thời gian, không gian địa lý lân cận trong (GIS) và dữ liệu khí hậu để dự báo dịch tả tại Hà Nội;

Đề xuất ứng dụng mô hình dự báo trong thực tiễn.

Đối tượng và phạm vi nghiên cứu:

Để xây dựng mô hình dự báo dịch tả ở Hà nội, luận án sử dụng các tập dữ liệu sau: Tập dữ liệu về dịch tả , tập dữ liệu về khí hậu, tập dữ liệu địa lý của Hà nội và tập dữ liệu về chỉ số giao động phía nam (SOI). Thông tin về tập dữ liệu này sẽ được mô tả trong Chương 1 của luận án. Bên cạnh việc hồi cứu dữ liệu phục vụ cho nghiên cứu, luận án cũng xem xét một số thuật toán và kỹ thuật học máy áp dụng trong dự báo, như hồi qui, phân lớp sử dụng cây quyết định, support vector machine, rừng ngẫu nhiên,... và các kỹ thuật phân tích không gian trong GIS.

Phạm vi không gian ứng dụng mô hình là toàn bộ thành phố Hà Nội. Đây là một trong những thành phố lớn nhất trong cả nước với diện tích là 3.328,9 km^2, dân số trung bình theo năm 2011 là 6.561.900 người, mật độ dân số là 2.013 người/km^2 với tỷ lệ nhập cư lớn và là cửa ngõ giao thông quan trọng của cả nước.

Phạm vi nghiên cứu và các giả thiết của luận án 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 không đáng kể.
- Các yếu tố xã hội và hành vi- thói quen ăn uống trong cộng đồng, sự can thiệp của các chương trình y tế được coi là không đáng kể.
- Người nhiễm bệnh đã hết bệnh thì người này không còn khả năng nhiễm bệnh trong cùng một khoảng thời gian dự báo.

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.

Cấu trúc của luận án

Ngoài phần Mở đầu và Kết luận, luận án có cấu trúc các chương sau:

Chương 1: Tổng quan về các mô hình dự báo dịch bệnh: Nội dung của chương mô tả khái niệm, những thuật ngữ cũng như tổng quan các công trình nghiên cứu về mô hình dự báo dịch bệnh trong y tế của cộng đồng nghiên cứu trong nước và thế giới.

Chương 2: Đề xuất mô hình dự báo dịch tả dựa trên khai phá luật kết hợp và học máy hồi qui, phân lớp: Nội dung chương đề xuất ứng dụng khai phá luật kết hợp, học máy hồi qui, phân lớp để dự báo dịch tả tại Hà Nội.

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. Nội dung chương đề xuất phân rã dữ liệu theo phương pháp cửa sổ trượt để dự báo và đánh giá độ ảnh hưởng của yếu tố khí hậu, không gian địa lý và thời gian trong mô hình.

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

1.1.1. Khái niệm

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. Khi tiến hành dự báo cần căn cứ vào việc thu thập, xử lý dữ liệu trong quá khứ và hiện tại để xác định xu hướng vận động của các hiện tượng trong tương lai dựa vào một số mô hình toán học (định lượng). Tuy nhiên, dự báo cũng có thể là một dự đoán chủ quan hoặc trực giác về tương lai (định tính) và để dự báo định tính được chính xác hơn, người ta thường cố gắng loại trừ tính chủ quan của người dự báo. Phân tích dự báo là quá trình khám phá ra mô hình mẫu thú vị và có ý nghĩa trong dữ liệu.

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. Mô hình có thể là một mô hình tĩnh biểu diễn một hệ thống "tại vị" hoặc là một mô hình động biểu diễn cho một quá trình [97]. Mô hình hóa hay xây dựng mô hình giúp chúng ta hiểu được các hiện tượng đang xảy ra, hiểu được các thành phần trong đó tương tác với nhau như thế nào, hoặc để dự đoán những gì có thể xảy ra khi các hiện tượng thay đổi hoặc tiến hóa.

1.1.2. Một số thuật ngữ liên quan

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. Ở đây, có hai yếu tố liên quan tới việc tiến hành hoạt động dự báo. Thứ nhất, dự báo được tiến hành chỉ khi có tính không chắc chắn; Ví dụ như dự báo ngày mai mặt trời có mọc hay không là không cần thiết do chắc chắn mặt trời mọc hàng ngày, song dự báo ngày mai có mưa hay không là rất cần thiết. Thứ hai, chủ thể dự báo không điều khiển được giá trị của đại lượng cần được dự báo; như vậy, không đặt ra việc dự báo về nhiệt độ trong phòng vì chủ nhân của nó có thể có

các phương tiện đảm bảo nhiệt độ của phòng ở một phạm vi cho phép, song lại cần dự báo về nhiệt độ ngoài trời.

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. Trong tiếng Anh, các thuật ngữ "*predict*", "*forecast*" là thông dụng và trong một số trường hợp thì thuật ngữ "*foresight*" (*nhìn trước*) được sử dụng, song *foresight* thường đề cập tới "phương pháp" dự báo. Trong nhiều trường hợp, có sự phân biệt ngữ nghĩa của ba thuật ngữ tiếng Anh này. "*Predict*" là dự báo trong phạm vi dữ liệu hiện có (tương tự như "dự đoán" trong tiếng Việt), "*forecast*" là dự báo ngoài miền dữ liệu đó. *Foresight* thường được sử dụng trong lĩnh vực kinh tế - xã hội mà trong nhiều trường hợp có ý nghĩa tương tự như "forecast" song đề cập tới khoảng thời gian dự báo xa (dài) hơn và liên quan tới các đại lượng có tính chiến lược.

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

Sự lan truyền dịch bệnh vừa là một quá trình xã hội vừa là một quá trình sinh học[35],[92]. Sự lan truyền dịch bệnh là một quá trình xã hội vì các cá nhân trong một quần thể lan truyền dịch bệnh cho nhau qua các quan hệ xã hội (di truyền, tiếp xúc trực tiếp, gián tiếp,..). Sự lan truyền dịch bệnh là một quá trình sinh học vì sự phát triển của các vi sinh vật gây bệnh dịch được sinh sôi, phát triển và lan truyền trong cộng đồng theo các quá trình sinh học tương ứng với vi sinh vật gây bệnh dịch. Nói chung, công việc dự báo dịch bệnh được tiến hành qua hai giai đoạn: mô hình hóa quá trình lan truyền dịch bệnh dựa trên các dữ liệu thu thập được và dự báo giá trị của các biến trong tương lai dựa trên mô hình đã được xây dựng.

Hầu hết các phương pháp dự báo dịch bệnh truyền thống đều dựa trên mô hình lan truyền dịch bệnh, nên mục sau đây sẽ tập trung giới thiệu mô hình dự báo dịch bệnh ở mức độ cơ bản nhất, điển hình là mô hình dịch tễ học toán học mà đại diện là mô hình SIR và sau đó là một số mô hình dự báo dịch bệnh bằng khai phá dữ liệu và

phân tích dự báo không gian.

1.2.1 Một số mô hình dự báo dịch bệnh

1.2.1.1 Mô hình dịch tễ học toán học

Fred Brauer và cộng sự [24], cho rằng hầu hết mô hình dịch bệnh dựa trên việc chia quần thể đang nghiên cứu thành một số lượng nhỏ các ngăn (*compartment*) tương ứng với số lượng trạng thái liên quan tới bệnh dịch mà các cá nhân trong quần thể có thể rơi vào; ở đây, mỗi ngăn chứa các cá nhân có tình trạng bệnh dịch giống hệt nhau. Đối với mỗi bệnh dịch, các cá nhân có thể trải qua các trạng thái trong vòng đời bệnh dịch. 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.

Để chuyển trạng thái từ trạng thái dễ bị nhiễm (*S*) sang trạng thái đang nhiễm bệnh (*I*), cá nhân đó phải tiếp xúc với các cá nhân đang nhiễm bệnh. Theo khung nhìn của quá trình xã hội (*mô hình mạng*), hai cá nhân tiếp xúc nhau khi họ là các "*nút láng giềng*" của nhau theo các quan hệ xã hội (di truyền, tiếp xúc trực tiếp, tiếp xúc gián tiếp qua đường nước hoặc các sinh vật trung gian...)[35]. Để chuyển trạng thái từ trạng thái nhiễm bệnh (*I*) sang trạng thái hồi phục (*R*), cá nhân đó được sử dụng văcxin hoặc bị tử vong.Trong mô hình dự báo dịch bệnh, các chữ cái *S, I, R* được dùng để chỉ số lượng cá nhân trong các ngăn *S, I, R* tương ứng. Trong nhiều trường hợp, số lượng người trong quần thể đang xem xét *N* (*N = S + I + R*) được giả thiết là một hằng số. Bài toán dự báo dịch bệnh xem xét việc biến đổi các giá trị *S, I, R* theo thời gian *t*, theo đó, *S(t), I(t), R(t)* là giá trị của *S, I, R* tương ứng tại thời điểm *t*. 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. Dạng đơn giản của mô hình *SIR* là hệ hai phương trình [24]:

$$\frac{dS}{dt} = -\beta SI \tag{1.1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{1.2}$$

trong đó, tốc độ truyền nhiễm (bình quân đầu người) là β và tỷ lệ hồi phục γ (vì vậy khoảng lây nhiễm trung bình là $1/\gamma$). Lưu ý, *I* không được viết một phương trình vi phân cho lượng cá thể bị biến mất. Tại thời điểm ban đầu, mọi cá thể ở trạng thái dễ bị nhiễm ($S(0)=N$), sau đó một cá thể bị nhiễm bệnh và có khả năng truyền bệnh cho các cá thể khác với tỷ lệ βN trong khoảng thời gian $1/\gamma$. Như vậy, cá nhân bị nhiễm bệnh đầu tiên đó có thể lây nhiễm tới $R_0 = \beta N/\gamma$ cá thể mới. R_0 được gọi là lượng tái nhiễm cơ bản (*basic reproduction number*) và đây là một đại lượng quan trọng nhất trong phân tích mọi mô hình dịch bệnh; số lượng nhiễm bệnh *I* chỉ tăng khi $R_0 > 1$. Để giải quyết mô hình *SIR* cơ bản, đầu tiên tích hợp hai phương trình (1.1) và (1.2) để nhận được:

$$\frac{dI}{dS} = \frac{\frac{dI}{dt}}{\frac{dS}{dt}} = \frac{\beta SIn - \gamma I}{-\beta SI} = -1 + \frac{\gamma}{\beta S} = -1 + \frac{S(0)}{R_0 S}$$

$$\frac{dI}{dS} = -1 + \frac{1}{R_0 S} \tag{1.3}$$

và sau đó lấy nguyên hàm:

$$I = I(0) + S(0) - S + \frac{1}{R_0} \ln \frac{S}{S_0} \tag{1.4}$$

Đây là một lời giải xác định tường minh cho I, nhưng lại đáng tiếc rằng nó là một hàm của S mà không phải là một hàm của t như mong muốn. Cho đến nay, vẫn chưa có một lời giải chính xác cho I là một hàm của t [24].

Có một số phương án xấp xỉ được đề xuất, trong đó có phương pháp Ơle: Với giả thiết là trong khoảng thời gian Δt đủ nhỏ thì dS/dt xấp xỉ bằng $\Delta S/\Delta t$ (xấp xỉ vi phân bằng sai phân), trong đó $\Delta S = S(t+\Delta t) - S(t)$; và như vậy, xấp xỉ số lượng cá thể dễ bị nhiễm tại thời điểm trong tương lai $t+\Delta t$ như sau:

$$S(t+\Delta t) = S(t) - \beta S(t)I(t)\Delta t \quad (1.5)$$

Tương tự, xấp xỉ số lượng cá thể dễ bị nhiễm tại thời điểm trong tương lai $t+\Delta t$ như sau:

$$I(t+\Delta t) = I(t) + \beta S(t)I(t)\Delta t - \gamma I(t)\,\Delta t \quad (1.6)$$

Cặp hai phương trình (1.5, 1.6) cung cấp một sơ đồ của giải pháp xấp xỉ mô hình *SIR* cơ bản. Để mô hình hóa dựa dịch bệnh dựa trên sơ đồ này, bước thời gian Δt cần được xác định đủ nhỏ và cung cấp các giá trị tham số về tốc độ lây lan và hồi phục (β và γ, hoặc R_0, N và γ) cũng như các giá trị khởi đầu ($R(0)$ và $I(0)$). Tham số tốc độ lây lan (β, hoặc lượng tái nhiễm dịch R_0) và hồi phục (γ) là những đại lượng không dễ dàng có được.

Một số phiên bản mở rộng mô hình SIR [24] được đề xuất trong những năm gần đây. Năm 2012, Jin Wang và Shu Liao [96] đề xuất một mô hình dịch tả tổng quát kết hợp mô hình SIR thông thường với một thành phần môi trường thông qua bốn phương trình vi phân:

$$\frac{dS}{dt} = bN - Sf(I,B) - bS$$

$$\frac{dI}{dt} = Sf(I,B) - (\gamma + b)I \qquad \frac{dR}{dt} = \gamma I - bR \quad (1.7)$$

$$\frac{dB}{dt} = h(I,B)$$

trong đó, S, I, R (như trong mô hình SIR) tương ứng chỉ dẫn các ngăn quần thể dễ bị nhiễm; B (thành phần môi trường) biểu thị nồng độ khuẩn tả (*V.vibrios*) trong nước bị ô nhiễm. Tổng dân số của quần thể $N = S + I + R$ được giả thiết không đổi. Tham

số *b* chỉ dẫn tỷ lệ sinh/tử tự nhiên của con người, và γ biểu thị tốc độ hồi phục từ bệnh tả. Trong mô hình tổng quát này, *f(I,B)* là hàm tỷ lệ mắc bệnh xác định tỷ lệ nhiễm mới: hàm này phụ thuộc vào số lượng người nhiễm bệnh *I* và thành phần môi trường *B*. Hàm *h(I, B)* mô tả tỷ lệ thay đổi các tác nhân gây bệnh trong môi trường, hàm này có thể ở dạng tuyến tính hoặc phi tuyến. Đặt *X = [S, I, R, B]T* thì hệ phương trình trên được viết dưới dạng vector là:

$$\frac{d}{dt}X = F(X) \tag{1.8}$$

Để mô hình hóa tổng quát dịch tả , các tác giả thừa nhận thành phần *B* có thể là đại lượng vô hướng hay vector. Mô hình này thừa nhận năm giả thiết sau đây:

1. *f(0,0) = 0;* *h(0,0) = 0* : đảm bảo rằng phương trình (1.8) có nghiệm duy nhất là $X_0 = (N, 0, 0, 0)T$.
2. $f(I, B) \geq 0$: đảm bảo rằng tỷ lệ mắc bệnh không âm.
3. $\frac{\partial f}{\partial I}(I,B) \geq 0$, $\frac{\partial f}{\partial B}(I,B) \geq 0$: đảm bảo rằng số cá thể sẽ nhiễm dịch đơn điệu tăng theo số lượng cá thể đã nhiễm dịch và nồng độ khuẩn tả V.vibrios trong môi trường.
4. $\frac{\partial h}{\partial I}(I,B) \geq 0$: đảm bảo rằng môi trường tăng độ nhiễn dịch khi số lượng cá thể nhiễm dịch tăng.
5. $\frac{\partial h}{\partial B}(I,B) \leq 0$: đảm bảo tỷ lệ tử vong không âm.

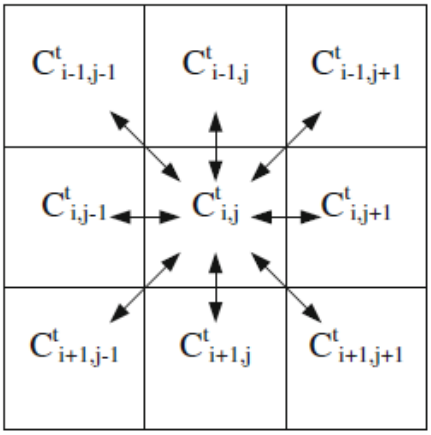
Jin Wang và Shu Liao[96] đã chứng tỏ mô hình được đề xuất là khung chung cho nhiều mô hình dịch tả đã có và như vậy, mỗi mô hình trong tập các mô hình dịch tả được xem xét là một trường hợp riêng của mô hình với việc chọn các tham số cụ thể. Dù mô hình ba ngăn này là nền tảng cho nghiên cứu dịch tễ, nhưng việc xác định các tham số chủ yếu nêu trên không hề dễ dàng và để trả lời các câu hỏi liên quan khác đòi hỏi các mô hình ngẫu nhiên phức tạp hơn. Nhiều mở rộng của mô hình SIR đã được đề xuất tùy theo góc nhìn của nhà nghiên cứu và theo mục tiêu lượng hóa các tham số quan tâm. Đầu tiên là thay đổi cấu trúc dân số bằng cách thêm vào lớp L

chứa các cá thể đang ủ bệnh, hay lớp T chứa các cá thể đang được điều trị... Giải pháp cho mô hình SIR mở rộng cũng như các mô hình dịch tễ học toán học có thể được tìm thấy trong nhiều tài liệu nghiên cứu, chẳng hạn như [13],[24],[35],[37],[49],[66],[83],[92].

1.2.1.2 Dự báo dịch bệnh dựa trên khai phá dữ liệu

Gần đây, các ứng dụng của khai phá dữ liệu đã được chứng minh là mang lại lợi ích cho nhiều lĩnh vực y học bao gồm chẩn đoán, tiên lượng và điều trị. Khai phá dữ liệu y tế có tiềm năng lớn để khám phá các mẫu ẩn trong các tập dữ liệu của ngành y. Những mẫu này có thể được sử dụng để chẩn đoán lâm sàng và dự báo [57]. Khai phá dữ liệu là một kỹ thuật liên quan đến việc trích xuất dự đoán ẩn thông tin từ một cơ sở dữ liệu lớn, nó sử dụng các thuật toán phức tạp cho quá trình phân loại với số lượng bộ dữ liệu và chọn ra thông tin có liên quan. Khai phá dữ liệu là một lĩnh vực trong khoa học máy tính tương đối trẻ và liên ngành, là quá trình trích xuất các mẫu từ các tập dữ liệu lớn bằng cách kết hợp các phương pháp từ thống kê, trí tuệ nhân tạo với quản lý cơ sở dữ liệu[91]. Nghiên cứu sử dụng các mô hình khai phá dữ liệu đã được áp dụng cho các bệnh như tiểu đường, hen suyễn, bệnh tim mạch, AIDS... Các kỹ thuật khai phá dữ liệu như phân lớp, mạng nơron nhân tạo, máy vector hỗ trợ, cây quyết định, hồi quy logistic… đã được sử dụng để phát triển các mô hình trong nghiên cứu y tế [90].

Yujuan Yue và cộng sự [102] đề xuất các mô hình dịch tả theo tác động của các yếu tố khí hậu tại khu vực cửa sông Châu Giang, Trung Quốc. Dữ liệu được lấy tại 24 điểm lấy mẫu (ký hiệu là Z1-Z24) thuộc 4 khu vực được giám sát nằm trong vùng 22-24 vĩ độ Bắc và 112-114 kinh độ Đông. Dữ liệu gồm tỷ lệ dương tính với V.vibrios, nhiệt độ nước, độ pH, nhiệt độ bề mặt đất được Trung tâm giám sát và ngăn ngừa dịch bệnh Trung Quốc cung cấp theo từng điểm lấy mẫu hàng tháng từ tháng 01/2008 tới tháng 12/2009. Dữ liệu về nhiệt độ không khí, lượng mưa, áp suất không khí, độ ẩm, số giờ nắng, tốc độ gió được thu thập hàng ngày từ hai trạm khí tượng Quảng Châu và Thẩm Quyến và sau đó được chuyển thành dữ liệu tháng.



Hình 1.2. Quá trình chuyển đổi tỷ lệ dương tính

Mô hình dịch tả (xem xét quan hệ của tỷ lệ dương tính với V.vibrios) theo mỗi yếu tố khí hậu tại điểm lấy mẫu (i, j) được cụ thể hóa bằng hai phương trình sau đây:

$$C_{i,j}^{t+1} = C_{i,j}^{t} + m[(C_{i-1,j}^{t} - C_{i,j}^{t}) + (C_{i+1,j}^{t} - C_{i,j}^{t}) + (C_{i,j-1}^{t} - C_{i,j}^{t}) + (C_{i,j+1}^{t} - C_{i,j}^{t})] + \quad (1.10)$$

$$md[(C_{i-1,j-1}^{t} - C_{i,j}^{t}) + (C_{i+1,j-1}^{t} - C_{i,j}^{t}) + (C_{i-1,j+1}^{t} - C_{i,j}^{t}) + (C_{i+1,j+1}^{t} - C_{i,j}^{t})]$$

$$C_{i,j}^{'t+1} = t1(T_{I(i,j)}^{t+1} - T_{I(i,j)}^{t}) + C_{i,j}^{t+1} \quad (1.11)$$

Theo Hình 1.2 và diễn giải cụ thể trong hai phương trình (1.10)-(1.11), ô (i, j) đang được quan sát thực hiện việc trao đổi với các ô lân cận của nó theo tám hướng một cách đồng bộ tại mỗi nhịp thời gian. Xem xét mô hình ảnh hưởng của nhiệt độ nước (các yếu tố khí hậu khác là hoàn toàn tương tự) tới dịch tả. Các quy tắc tính toán theo ô được thực thi theo các công thức (1.10) và (1.11), trong đó t_1 là hệ số nhiệt độ nước, m là hệ số chuyển giao V.vibrios giữa các ô lân cận và d là hệ số đường chéo theo sự khác biệt giữa chuyển giao V.vibrios giữa các ô đường chéo và chuyển giao V.vibrios giữa các ô lân cận. $T_{I(i,j)}^{t}$ và $T_{I(i,j)}^{t+1}$ là giá trị nhiệt độ nước tại ô (i, j) tại các nhịp thời gian t và t+1, tương ứng. $C_{I(i,j)}^{t}$ và $C_{I(i,j)}^{t+1}$ là các giá trị nồng độ V.vibrios tại ô (i, j) tại các nhịp thời gian t và t+1, tương ứng. $C_{I(i,j)}^{'t+1}$ là giá trị kết luận về nồng độ V.vibrios tại ô (i, j) vào nhịp thời gian t+1. Giá trị các tham số t, m, d được xác định sử dụng hồi qui tuyến tính. Nghiên cứu 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. Lý do đơn giản khiến cho mô hình có tính khoa học là vi khuẩn khởi đầu cho sự lây nhiễm và sinh sản trong một cá thể đơn lẻ sẽ rất có thể lây lan đến hàng triệu các cá thể khác và tạo thành một đại dịch nếu các điều kiện địa phương và khí hậu thuận lợi cho chúng tồn tại. Các tác giả cũng nhận định rằng mô hình hiện tại còn đơn giản và cần cải thiện hơn nữa.

Martin Mabangiz và cộng sự đã thực hiện nghiên cứu sử dụng kỹ thuật học máy để dự báo dịch tả ở những khu vực khác nhau ở Uganda bằng thuật toán Bayesians dựa trên số liệu dịch tả trong quá khứ [106].

Năm 2012, R. Chunara và cộng sự [79] xây dựng một mô hình hỗ trợ dự báo sớm dịch tả sử dụng dữ liệu từ mạng xã hội Twitter (*http://www.twitter.com*). Các tác giả nhận định rằng phân tích dữ liệu dựa trên dữ liệu báo cáo từ các nguồn y tế công cộng thường bị giới hạn về thời gian và các nguồn dữ liệu khác có thể cung cấp một cơ hội thu thập thông tin sớm về phương thức một dịch bệnh đang diễn ra, và do đó tạo cơ hội cho việc thực hiện các biện pháp can thiệp kịp thời và hiệu quả hơn. Ở đây, các tác giả sử dụng hai nguồn thông tin không chính thức từ HealthMap (http://www.HealthMap.org) và Twitter cùng với nguồn thông tin chính thức từ Bộ Y tế Haiti. Dữ liệu được thu thập trong thời gian 100 ngày, từ 20/10/2010 đến 28/01/2011. Các tác giả tập trung vào các khoảng thời gian bùng phát dịch bệnh, và phát hiện dữ liệu chuỗi thời gian phù hợp với một phân phối mũ. Một công thức đơn giản sau được sử dụng để tính toán số nhiễm bệnh dựa trên mô hình SIR:

$$Re = 1 + rTc \tag{1.12}$$

trong đó, $Tc = 1/b$ (b là tỷ lệ chuyển dịch từ nhiễm bệnh mô hình SIR) và r tốc độ tăng trưởng. Kết quả cho thấy có mối tương quan cao về khối lượng theo thời gian giữa dữ liệu không chính thức và dữ liệu chính thức trong giai đoạn đầu của một ổ dịch hoặc sự kiện có liên quan. Hơn nữa, sự tương quan tốt nhất với độ trễ một ngày chứng tỏ khả năng sử dụng các dữ liệu không chính thức trong việc phát hiện sớm một ổ dịch để đạt được cái nhìn sâu sắc vào việc ước tính số nhiễm bệnh dịch tả trong giai đoạn phát triển ban đầu của dịch bệnh. Điều này càng có ý nghĩa rất quan trọng

để tiến hành các biện pháp kiểm soát dịch bệnh khi mà dữ liệu chính thức được công bố trễ hai tuần trong trường hợp dịch tả Haiti năm 2010. Các tác giả cũng cho rằng mô hình đề xuất có khả năng phù hợp với các bệnh dịch khác. Tuy nhiên, R. Chunara và cộng sự cũng chỉ ra một số hạn chế của phương pháp sử dụng dữ liệu truyền thông xã hội cho dự báo dịch bệnh. Thứ nhất, hạn chế từ trình độ sử dụng truyền thông xã hội thấp kém ở những vùng dịch bệnh và điều này có thể được khắc phục trong tương lai. Thứ hai, hạn chế về nhân khẩu học cung cấp dữ liệu cá nhân trên các truyền thông xã hội (ví dụ như blog, điện thoại di động, v.v.). Thứ ba, một sai lệch tiềm ẩn là thông điệp truyền thông xã hội có thể chứa các sai lệch do dựa trên các cảnh báo sai, tin đồn, hoặc báo cáo sai, đặc biệt là trong các tình huống của sự sợ hãi hoặc hoảng sợ. Cuối cùng, độ tương quan giữa dữ liệu nguồn truyền thông xã hội với báo cáo chính thức vào khoảng thời gian cuối dịch bệnh là rất thấp.

Ứng dụng các kỹ thuật khai phá dữ liệu như Cây quyết định, Naïve Bayes, Mạng nơ-ron, K-means, liên kết phân loại, máy vector hỗ trợ (SVM) và thuật toán MAFIA để dự đoán bệnh tim trên cơ sở phân tích dữ liệu về bệnh tim đã được Ramandeep Kaur và cộng sự thực hiện[53]. Nhóm tác giả đã khẳng định việc sử dụng những kỹ thuật khai phá dữ liệu đã làm giảm đáng kể thời gian xây dựng mô hình và làm cho quá trình dự đoán bệnh tim nhanh hơn đáng kể với độ chính xác cao giúp cải thiện sức khỏe bệnh nhân.

1.2.1.3 Dự báo dịch bệnh với yếu tố không gian

Năm 2008, Osei và Duker đã sử dụng các mô hình hồi qui không gian để khám phá sự phụ thuộc về không gian của tỷ lệ mắc bệnh tả vào một yếu tố môi trường địa phương quan trọng (các bãi rác lộ thiên) ở Kumasi, Ghana [23]. Kết quả nghiên cứu cho thấy những vùng có mật độ cao các bãi rác lộ thiên có tỷ lệ mắc bệnh tả cao hơn những vùng có mật độ thấp các bãi rác lộ thiên. Hơn nữa, những vùng gần bãi rác có tỷ lệ mắc cao hơn những vùng ở xa.

Tương tự, năm 2010, Osei và đồng nghiệp đã sử dụng các mô hình hồi qui không gian để khám phá sự phụ thuộc không gian của bệnh tả vào các thủy vực có tiềm năng bị ô nhiễm [22-23].

Năm 2013, Nkeki và Osirike [70] đã sử dụng hai phương pháp hồi qui trọng số không gian – GWR (G*eographicaly Weighted Regression*) trong GIS và hồi qui tuyến tính (*Ordinary Least Square- OLS*) để phân tích các mối quan hệ giữa sự xuất hiện của dịch tả và các nguồn cấp nước cho các hộ gia đình. Nghiên cứu sử dụng dữ liệu bản đồ các tiểu bang của Nigeria và số liệu thống kê về các trường hợp mắc bệnh tả, nguồn cung cấp nước cho các hộ gia đình và dữ liệu dân số. Kết quả cho thấy dịch tả xảy ra trong khu vực nghiên cứu có liên quan đáng kể đến các nguồn cung cấp nước cho các hộ gia đình và thay đổi theo các khu vực khác nhau.

Nghiên cứu khai phá dữ liệu không gian với các giải thuật Chaid, Quest, C5.0, Neuural Net, so sánh và tìm kiếm giải thuật phù hợp cho mô hình dự báo dịch tả tại Ấn độ đã được Nagabhushara và công sự thực hiện. Trong các thuật toán khai phá này thì CHAID là thuật toán được đánh giá là hiệu quả và phù hợp nhất [81].

Năm 2014, Rasam và cộng sự [107] đã tiến hành nghiên cứu tích hợp GIS và các kỹ thuật phân tích dịch tễ học trong phân tích mô hình không gian của bệnh tả tại huyện Sabah, Malaysia. Kết quả cho thấy bệnh tả có xu hướng tập trung quanh khu vực người bị nhiễm khoảng 1.500 mét. Các ổ dịch tả thường xuất hiện tại các khu vực đông người, môi trường mất vệ sinh, và gần với nguồn nước bị ô nhiễm. Ngoài ra, bệnh tả cũng có quan hệ chặt chẽ với các khu vực ven biển.

Leckebusch and Abdussalam [43] tiến hành nghiên cứu ảnh hưởng của các yếu tố khí tượng và kinh tế xã hội đến sự biến đổi không gian - thời gian của bệnh tả ở Nigeria. Mô hình hồi qui đa biến từng bước (Stepwise multiple regression) và mô hình tổng quát phụ (generalised additive models) được thiết lập cho từng tiểu bang cũng như đối với ba nhóm bang dựa trên lượng mưa hàng năm. Các biến khí tượng khác nhau được phân tích có xem xét đến yếu tố kinh tế - xã hội ẩn chứa khả năng dễ bị tổn thương (ví dụ như tỉ lệ nghèo đói, biết chữ, tiếp cận nguồn nước). Kết quả định lượng cho thấy ảnh hưởng của cả các biến khí hậu và các biến kinh tế - xã hội trong việc giải thích sự thay đổi không gian và thời gian của các ca mắc và tử vong do bệnh tả. Tầm quan trọng của các yếu tố khác nhau được đánh giá cho phép có cái nhìn sâu sắc vào quá trình phát triển dịch bệnh. Ngoài ra, các mô hình kiểm định cho thấy khả

năng dự đoán dịch bệnh, nhờ đó giúp chính quyền đưa ra các biện pháp kiểm soát dịch bệnh kịp thời, hiệu quả.

Ngày nay, việc ứng dụng GIS trong các nghiên cứu ngày càng trở nên phổ biến và mang lại các kết quả gia tăng từ việc phân tích nguồn gốc các yếu tố phát sinh, cơ chế lây truyền và diễn biến dịch trên cả hai phương diện không gian và thời gian mà các phương pháp truyền thống khác khó có thể mang lại được. Ở Việt Nam, có thể nói sản phẩm "Hệ thống thông tin phòng chống thảm họa" của tác giả Nguyễn Hòa Bình (Peacesoft) do Hồi đồng tư vấn chuyên môn Y học thảm họa & Bỏng của Bộ Y tế chủ trì và triển khai ứng dụng tại các sở y tế của 5 tỉnh: Hà Nội, Hà Nam, Nam Định, Ninh Bình và Thái Bình, cùng công trình nghiên cứu "Ứng dụng công nghệ viễn thám và GIS trong dự báo nguy cơ sốt rét tại Bình Thuận năm 2002" của tác giả Nguyễn Ngọc Thạch được coi là những ứng dụng GIS đầu tiên trong y tế[7]. Cho đến nay, một số đơn vị triển khai ứng dụng GIS trong công tác chuyên môn y tế bước đầu thu được kết quả khả quan như đề tài "Ứng dụng GIS trong quản lý và phòng chống HIS/AIDS" của sở Y tế thành phố Hồ Chí Minh[2],[3].

Nhìn chung việc nghiên cứu và ứng dụng GIS trong y tế tại Việt Nam còn hạn chế, chủ yếu tập trung vào khả năng biểu diễn của GIS. Theo khảo sát của nghiên cứu sinh, chưa có các nghiên cứu về mô hình hóa mối quan hệ không gian giữa bệnh dịch và các yếu tố rủi ro trong môi trường sống, cũng như gợi ý các yếu tố nên xét để đưa vào trong mô hình dựa trên GIS.

1.2.2 Một số kỹ thuật xây dựng mô hình dự báo phổ biến

1.2.2.1 Dự báo dựa trên khai phá luật kết hợp

Một trong các hướng tiếp cận hiệu quả trong khai phá dữ liệu (KPDL) là sử dụng luật kết hợp (association rule). Đây là dạng luật biểu diễn tri thức ở dạng khá đơn giản. Phương pháp này nhằm phát hiện ra các luật kết hợp giữa các thành phần trong cơ sở dữ liệu (CSDL). Mẫu đầu ra của giải thuật KPDL là tập luật kết hợp. Luật kết hợp là những luật có dạng như "75% bệnh nhân hút thuốc lá và sống ven vùng ô nhiễm thì bị ung thư phổi, trong đó 25% số bệnh nhân vừa hút thuốc lá, sống ven

vùng ô nhiễm vừa ung thư phổi" [59]. "Hút thuốc lá và sống ven vùng ô nhiễm" ở đây được xem là vế trái (tiền đề - antecedent) của luật, còn "ung thư phổi" là vế phải (kết luận - consequent) của luật. Những con số 25% là độ hỗ trợ của luật (support - số phần trăm các giao dịch chứa cả vế trái lẫn vế phải), còn 75% là độ tin cậy của luật (confidence - số phần trăm các giao dịch thỏa mãn vế trái thì cũng thỏa mãn vế phải).

Lấy I={I_1,I_2,...I_m}, F là tập hợp của m tính chất riêng biệt. Giả sử D là CSDL, với các bản ghi chứa một tập con T các tính chất (có thể coi như T là tập con của I), các bản ghi đều có chỉ số riêng. Một luật kết hợp là một mệnh đề có dạng $X \rightarrow Y$, trong đó X và Y đều là tập con của I, thỏa mãn điều kiện $X \cap Y = \emptyset$. Các tập X và Y được gọi là các tập mục (itemset). Về mặt xác suất, độ tin cậy c của một luật là xác suất (có điều kiện) xảy ra Y với điều kiện đã xảy ra X. Một luật được xem là tin cậy nếu độ tin cậy c của nó lớn hơn hoặc bằng một ngưỡng *minconf* nào đó do người dùng xác định: c ≥ *minconf* [15]. Bài toán khai phá luật kết hợp ở dạng đơn giản nhất được đặt ra như sau:

Hãy tìm kiếm tất cả các luật kết hợp có dạng $X \rightarrow Y$ thỏa mãn độ hỗ trợ $s(X \cup Y) \geq$ *minsup* (*minsup* là giá trị cho trước của người dùng) và độ tin cậy của luật $c(X \rightarrow Y) = s(X \cup Y) / s(X) \geq$ *minconf*. Hầu hết các thuật toán được đề xuất để khai phá luật kết hợp thường chia bài toán này thành hai giai đoạn [16], [44], [64], [68], [104]:

Giai đoạn 1: Tìm tất cả các tập mục phổ biến từ CSDL tức là tìm tất cả các tập mục X thỏa mãn $s(X) \geq$ *minsup*. Đây là giai đoạn có yêu cầu cao về tài nguyên tính toán.

Giai đoạn 2: Sinh các luật tin cậy từ các tập phổ biến đã tìm thấy ở giai đoạn thứ nhất. Giai đoạn này tương đối đơn giản và yêu cầu tài nguyên tính toán thấp hơn so với giai đoạn trên.

Độ hỗ trợ (*Support*), độ tin cậy (*Confidence*) và độ chắc chắn thống kê (*Lift*) là các độ đo dùng để đo lường luật kết hợp. Độ hỗ trợ của luật kết hợp $X \rightarrow Y$ là xác suất xuất hiện tất cả các đối tượng trong cả hai tập X và Y. Công thức để tính độ hỗ trợ của luật $X \rightarrow Y$, ký hiệu Supp($X \rightarrow Y$) như sau:

$$\text{Supp}\ (X \rightarrow Y) = P\ (X \cup Y) = \frac{n(X \cup Y)}{N} \quad (1.13)$$

trong đó *N* là tổng số sự kiện, n(*X* ∪ *Y*) là số sự kiện chứa cả *X* và *Y*.

Độ tin cậy của luật kết hợp *X* →*Y*, ký hiệu Conf(*X*→*Y*) là xác suất xảy ra *Y* khi đã biết *X*. Công thức để tính độ tin cậy của luật kết hợp $X \rightarrow Y$ là xác suất có điều kiện *Y* khi đã biết *X* như sau:

$$\text{Conf}\ (X \rightarrow Y) = P\ (Y \mid X) = \frac{n(X \cup Y)}{n(X)} \quad (1.14)$$

trong đó n(*X*) là số sự kiện chứa *X*.

Độ chắc chắn thống kê của luật kết hợp *X*→*Y*, ký hiệu Lift(*X*→*Y*), được định nghĩa là:

$$\text{Lift}(X \rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}\ (Y)} \quad (1.15)$$

trong đó supp(*X*) là độ hỗ trợ của tập đối tượng *X*, được định nghĩa là tỷ lệ các sự kiện chứa các đối tượng trong *X* trên tổng số sự kiện.Tương tự, supp(*Y*) là độ hỗ trợ của tập đối tượng *Y*, được định nghĩa là tỷ lệ các sự kiện chứa các đối tượng trong *Y* trên tổng số sự kiện. Giá trị của Lift(*X*→*Y*) càng lớn, ý nghĩa thống kê của luật càng cao.

1.2.2.2 Dự báo bằng kỹ thuật học máy hồi qui và phân lớp

Học máy (Machine Learning) là một lĩnh vực khoa học nghiên cứu các thuật toán cho phép máy tính có thể học được các khái niệm. Hai kỹ thuật điển hình trong học máy ứng dụng trong dự báo là hồi qui và phân lớp. Hồi qui tương ứng với miền giá trị của biến đầu ra liên tục còn phân lớp tương ứng với miền giá trị của biến đầu ra rời rạc. Bài toán xây dựng mô hình dự báo được hình thức hóa như sau: Gọi *D* là tập tất cả các điểm dữ liệu có thể có trong miền ứng dụng liên quan tới công việc dự báo đang được quan tâm, *D = {điểm dữ liệu d}*. Thông thường, mỗi điểm dữ liệu d bao gồm *n+1* thành phần, trong đó thành phần *n+1* (ký hiệu là *y*) là một thành phần đặc biệt mà giá trị của nó cần được dự báo và được gọi là biến đầu ra (biến mục tiêu) và tập *n* thành phần còn lại (ký hiệu là các biến đầu vào $x_1, x_2, .., x_n$) được gọi là biến đầu vào. Ta có $d=(x_1, x_2, .., x_n, y)$. Gọi *X* là không gian các biến đầu vào tương ứng với *n* thành phần đầu vào và *Y* là không gian biến đầu ra. Như vậy, $D \subseteq X \times Y$ (tích đề

các của *X* và *Y*). Gọi $D_{example}$ là tập các dữ liệu đã thu thập được. $D_{example}$ được gọi là *tập dữ liệu ví dụ* (example set) và nó là tài nguyên cơ bản để xây dựng mô hình dự báo.

Bài toán xây dựng mô hình được phát biểu như sau "Cho trước tập dữ liệu ví dụ $D_{example}$, hãy tìm một ánh xạ f: $X \rightarrow Y$ sao cho ánh xạ f *phù hợp với tập dữ liệu ví dụ* $D_{example}$". Bài toán xây dựng mô hình được gọi là *bài toán hồi qui* (regression) khi tập giá trị *Y* của biến mục tiêu là liên tục và được gọi là *bài toán phân lớp* (classification) khi tập giá trị *Y* của biến mục tiêu là hữu hạn. Ánh xạ kết quả tìm được f chính là mô hình dự báo, theo đó khi cho biết giá trị các biến đầu vào thì f sẽ chỉ ra được giá trị cần dự báo của biến đầu ra.

Một số các kỹ thuật học máy được áp dụng phổ biến như hồi qui tuyến tính, hồi qui và phân lớp rừng ngẫu nhiên, máy vector hỗ trợ, Naïve Bayes,... sẽ được mô tả ngắn gọn trong phần tiếp theo.

Hồi qui tuyến tính (Linear Regression –LM): Các phương pháp dự báo đều xem xét sự biến động của đại lượng cần dự báo theo thời gian thông qua số liệu thống kê được trong quá khứ. Tuy nhiên, trong thực tế đại lượng cần dự báo còn có thể bị tác động bởi các nhân tố khác. Đại lượng cần dự báo là biến phụ thuộc còn nhân tố tác động lên nó là biến độc lập. Biến độc lập có thể gồm một hoặc nhiều biến. Mô hình hồi quy tương quan được sử dụng phổ biến nhất trong dự báo là mô hình hồi quy tương quan tuyến tính. Đại lượng dự báo được xác định theo công thức sau:

$$Y_t = a+bX \qquad (1.16)$$

Trong đó:

Y_t - mức nhu cầu dự báo cho thời điểm t

X - Biến độc lập (nhân tố ảnh hưởng đến đại lượng dự báo)

a, b - Các hệ số (a hệ số chặn , b - độ dốc)

Để đánh giá mối liên hệ giữa hai biến số trong mô hình hồi qui tương quan cần tính "Hệ số tương quan". Hệ số này biểu hiện mức độ hoặc cường độ của mối quan hệ tuyến tính. Hệ số tương quan nhận giá trị giữa -1 và 1.

Tuỳ theo các giá trị của hệ số tương quan, mối quan hệ giữa hai biến X và Y có thể gồm các khả năng như sau:

- Khi hệ số tương quan = ±1, giữa x và y có quan hệ chặt chẽ

- Khi hệ số tương quan = 0, giữa x và y không có liên hệ gì

- Khi hệ số tương quan càng gần ±1, mối liên hệ tương quan giữa x và y càng chặt chẽ

- Khi hệ số tương quan mang dấu dương ta có tương quan thuận, ngược lại mang dấu âm ta có tương quan nghịch.

Cây quyết định (Decission Trees- DT): Cây quyết định là một đồ thị của các quyết định và các hậu quả có thể của nó. Cây quyết định được sử dụng để xây dựng một kế hoạch nhằm đạt được mục tiêu mong muốn. Trong lĩnh vực học máy, cây quyết định là một kiểu mô hình dự báo, nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật/hiện tượng. Mỗi một nút trong tương ứng với một biến; đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó. Cây quyết định là mô hình học máy tự động được sử dụng rất nhiều trong khai phá dữ liệu do tính đơn giản mà hiệu quả [56], [99],[34].

Algorithm 1: Decission Tree

```
1. node LearnTree(examples, targetAttribute, attributes)
2. examples is the training set
3. targetAttribute is what to learn
4. attributes is the set of available attributes
5. returns a tree node
6. begin
7. if all the examples have the same targetAttribute value,
      a. return a leaf with that value
8. else if the set of attributes is empty
      a. return a leaf with the most common targetAttribute
         value among examples
9. else begin
```

```
        a. A = the "best" attribute among attributes having a
           range of values v1, v2, …, vk
        b. Partition examples according to their value for A
           into sets S1, S2, …, Sk
        c. Create a decision node N with attribute A
        d. for i = 1 to k
             i. begin
                  1. Attach a branch B to node N with test Vi
                  2. if Si has elements (is non-empty)
                       a. Attach B to LearnTree(Si,
                          targetAttribute, attributes –
                          {A});
                  3. Else
                       a. Attach B to a leaf node with most
                          common targetAttribute
            ii. end
        e. return decision node N
     10. end
     11. End
```

Rừng ngẫu nhiên (Random Forests- RF): giải thuật rừng ngẫu nhiên là thành viên trong chuỗi thuật toán cây quyết định. Ý tưởng của Random Forest là tạo ra vô số cây quyết định với các câu hỏi cho từng thuộc tính. Để tạo mới cây quyết định, thuật toán Random Forest luôn luôn bắt đầu với một cây quyết định rỗng. Đó là cây quyết định chỉ có điểm bắt đầu và liên kết thẳng tới câu trả lời. Mỗi khi thuật toán tìm được một câu hỏi tốt để hỏi, nó sẽ tạo ra 2 nhánh (trái và phải) của cây. Khi không còn câu hỏi nào nữa, thuật toán sẽ dừng lại và kết thúc quá trình xây dựng cây quyết định. Để tìm được câu hỏi đầu tiên tốt nhất, thuật toán sẽ cố gắng thử hết tất cả các câu hỏi có thể. Sau đó ứng với mỗi câu hỏi, thuật toán sẽ xác minh câu hỏi này có dùng được để phân loại cho các đối tượng cần theo dõi không? Câu hỏi được chọn không cần thiết là hoàn hảo, nhưng nó nên tốt hơn các câu khác[89].

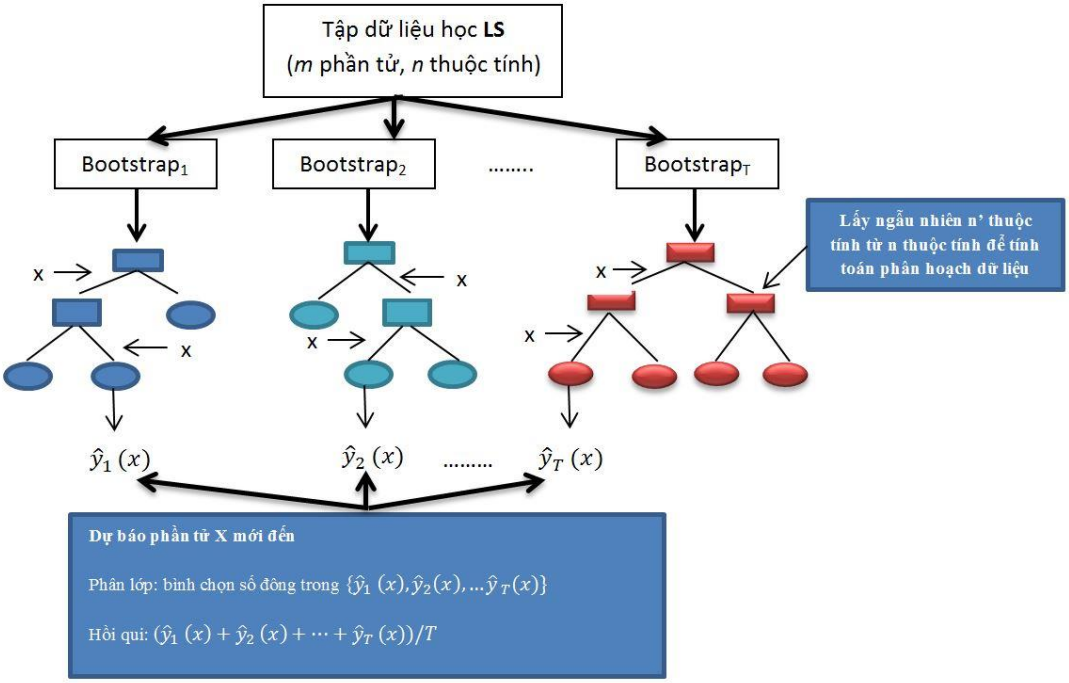
Thông thường để xác định thế nào là câu hỏi tốt, các thuật toán sẽ tính toán "information gain" – đây là cách để chấm điểm từng câu hỏi. Và câu hỏi nào có "information gain" cao nhất, sẽ là câu hỏi tốt nhất. Điều đặc biệt của Random Forest là việc tạo ra mỗi cây quyết định có thể bỏ phiếu độc lập. Khi kết thúc việc bỏ

phiếu, câu trả lời có lượng bỏ phiếu cao nhất, sẽ được chọn bởi Random Forest [89]. Tuy nhiên tồn tại vấn đề là: nếu tất cả các cây quyết định đều được sử dụng cùng một cách, chúng sẽ giống nhau. Để chắc chắn rằng tất cả các cây quyết định là không giống nhau, Random Forest sẽ tự động thay đổi ngẫu nhiên đối tượng cần theo dõi. Nói một cách chính xác hơn, thuật toán sẽ xóa ngẫu nhiên 1 vài đối tượng, và nhân bản 1 vài đối tượng khác. Tiến trình này được gọi là "**bootstrapping**". Ngoài ra để đảm bảo rằng các cây quyết định có sự khác biệt, Random Forest sẽ ngẫu nhiên loại bỏ có mục đích một vài câu hỏi khi xây dựng cây quyết định. Trong trường hợp này, nếu câu hỏi tốt nhất không được kiểm tra, thì các câu hỏi khác sẽ được chọn để tạo ra cây- Tiến trình này được gọi là "**attribute sampling**".

Algorithm 2: Random Forest [89]

Precondition: A training set $S := (x_1, y_1), \ldots, (x_n, y_n)$, features F, and number of trees in forest B.

```
1. function RandomForest(S , F)
2. H ← ∅
3.  for i ∈ 1,... , B do
4.    S (i) ← A bootstrap sample from S
5.    hi ← RandomizedTreeLearn(S(i),F)
6.    H ← H ∪ {hi}
7.  end for
8. return H
9. end function
10. function RandomizedTreeLearn(S,F)
11.   At each node:
12.   f ← very small subset of F
13.   Split on best feature in f
14.  return The learned tree
15. end function
```



Hình 1.3: Giải thuật rừng ngẫu nhiên.

Giải thuật rừng ngẫu nhiên xây dựng cây không cắt nhánh nhằm giữ cho thành phần lỗi *bias* thấp và dùng tính ngẫu nhiên để điều khiển tính tương quan thấp giữa các cây trong rừng. Tiếp cận rừng ngẫu nhiên có độ chính xác cao, học nhanh, chịu nhiễu tốt và không bị tình trạng học vẹt và đáp ứng được yêu cầu thực tiễn cho vấn đề phân loại, hồi quy [25].

Máy vector hỗ trợ (Support Vector Machines - SVM): Đây là một phương pháp học máy có giám sát nhằm thực hiện phân loại và phân tích hồi quy. Phương pháp này được coi là một phương pháp mạnh và chính xác trong các phương pháp phân loại dữ liệu. Máy vector hỗ trợ (SVM) là mô hình hiệu quả và phổ biến cho vấn đề phân lớp, hồi qui cho những tập dữ liệu có số chiều lớn. Ý tưởng chính của SVM: Là chuyển tập mẫu từ không gian biểu diễn Rn của chúng sang một không gian Rd có số chiều lớn hơn. Trong không gian Rd, tìm một siêu phẳng tối ưu để phân hoạch tập mẫu này dựa trên phân lớp của chúng, cũng có nghĩa là tìm ra miền phân bố của từng lớp trong không gian Rn để từ đó xác định được phân lớp của 1 mẫu cần nhận dạng.

Ta có thể hiểu, siêu phẳng là một mặt hình học f(x) trong không gian N chiều, với $x \in RN$ [42].

Naïve Bayes: Thuật toán Bayes là một trong những thuật toán phân lớp điển hình trong học máy và khai phá dữ liệu.Ý tưởng chính của thuật toán là tính xác suất hậu nghiệm của sự kiện *c* xuất hiện sau khi sự kiện *x* đã có trong không gian ngữ cảnh t thông qua tổng hợp các xác suất tiên nghiệm của sự kiện *c* xuất hiện khi sự kiện *x* đã có trong tất cả các điều kiện T thuộc không gian t:

$$p(c|x,t) = \sum p(c|x,T)p(T|x)(\text{với T trong t}) \quad (1.17)$$

Gọi X=$\{x_{1,}x_{2,}, \dots x_{n,}\}$ là một mẫu, các thành phần của nó biểu diễn các giá trị được tạo ra trên một tập *n* thuộc tính. Theo phương pháp Bayesian, *X* được xem là "bằng chứng" hay "dấu hiệu". *H* là một giả thuyết nào đó, chẳng hạn như dữ liệu *X* thuộc một lớp cụ thể *C*. Với các bài toán phân lớp, mục tiêu là xác định P(*H*|*X*), xác suất mà giả định *H* xảy ra với các dấu hiệu cho trước. Nói một cách khác, chúng ta đi tìm xác suất để mẫu *X* thuộc về lớp *C* khi đã biết được các thuộc tính mô tả mẫu *X*. Theo định lý Bayes, xác suất mà chúng ta muốn tính P(*H*|*X*) có thể được biểu diễn qua các xác suất P(*H*), P(*X*|*H*) và P(*X*) như sau:

$$P = \frac{P(X|H)P(H)}{P(X)} \quad (1.18)$$

Và các xác suất này có thể được thiết lập từ tập dữ liệu cho trước [76].

1.2.2.3 Dự báo bằng phân tích 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. Phần tiếp theo sẽ trình bày vắn tắt các kỹ thuật này.

Nội suy không gian: Nội suy không gian là quá trình tính toán giá trị của các điểm chưa biết từ điểm đã biết trên miền bao đóng của tập giá trị đã biết bằng một phương pháp hay hàm toán học nào đó. Hiện nay, có nhiều thuật toán nội suy khác nhau như: nội suy điểm, nội suy bề mặt, nội suy toàn diện, nội suy địa phương, nội suy chính xác, nội suy gần đúng. Trong luận án sử dụng phương pháp nội suy thông dụng trong công cụ ArcGIS đó là IDW. Phương pháp nội suy IDW (Inverse Distance Weight) xác định giá trị của các điểm chưa biết bằng cách tính trung bình trọng số khoảng cách các giá trị của các điểm đã biết giá trị trong vùng lân cận của mỗi pixel. Những điểm càng cách xa điểm cần tính giá trị càng ít ảnh hưởng đến giá trị tính toán. Công thức nội suy IDW như sau:

$$\boldsymbol{z = \frac{\sum_{i=1}^{n}(w_i * z_i)}{\sum_{i=1}^{n} w_i}} \qquad (1.19)$$

với $\boldsymbol{w = \frac{1}{d^k}}$

Trong đó, z là giá trị chưa biết tại điểm cần nội suy, i là số thứ tự điểm được sử dụng để nội suy (i = 1, 2,..., n), n là tổng số điểm được sử dụng để nội suy, w_i là trọng số nghịch khoảng cách, z_i là giá trị đã biết tại điểm i, d là khoảng cách từ điểm cần nội suy đến điểm i, k là hằng số ảnh hưởng khoảng cách (thông thường k = 2). Phương pháp này được nhận định là nhanh và dễ thực hiện[29],[40].

Phân tích điểm nóng: Đây là một phương pháp phân nhóm không gian. Nó tính toán thống kê Getis-Ord Gi* **[12]**, **[72]** cho mỗi đối tượng trong tập dữ liệu GIS và cho biết mức độ phân nhóm giá trị cao hay thấp về mặt không gian. Phương pháp này tính toán bằng cách xem xét từng đối tượng trong quan hệ với các đối tượng lân cận. Một đối tượng có giá trị cao chưa hẳn là một điểm nóng có ý nghĩa về mặt thống kê. Để trở thành một điểm nóng về mặt thống kê, một đối tượng cần có giá trị cao và được bao quanh bởi các đối tượng khác cũng có giá trị cao. Thống kê Getis-Ord Gi* được biểu diễn theo công thức như sau [12]**:**

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}} \quad (1.20)$$

Trong đó, x_j là giá trị của đối tượng j; $w_{i,j}$ là trọng số không gian giữa đối tượng i và j; n là tổng số đối tượng; và

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n}$$
$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2} \quad (1.21)$$

Hồi qui ước lượng bình phương nhỏ nhất - OLS (Ordinary Least Square): là phương pháp mô tả và đánh giá mối quan hệ giữa một biến (gọi là biến phụ thuộc - ký hiệu là *Y*) với một hay nhiều biến khác (gọi là biến độc lập). Trong mô hình, chúng ta coi biến độc lập và biến phụ thuộc là hoàn toàn khác nhau. Biến *Y* được giả thiết là có tính ngẫu nhiên, còn biến *X* được giả thiết là cố định (nhận giá trị cố định). Mô hình hồi qui cho phép ước lượng và suy diễn thống kê các tham số tổng thể. Dạng tổng quát của mô hình hồi qui tuyến tính đơn giản là:

$Y = \alpha + \beta x + u$ (1.22)

Trong đó: Y là biến phụ thuộc -

x là biến độc lập

α là hệ số chặn

β là độ dốc

u là sai số của đường hồi qui tổng thể

Phương pháp hồi qui (*OLS*) được dùng để ước lượng các tham số tổng thể trên cơ sở một mẫu số liệu. Gọi $\{(x_i,y_i): i=1;....;n\}$ là một mẫu ngẫu nhiên, có cỡ là *n* mà ta thu được từ tổng thể. Với mỗi quan sát trong mẫu này, ta sẽ có $Y_i = \alpha + \beta x_i$ (1.23)

Để ước lượng với phương pháp bình phương cực tiểu, giả thuyết chính trong phương pháp này là u và x hoàn toàn không có quan hệ với nhau, nghĩa là $E(u|x) =$

E(u) = 0. Cần tìm đường phù hợp nhất thông qua xây dựng bài toán cực tiểu nghĩa là tìm các tham số sao cho biểu thức dưới đây đạt giá trị cực tiểu:

$$\sum_{i=1}^{n}(\widehat{U}_i)^2 = \sum_{i-1}^{n}(y_i - (\hat{\alpha} + \hat{\beta}x_i))^2 \quad (1.24.)$$

$$\frac{\partial L}{\partial \hat{\alpha}} = -2\sum_t (y_t - \hat{\alpha} - \hat{\beta}x_t) = 0$$

$$\frac{\partial L}{\partial \hat{\alpha}} = -2\sum_t x_t(y_t - \hat{\alpha} - \hat{\beta}x_t) = 0$$

Sử dụng đạo hàm để giải bài toán cực tiểu này, lấy đạo hàm bậc 1 theo α và β và giải phương trình. Qua đó có thể ước lượng được tham số của mô hình hồi qui.

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2} \quad (1.25)$$

S_{XY} = đồng phương sai của (X, Y)

S_X^2 = phương sai của (X)

Về mặt trực giác, OLS là việc ước lượng đường thẳng qua các điểm số liệu trong mẫu sao cho tổng khoảng cách bình phương sai số là nhỏ nhất, nên có tên là bình phương cực tiểu.

Hồi qui trọng số không gian GWR (Geographically Weighted Regression) [27]: Phương pháp này xem xét tính không đồng nhất của các mối quan hệ theo không gian. Nói cách khác, nó mô hình hóa các mối quan hệ thay đổi theo các vị trí không gian khác nhau thông qua ma trận quyền số không gian. Mô hình dựa vào kỹ thuật hồi qui quyền số địa lý hay còn gọi là phân tích hồi qui theo vùng. Sử dụng một khung phân tích hồi qui cửa sổ chuyển động (moving window regression framework). Các quan sát giữa các cửa sổ hồi qui vùng được tính theo khoảng cách tới các điểm hồi qui. Các quan sát gần với điểm hồi qui x_i sẽ có trọng số cao hơn so với các quan sát ở xa hơn. Cửa sổ hồi qui quyền số này sẽ được dịch chuyển đến các điểm hồi qui tiếp

theo cho đến khi tất cả các điểm hồi qui được nằm trong đó. Trong mô hình này dựa trên khung hồi qui cổ truyền, nên kỹ thuật sẽ tạo ra kết quả hồi qui chuẩn cho từng điểm hồi qui. Điều này cho phép kết quả hồi qui có thể hiện thị trên bản đồ cho thấy sự khác nhau giữa các vùng, kỹ thuật này đặc biệt hữu ích đối với việc phân tích quan hệ giữa các dữ liệu về không gian.

Mô hình hồi qui trọng số được viết dưới dạng sau:

$$y_i = a_0 + \sum_j x_{ij}\, a_{ij} + \varepsilon \tag{1.26}$$

trong đó y là biến phụ thuộc, x là biến độc lập, a là hệ số hồi quy, i là chỉ số cho các vị trí (location), j là chỉ số cho biến độc lập, và ε là sai số cho mỗi hồi qui tại các điểm i, Cho mỗi hồi qui tại các điểm i, quyền số của các quan sát được lựa chọn phục thuộc vào khoảng cách tới điểm hồi qui j. Hàm phân tách khoảng cách Gauss áp dụng trong phân tích này có thể viết như sau:

$$w_{ij} = exp\left[-\frac{1}{2}\left(d_{ij}|b\right)^2\right] \tag{1.27}$$

Trong đó w là quyền số,

d là khoảng cách từ các điểm hồi qui,

b là phạm vi hay bán kinh của sự ảnh hưởng xung quanh mỗi quan sát.

1.2.3 Nhận xét về các mô hình dự báo dịch bệnh hiện có

Từng mô hình trong các mô hình dự báo dịch bệnh đề cập trong Mục 1.2.2 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. Lớp đầu tiên bao gồm các mô hình dịch tễ học toán học mà điển hình là các mô hình SIR và biến thể của mô hình dự báo SIR-SIS. Lớp thứ hai bao gồm các mô hình học máy mà điển hình là các mô hình hồi quy, mô hình (tự) hồi quy, mô hình hồi qui không gian và các mô hình khai phá dữ liệu truyền thông xã hội. Lớp cuối cùng gồm các mô hình còn lại.

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

Mô hình SIR hay mô hình lan truyền theo ngăn chuẩn trong dịch tễ học toán học được dùng để phân chia một dân số (hay một phần/vùng dân số) thành ba thành phần dân số con (3 ngăn) cho một căn bệnh truyền nhiễm trong một tổng dân số. Nhiều chi tiết quan trọng của sự tiến triển lây nhiễm sẽ bị lược bỏ. Để xây dựng mô hình SIR, thực hiện mô hình hóa sơ lược sự biến động giữa các ngăn thông qua các phương trình vi phân và tích phân. Dù mô hình ba ngăn này là nền tảng cơ bản cho nghiên cứu dịch tễ, nhưng việc xác định các tham số chủ yếu không hề dễ dàng và để trả lời các câu hỏi liên quan khác đòi hỏi các mô hình ngẫu nhiên phức tạp hơn và có

hiệu quả hơn khi áp dụng tin học và phân tích để xử lý một khối lượng rất lớn dữ liệu y tế. Vì lý do trong thực tế nếu xem xét mỗi tỉnh có một mô hình SIR riêng biệt thì Việt Nam sẽ có 63 mô hình SIR. Mô hình nào có thể diễn đạt hay theo dõi được sự lây nhiễm động của quá trình khi giữa các lớp S_i có sự tương tác với nhau tương tự cho các lớp I_i. Như vậy, yêu cầu cần có những mô hình khác để xem xét vì ít nhất hai lý do sau:

- Các khối dân số lớn chứa một cấu trúc có tính đa tầng, các tầng này tương tác với nhau;
- Sự di chuyển của các cá thể giữa các tầng (tỉnh/thành phố) là tiềm năng cho sự lan truyền dịch bệnh.

Yếu tố không gian là một trong những yếu tố chính của dịch tễ học. Dữ liệu không gian thường có hai tính chất là:

- Tự tương quan về không gian (những đối tượng gần nhau có xu hướng tương tự nhau hơn những đối tượng ở xa) và
- Không dừng về không gian (biến đổi theo vùng).

Các tính chất này có thể làm cho các ước lượng mô hình hồi qui truyền thống không hiệu quả. Vì vậy, cần các nghiên cứu các phương pháp đưa yếu tố không gian vào để phù hợp với đặc tính thực tế này hoặc kết hợp những tính chất đặc biệt này của dữ liệu không gian để cải thiện khả năng mô hình hóa các mối quan hệ dữ liệu. Một số phương pháp hồi qui không gian xử lý hiệu quả vấn đề tự tương quan không gian; một số khác lại xử lý hiệu quả tính không dừng về không gian. Hiện tại, chưa có phương pháp hồi qui không gian nào xử lý hiệu quả cả hai vấn đề trên[40].

Bên cạnh đó, cùng với sự phát triển của khoa học, việc thu thập và lưu trữ dữ liệu ngày càng thuận tiện hơn. Chúng ta ngày càng hiểu biết nhiều hơn về sự thay hình, đổi dạng của virus, sự hòa trộn nhân chủng học, môi trường, khí hậu và mạng lưới tương tác phức tạp của con người có ảnh hưởng ra sao đến sự lan truyền dịch bệnh. Xu hướng nghiên cứu đã dịch chuyển từ hướng nghiên cứu trên tập mẫu dữ liệu đại diện để dự báo sang việc phân tích dữ liệu lớn để tìm ra qui luật phục vụ dự báo. Trên những tập dữ liệu thu thập được đòi hỏi có sự kết hợp và đa dạng hóa các

kiểu dữ liệu, các phương pháp phân tích. Kết quả của các phân tích này sẽ làm cơ sở cho việc quyết định phương pháp mô hình hóa phù hợp trong các mối quan hệ giữa bệnh tật và các yếu tố rủi ro từ môi trường cũng như gợi ý các yếu tố nên xem xét đưa vào mô hình. Ở Việt Nam, đã có một số nghiên cứu đưa ra mô hình dự báo tỷ lệ mắc bệnh dựa trên cơ sở phân tích mối quan hệ giữa dịch bệnh và các yếu tố động lực/rủi ro từ môi trường [1],[11]. Tuy nhiên các nghiên cứu này đều chưa phân tích mô hình phân bố dịch bệnh theo không gian. Hay nói cách khác, các nghiên cứu mới chủ yếu tập trung vào chiều thời gian mà chưa quan tâm đến dữ liệu không gian. Do đó, nghiên cứu trong luận án này tập trung vào việc tìm kiếm giải pháp mô hình hóa dự báo dịch bệnh có sử dụng dữ liệu không gian bằng các kỹ thuật khai phá và học máy.

1.3 Dịch tả và nhu cầu dự báo dịch tả

Theo R.R Colwell [31] thuật ngữ bệnh tả ("cholera") có nguồn gốc từ tiếng Hy Lạp, theo đó "cholera" là từ ghép của "chole" ("mật") và "rein" ("dòng chảy") có nghĩa là dòng chảy mật, hoặc là "máng xối của mái nhà". Hiện nay, bệnh tả vẫn là một mối đe dọa lớn ở quy mô toàn cầu. Bệnh tả có thể gây ra tình trạng mất nước nghiêm trọng và dẫn đến tử vong nếu không được điều trị đúng cách thông qua bù nước. Năm 1883, Rocbert Kock đã phân lập được vi khuẩn tả từ phân người bệnh và từ niêm mạc ruột của những xác chết vì bệnh tả.Vi khuẩn tả Vibrio cholerae (V.vibrios) thuộc giống Vibrio, chúng có thể tồn tại lâu ngày trong phân, đất ẩm, nước và thực phẩm. Dịch tả là một trong những bệnh truyền nhiễm tạo nên nhiều đại dịch lớn nhất đe dọa loài người[28], [48], [55],[69].

Theo Tổ chức Y tế Thế giới, bệnh tả thường lây truyền qua môi trường nước hoặc thức ăn bị lây nhiễm phân và vẫn duy trì như một nguy cơ có thể xuất hiện bất cứ lúc nào tại các quốc gia. Các vụ bùng phát mới có thể xuất hiện không thường xuyên tại bất cứ vùng nào của thế giới như nơi nguồn cấp nước, tình trạng vệ sinh an toàn thực phẩm không được đảm bảo. Nguy cơ lớn nhất xuất hiện tại các cộng đồng dân cư đông đúc với các đặc điểm điều kiện vệ sinh nghèo nàn, nguồn nước uống không hợp vệ sinh và tỷ lệ lây lan giữa người với người gia tăng. Vì thời gian ủ bệnh

là rất ngắn (chỉ từ 2 giờ đến 5 ngày) nên số lượng các trường hợp tăng lên rất nhanh. Việc ngăn chặn bệnh tả không cho thâm nhập vào một khu vực là không thể - song tốc độ lan truyền của căn bệnh trong một phạm vi là có thể kiểm soát được thông qua việc phát hiện và khẳng định sớm về các trường hợp mắc bệnh. Vì bệnh tả có thể là một vấn đề khẩn cấp đối với sức khỏe cộng đồng- với tỷ lệ tử vong cao, khả năng lây truyền nhanh chóng và có thể lan tràn trên khắp thế giới, ảnh hưởng nghiêm trọng tới du lịch và thương mại – do đó việc dự báo sớm, thích ứng kịp thời và hiệu quả là vô cùng quan trọng[83]. 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 [26],[28],[31], [38],[102]. Ali và cộng sự [58] đã phân tích dữ liệu ca bệnh Tả tại Matlab, Bangladesh từ năm 1988 đến năm 2001 và rút ra kết luận: Số ca dịch tả tại Matlab chịu ảnh hưởng mạnh của nhiệt độ tại thành phố và nhiệt độ bề mặt nước biển. Nghiên cứu này dự báo số ca mắc tả trên toàn vùng dựa trên phương pháp phân tích chuỗi thời gian.

R. C. Reiner và cộng sự [82] đã xây dựng mô hình dự báo số ca mắc tả trước 11 tháng tại Matlab, Bangladesh. Dữ liệu được sử dụng trong nghiên cứu này là các tham số khí tượng, chỉ số dao động phía Nam (SOI) và số ca mắc tả của Matlab từ năm 1995 đến năm 2008. Chỉ số dao động phía Nam và tình trạng ngập lụt ở Matlab là các yếu tố khí hậu cục bộ có ảnh hưởng lớn nhất đến số ca mắc tả. Ngoài ra, nghiên cứu này đã dự báo số ca theo đơn vị *thanas* và có một kết luận quan trọng là các *thanas* tại trung tâm Matlab có vai trò trong việc lây lan bệnh ra toàn thành phố. Kỹ

thuật xây dựng mô hình dự báo được sử dụng trong nghiên cứu này là mô phỏng bằng mô hình Markov đa chiều không đồng nhất (Multi Dimensional Inhomogeneous Markov Chain – MDIMC).

Xu Min và cộng sự [67] sử dụng mô hình MaxEnt – một mô hình dựa trên mô hình kỳ vọng cực đại – để phân tích ảnh hưởng của khí hậu đến bệnh tả ở Trung Quốc từ năm 2001-2008. Theo kết quả của nghiên cứu này, lượng mưa, nhiệt độ và độ cao so với mặt biển có ảnh hưởng mạnh nhất tới số ca bệnh tả. Khoảng cách tới bờ biển, độ ẩm tương đối và khí áp cũng có ảnh hưởng. Tuy nhiên số giờ nắng và quá trình giảm mức nước sông hầu như không có ảnh hưởng đến số ca bệnh.

Nguyên cứu phương pháp để lấy dữ liệu từ các nguồn khác nhau và áp dụng các kỹ thuật học máy để dự đoán nguy cơ bùng phát dịch tả theo thời gian ở các khu vực khác nhau ở Uganda của Martin [106], phân tích các khu vực có động lực tương tự về tỷ lệ dịch tả theo thời gian. Sau đó xây dựng một mô hình xác suất để dự đoán các trường hợp bệnh tả trong tương lai.

M.Nagabhushana Rao cùng cộng sự đã tiến hành nghiên cứu sử dụng công cụ và thuật toán khai phá dữ liệu để dự báo dịch tả tại Ấn độ. Nghiên cứu được thực hiện trên nhân khẩu học dữ liệu về sức khỏe. Bằng cách áp dụng quy tắc sắp xếp thứ tự, các khu vực bị ảnh hưởng của dịch tả rồi tiến hành phân tích thông qua công cụ khai thác dữ liệu để thống kê, tính toán. Mô hình được thiết kế sử dụng các thuật toán CHAID, C5.0, NeuralNet & QUEST. Trong số đó, thuật toán CHAID được chứng minh là hiệu quả hơn trong việc dự đoán dịch tả [81]. Ngoài ra, còn có một số công trình nghiên cứu dự báo khác như Prieto VM và cộng sự [95], José Carlos Santos và Sérgio Matos [86], Yusheng Xie và cộng sự [100],.

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, trong đó trường hợp mắc bệnh tả đầu tiên được báo cáo vào ngày 23/11/2007 tại Hà Nội. Đến ngày 11/4/2008, tổng số ca mắc tả tích lũy là 3.271 được báo cáo từ 18 tỉnh phía Bắc, trong đó Hà Nội chiếm đa phần người nhiễm bệnh. Sự bùng phát mạnh của dịch tả ở miền Bắc và đặc biệt là ở Hà Nội đã thúc đẩy việc nghiên cứu về

bệnh Tả tại Việt nam [73]. Một số nghiên cứu về dịch tả vào các năm 2007-2008 tại Việt Nam đã được công bố [71],[19],[36]. Tuy nhiên những công bố này chưa đề cập tới các yếu tố biến đổi khí hậu tác động tới dịch tả , cũng như chưa đề cập tới mô hình hóa dự báo dịch tả. Tại Việt Nam, dịch tả vẫn diễn ra phức tạp vì vậy công tác theo dõi, giám sát và dự báo 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

Qua phân tích các mô hình dự báo trong phần tổng quan và các kỹ thuật áp dụng trong dự báo, nghiên cứu sinh nhận định mô hình dự báo cần được thiết lập phù hợp với các dữ liệu thu thập được và với đặc thù của Việt Nam. Việc xây dựng mô hình dự báo dịch tả tại Hà Nội cần được thực hiện theo các định hướng sau:

- Giải pháp xây dựng mô hình dự báo theo tiếp cận mô hình thống kê cũng như mô hình khai phá dữ liệu cần được xem xét đồng thời.
- Trong tiếp cận mô hình, cần thử nghiệm cả hai tiếp cận mô hình hóa dựa trên hồi qui và phân lớp với phân vùng không gian để tìm kiếm và đánh giá mô hình phù hợp nhất.
- Nghiên cứu giải pháp xây dựng mô hình dự báo dịch tả dựa trên các kỹ thuật phân tích không gian của hệ thống thông tin địa lý GIS. Mô hình này không chỉ cung cấp một phương tiện trực quan hóa các sự kiện dịch tả mà còn là nguồn cung cấp các dữ liệu phục vụ việc mô phỏng dịch tả.

Từ các định hướng trên, luận án tập trung nghiên cứu các vấn đề sau:

- Về lý thuyết: Nghiên cứu cơ sở khoa học của dự báo và phân tích dự báo;
- Về xây dựng mô hình: 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: (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 toàn diện bài toán dự báo dịch bệnh.

1.5. Dữ liệu sử dụng trong nghiên cứu và tiền xử lý dữ liệu

1.5.1 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 bao gồm các số liệu về số ca dịch tả , về khí hậu và thủy văn khu vực Hà Nội. Trong phần này sẽ mô tả các tập số liệu được hồi cứu phục vụ cho nghiên cứu:

Số liệu dịch tả : Hồi cứu toàn bộ số ca tả dựa trên báo cáo tháng, báo cáo năm của Trung tâm Y học Dự phòng Hà Nội trong giai đoạn từ ngày 01/01/2001 đến 31/12/2012. Tiêu chuẩn lựa chọn là các ca tả có địa chỉ thường trú tại các quận/huyện trong thành phố Hà Nội. Tiêu chuẩn loại trừ là các ca tả không đầy đủ thông tin địa chỉ hoặc bệnh nhân do y tế các tuyến dưới gửi lên. Tập dữ liệu này bao gồm các trường Họ tên bệnh nhân, tuổi, giới tính, phường, quận, ngày mắc, ngày vào viện, tên bệnh viện. Số liệu sau khi thu thập về được kiểm tra đảm bảo đầy đủ và chính xác, được nhập vào máy tính bằng phần mềm Excel.

Số liệu mức nước các sông: Dữ liệu mực nước tại bốn trạm đo gồm có:

- Trạm Hà Nội giai đoạn các năm 1960 – 2012,
- Trạm Sơn Tây và trạm Thượng Cát giai đoạn các năm 1960-2013,
- Trạm Hà Đông giai đoạn các năm 1998-2003.

Số liệu mực nước được đo theo tháng phù hợp với đơn vị thời gian trong mô hình dự báo là tháng.

Số liệu khí hậu- thời tiết: Số liệu khí hậu - thời tiết được đo tại năm trạm khí tượng là Ba Vì, Sơn Tây, Láng, Hoài Đức và Hà Đông thuộc địa bàn Hà Nội trong giai đoạn 2001-2012 từ Trung tâm Nghiên cứu Khí Tượng Thủy Văn Trung Ương. Các thông số khí hậu gồm có: (1) Nhiệt độ không khí: trung bình ngày, cao nhất ngày, thấp nhất ngày. Từ các số liệu nhiệt độ ngày tính toán để có số liệu nhiệt độ trung bình tháng, cao nhất tháng, thấp nhất tháng; (2) Độ ẩm không khí: trung bình ngày, cao nhất ngày, thấp nhất ngày. Từ các số liệu ẩm độ ngày tính toán để có số liệu độ ẩm trung bình tháng, cao nhất tháng, thấp nhất tháng; (3) Lượng mưa: lượng mưa hàng ngày, từ đó tính toán để có lượng mưa tháng, số ngày mưa trong tháng. Ngoài

ra, các thông số (4) số giờ nắng hàng ngày và (5) tốc độ gió trung bình ngày cũng được ghi nhận và tính toán theo phương pháp tương tự.

Số liệu không gian thông tin địa lý: Luận án sử dụng tập số liệu bản đồ Hà Nội với bản đồ hành chính thể hiện ranh giới địa lý hành chính của 29 quận/huyện và các lớp đường phố, sông hồ, diện tích mặt nước với tỷ lệ 1:50 000. Tập dữ liệu này được thu thập từ Trung Tâm Nghiên Cứu Môi Trường thuộc Bộ Tài Nguyên Môi Trường. Việc xác định toàn bộ các quận/huyện lân cận của một quận/huyện trong luận án được thực hiện bằng truy vấn không gian trong tập số liệu này.

Số liệu về chỉ số dao động phía Nam (Southern Oscillation Index- SOI): Số liệu này để đo sự tiến triển và cường độ của El Nino và La Nina ở Thái Bình Dương, được đo theo tháng từ 1/2001 đến 12/2012. Tập dữ liệu này được lấy từ nguồn của chính quyền bang Queensland, Úc. Dữ liệu SOI được thể hiện bằng một số thực [1].

1.5.2 Tiền xử lý dữ liệu

Do dữ liệu về các đơn vị hành chính phường/xã, quận huyện của bệnh nhân tả chưa được chuẩn hóa, khuôn dạng dữ liệu khí tượng ở 5 trạm đo chưa được thống nhất và định dạng chuẩn, nên các dữ liệu được tiền xử lý qua các bước sau:

- Định dạng lại tệp dữ liệu chứa thông tin bệnh nhân để có thể xử lý tự động bằng chương trình trên máy tính.
- Xử lý thủ công một số tệp bảng tính Excel chứa số liệu khí tượng để thống nhất về qui cách và định dạng để đưa vào xử lý tự động.
- Chuyển đổi dữ liệu SOI sang định dạng bảng tính Excel.
- Tạo lập ra một bảng dữ liệu kết hợp dữ liệu khí tượng và ca bệnh tả với cấu trúc như sau: Mỗi dòng của bảng ứng với một ngày, từ 1/1/2001 đến 31/12/2012.Các cột của bảng gồm các nhóm: (1) Nhóm thuộc tính khí tượng

---

[1] Tập dữ liệu hệ số dao động phía Nam SOI ghi nhận theo ngày cho các thủ đô trên toàn cầu của chính quyền bang Queensland – Úc.
https://www.longpaddock.qld.gov.au/seasonnalclimateoutlook/southernoscillationindex/soidatafiles/DailySOI1887-1889Base.txt

của 5 trạm đo như đã mô tả ở trên, (2) thuộc tính SOI và (3) nhóm 29 thuộc tính mô tả số ca mắc tả tại mỗi quận/huyện trong địa bàn thành phố Hà Nội. Bảng dữ liệu theo ngày được gọi tắt là DL1.

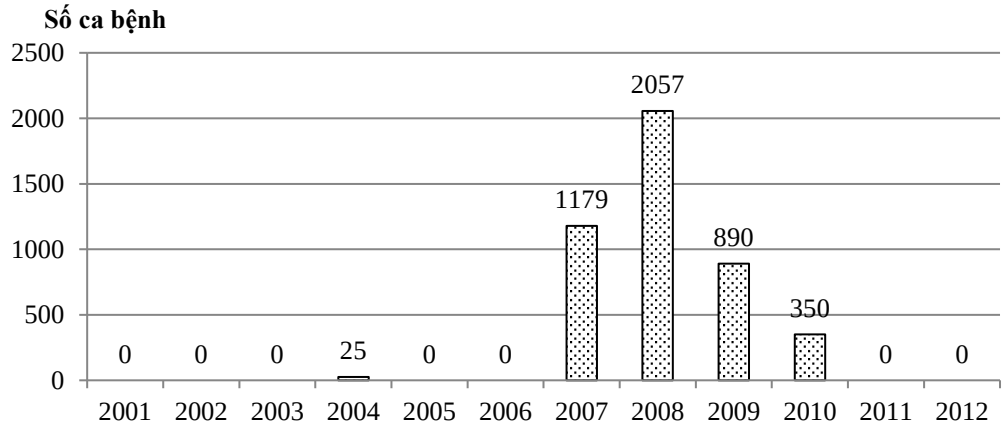
Từ bảng dữ liệu theo ngày DL1, tổng hợp lại để có bảng dữ liệu theo tháng (gọi tắt là bảng DL2) với cấu trúc: Mỗi dòng của bảng ứng với một tháng, từ 1/2001 đến 12/2012. Các cột của bảng dữ liệu theo tháng tương tự bảng dữ liệu theo ngày.

Việc phân tích dữ liệu, dự báo sẽ được căn cứ chủ yếu vào hai bảng dữ liệu ngày và tháng nói trên. Sau khi tạo bảng DL1, có 5 cột dữ liệu thuộc nhóm khí tượng không thể sử dụng được do bị thiếu hoặc hoàn toàn không có dữ liệu. Các cột dữ liệu này bị loại bỏ, bao gồm:

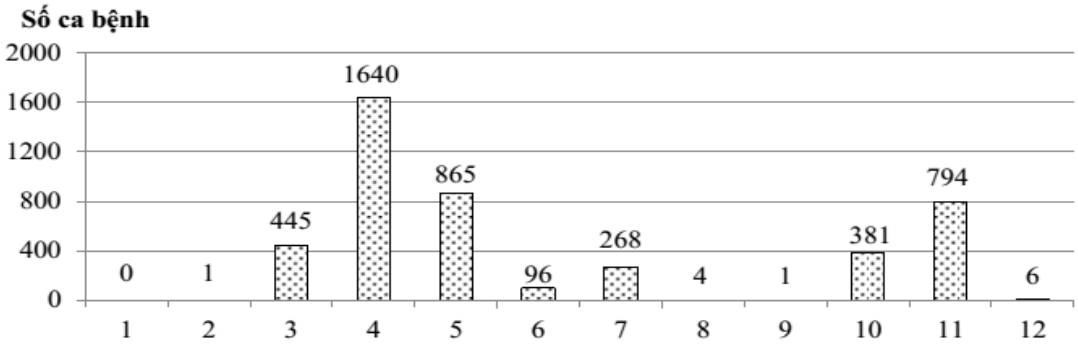
- Tốc độ gió của trạm Hoài Đức
- Độ ẩm cao nhất ngày của trạm Hoài Đức
- Nhiệt độ thấp nhất ngày của trạm Hoài Đức
- Độ ẩm cao nhất ngày của trạm Ba Vì, và
- Nhiệt độ cao nhất ngày của trạm Ba Vì

Tại trạm Hà Đông, số liệu mực nước không có trong giai đoạn 2007-2010 cho nên chỉ sử dụng được số liệu mực nước tại ba trạm còn lại.

Dữ liệu chi tiết của các ca bệnh tả được tổng hợp lại theo số lượng ca bệnh của các quận theo từng ngày. Sau khi tổng hợp dữ liệu là một bảng có 4383 bản ghi, mỗi bản ghi có 29 trường tương ứng với 29 quận/huyện của Hà Nội. Dữ liệu nhiệt độ, độ ẩm, lượng mưa, số giờ nắng và tốc độ gió: bao gồm giá trị thấp nhất, giá trị cao nhất và giá trị trung bình theo ngày. Đây là tập dữ liệu có nguồn từ 5 trạm khí tượng Láng, Ba Vì, Hà Đông, Hoài Đức và Sơn Tây ghi nhận theo ngày.



Biểu đồ 1.1: Phân bố ca bệnh Tả của Hà nội giai đoạn 2001-2012 theo năm



Biểu đồ 1.2 : Phân bố ca bệnh Tả của Hà nội theo tháng

Thống kê các quận huyện hàng năm có trên 100 ca dịch tả hoặc thuộc tốp 5 số ca dịch tả cho thấy xuất hiện 4 lần có Đống Đa, Hai Bà Trưng, 3 lần có Thanh Xuân, Hoàng Mai, 2 lần có Ba Đình, 1 lần có Cầu Giấy, Hà Đông, Thạch Thất, Thường Tín.

Theo bộ số liệu này, Hà Nội có 4 đợt bùng phát dịch tả vào các năm 2004, 2007, 2008, 2009 và 2010. Số ca bệnh tả trong giai đoạn 2001-2012 trên toàn thành phố được mô tả trên các biểu đồ 1.1, 1.2 và 1.3. Trong các đợt bùng phát dịch năm 2007, 2008, 2009 và 2010, hầu hết tất cả các quận trong thành phố đều có ca bệnh và khoảng thời gian xuất hiện các ca bệnh khá giống nhau. Năm 2004 có số ca bệnh tả thấp nhất và các ca bệnh chỉ có ở các quận Ba Đình, Hai Bà Trưng và Hoàng Mai. Như vậy dữ liệu số ca bệnh tả từ 2001 đến 2012 là không cân bằng (số ngày có ca bệnh là 185 trên tổng số 4383 ngày của 12 năm, chiếm 4,22%, hoặc 13% nếu tính theo tháng).

Dữ liệu của các trạm có độ tương quan rất cao giữa các biến cùng loại, ví dụ tương quan của nhiệt độ trung bình ngày đo tại trạm Ba Vì và tại trạm Láng là 0.95. Đồng thời tương quan giữa nhiệt độ cao nhất, nhiệt độ trung bình và nhiệt độ thấp nhất ngày cũng có tương quan rất cao (tương tự với độ ẩm).

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 và vấn đề tiền xử lý dữ liệu.

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

Trong chương này, luận án sẽ lần lượt đề xuất các mô hình dự báo dịch tả dựa trên khai phá luật kết hợp và học máy hồi qui và phân lớp. Trên cơ sở các mô hình dự báo dựa trên khai phá luật kết hợp, học máy hồi qui và phân lớp đề xuất, các thực nghiệm được thực hiện để đánh giá khả năng dự báo dịch tả tại khu vực Hà Nội.

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

Qua phân tích, quan sát trực quan trên bản đồ ca bệnh cho thấy số ca bệnh tả có xu hướng xuất hiện tập trung quanh các con sông đang bị ô nhiễm trong địa bàn Hà Nội. Câu hỏi đặt ra là có mối liên quan nào giữa các con sông này và các địa điểm có con sông chảy qua với việc xuất hiện ca bệnh tả không? Để có câu trả lời, luận án đã tiến hành dự đoán khả năng xuất hiện bệnh tả trên địa bàn thành phố Hà Nội dựa trên việc sinh các luật kết hợp từ bộ dữ liệu các ca bệnh tả tại các quận huyện ở Hà Nội trong giai đoạn từ năm 2001 đến năm 2012.

2.1.1 Khai phá luật kết hợp sử dụng thuật toán Apriori

Các bộ dữ liệu DL1 và DL2 đã được mô tả trong Chương 1 được sử dụng cho thực nghiệm dự báo dịch tả dựa trên khai phá luật kết hợp. Trên cơ sở sử dụng ngôn ngữ R [17], [87] để tạo ra một bảng dữ liệu các ca mắc tả của từng quận, huyện trong thành phố Hà Nội, 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 đề xuất sử dụng phương pháp dự đoán khả năng xuất hiện bệnh tả trên địa bàn thành phố Hà Nội dựa trên 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. Cụ thể, thuật toán Apriori [17] được mô tả như sau :

Algorithm 3: Apriori

Đầu vào:

Tập các giao dịch D, ngưỡng support tối thiểu min_sup

Đầu ra:

Các tập phổ biến trong D

Thuật toán Apriori:

```
1.  L1 = {large 1-itemsets}
2.  for (k = 2; Lk-1 ≠ Ø; k++) do begin
3.      Ck = apriori-gen(Lk-1);
4.      Ct = subset(Ck, t);
5.      forall candidates c € Ct do
6.      c.count++;
7.  end
8.  Lk = {c € Ck | c.count ≥ min_sup} end
9.  Answer = ∪k Lk;
```

Hàm *apriori-gen()* trong thuật toán Apriori [17] gồm Bước nối và Bước tỉa, cụ thể như sau:

- ***Bước nối***: Sinh các tập mục là ứng viên tập phổ biến bằng cách kết hợp hai tập phổ biến có độ dài k và trùng nhau ở k-1;
- ***Bước tỉa***: Giữ lại tất cả các luật thỏa tính chất nghĩa là đã loại (tỉa) bớt đi mọi ứng viên không đáp ứng.

Sử dụng thuật toán sinh luật để sinh ra các luật kết hợp [103].

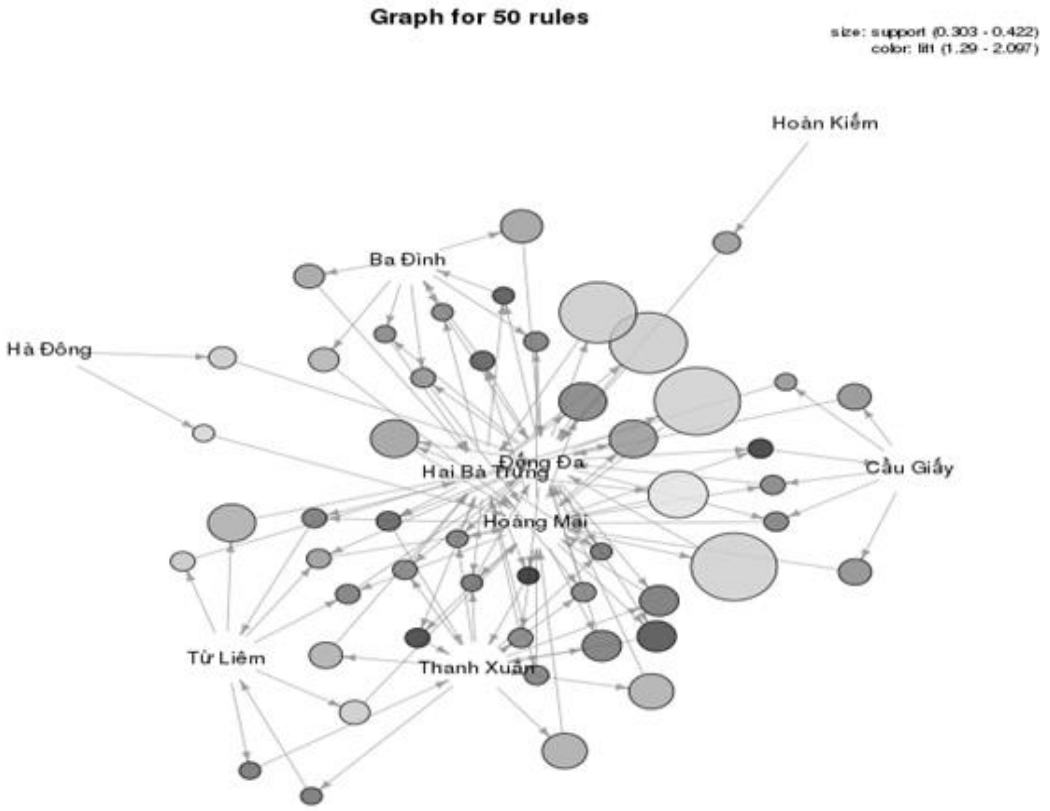
Algorithm 4: Generate rule

```
Generate_rules (L)
1. forall large k-itemsets Lk, k ≥ 2 do
2. begin
```

```
3.    H1 = {tập 1-item để sinh ra tập item tiếp theo}
4. call ap-genrules(Lk, H1);
5. end
6. procedure ap-genrules(Lk, Hm)
7.       if (k > m+1) then begin
8.            Hm+1 = apriori-gen(Hm);
9.           forall hm+1 € Hm+1 do begin
10.                  conf = supp(Lk) / supp(Lk - hm+1);
11.                 if (conf  ≥ minconf) then
12.                     ouput the rule (Lk - hm+1) => hm+1
13.                       with the confidence = conf and
14.                       support = supp(Lk)
15.                 else
16.                     delete hm+1 from Hm+1;
17.             end
18.             call ap-genrules(Lk, Hm+1);
19.       end.
```

2.1.2. Kết quả thử nghiệm

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 và Hình 2.1. Chi tiết 50 luật thu nhận được trong thử nghiệm được thể hiện trong Phụ lục 1 của Luận án. 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.



Hình 2.1. 50 luật thu được với độ đo thống kê lớn hơn 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	{Thanh Xuân}	{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

Các quận xuất hiện trong cả vế trái và vế phải của 50 luật kết hợp bao gồm 9 quận/huyện: Đống Đa, Hai Bà Trưng, Hoàng Mai, Thanh Xuân, Từ Liêm, Hà Đông, Ba Đình, Cầu Giấy và Hoàn Kiếm, trong đó chỉ có quận Hoàn Kiếm không có con

sông nào chảy qua địa bàn. Xem xét các yếu tố thủy văn Hà Nội có ảnh hưởng đến sự lây lan của dịch tả , có 3 con sông bị ô nhiễm nặng chảy qua thành phố Hà Nội, bao gồm sông Tô Lịch, sông Kim Ngưu và sông Nhuệ [8]. Các con sông này chảy qua một số quận/huyện như thể hiện trong Bảng 2.2. Bảng 2.2 cũng cho biết danh sách các quận/huyện tiếp giáp với quận/huyện bị ô nhiễm bởi các con sông chảy qua.

Bảng 2.2. Các quận/huyện có sông ô nhiễm chảy qua và các quận/huyện tiếp giáp

Quận/Huyện	**Các sông chảy qua**	**Quận/huyện tiếp giáp**
Ba Đình	Tô Lịch	Hoàn Kiếm, Cầu Giấy, Đống Đa
Cầu Giấy	Tô Lịch	Từ Liêm, Ba Đình, Cầu Giấy, Đống Đa
Đống Đa	Tô Lịch	Hoàn Kiếm, Cầu Giấy, Ba Đình, Hai Bà Trưng, Thanh Xuân
Hà Đông	Nhuệ	Từ Liêm, Thanh Xuân
Hai Bà Trưng	Kim Ngưu	Hoàn Kiếm, Hoàng Mai, Thanh Xuân, Đống Đa
Hoàng Mai	Kim Ngưu, Tô Lịch	Hai Bà Trưng, Thanh Xuân
Hoàn Kiếm		Ba Đình, Hai Bà Trưng, Đống Đa
Thanh Xuân	Tô Lịch, Kim Ngưu	Cầu Giấy, Hà Đông, Hoàng Mai, Đống Đa
Từ Liêm	Nhuệ	Cầu Giấy, Hà Đông, Thanh Xuân

2.1.3. Nhận xét

Nghiên cứu này khai phá các luật kết hợp số ca bệnh tả với dữ liệu thủy hệ của Hà Nội và 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%).

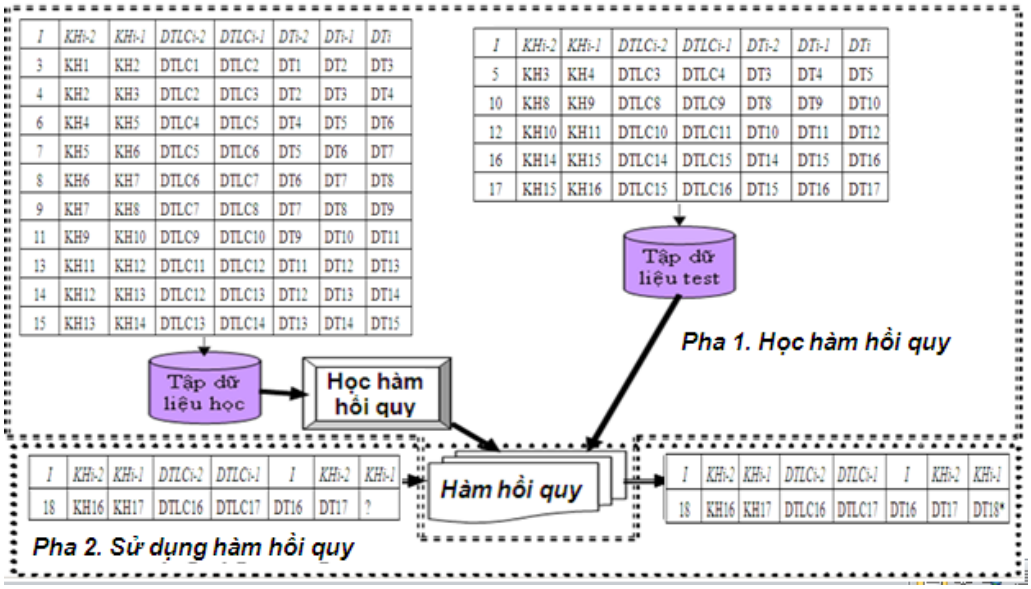
Kết quả nghiên cứu cho thấy xu hướng cùng xuất hiện ca bệnh tả tại các quận/huyện có sông ô nhiễm chảy qua tại Hà Nội dựa trên khai phá luật kết hợp tương đồng với kết quả các nghiên cứu về bệnh tả trên thế giới và Việt Nam trước đây [54], [88] [85], [80], [5], [9]. Điều này khẳng định khai phá luật kết hợp phù hợp với mô hình dự báo dịch tả trong điều kiện dữ liệu phân bố không chuẩn và không có sự khác biệt nhiều về điều kiện tự nhiên khí hậu giữa các vùng miền. Mặc dù các kết quả

nghiên cứu mới là bước đầu, nhưng với kết quả là tập các luật kết hợp với độ tin cậy và độ chắc chắn khá cao có thể sử dụng như là một trong các yếu tố hỗ trợ ra quyết định trong công tác phòng chống dịch tại thành phố Hà Nội. 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. Nghiên cứu này đã đăng trong kỷ yếu hội nghị quốc tế về Truyền thông quản lý và viễn thông 2015 (ComManTel2015) tại Đà Nẵng - Việt Nam.

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

2.2.1 Bài toán dự báo với kỹ thuật hồi qui

Kỹ thuật hồi qui được chia thành hai lớp chính là hồi qui tuyến tính và hồi qui phi tuyến theo dạng của hàm dự báo f trong mô hình dự báo. Kỹ thuật hồi qui (tuyến tính hay phi tuyến) đều hướng tới mô hình hồi qui khớp nhất với tập dữ liệu D_{learn} có nghĩa là quá trình xây dựng hàm hồi qui được quy về một bài toán xác định tham số với ràng buộc sai số giá trị biến đầu ra thực tế với giá trị biến đầu ra theo mô hình là cực tiểu. Hình 2.2 mô tả minh họa một ví dụ sử dụng kỹ thuật hồi qui xây dựng mô hình dự báo dịch tả. Giả sử, với đơn vị thời gian là một tháng, sau bước khảo sát dữ liệu dịch tả và khí hậu, chúng ta lựa chọn các biến sau đây (KH_i, DT_i) là giá trị khí hậu (KH_i) và giá trị dịch tả (DT_i) vào thời điểm thứ *i* tại quận/huyện đang được xem xét. Giá trị dịch tả của các quận/huyện lân cận với quận/huyện đang xét tại thời điểm *i* được ký hiệu là $DTLC_i$. Giả sử cần dự báo cho một tháng tiếp theo đối với quận huyện đang xem xét. Phân tích bài toán cho thấy giá trị biến dịch tả vào thời điểm thứ t là DT_t phụ thuộc vào các giá trị: (i) giá trị dịch tả của quận/huyện đang xem xét ở thời điểm trước đó DT_{t-2}, giá trị biến dịch tả ở vùng phụ cận ở thời điểm trước đó $DTLC_{t-2}$, giá trị biến khí hậu của quận/huyện đang xem xét thời điểm trước đó KH_{t-2}.



Hình 2.2. Quá trình học và sử dụng hàm hồi quy

Một trường hợp riêng của lớp mô hình hồi qui phi tuyến là mô hình hồi qui logarit, trong đó dữ liệu được thay thế bằng giá trị logarit của chúng thì phù hợp với mô hình hồi qui tuyến tính. Xây dựng mô hình hồi qui tuyến tính cho giá trị logarit, sau đó sử dụng hàm mũ để chuyển đổi giá trị kết quả trở về giá trị dạng thông thường của dữ liệu.

Kiểm thử trong hồi quy

Mô hình hồi qui hầu như bao giờ cũng có sai số vì hiện tượng tự nhiên và xã hội phụ thuộc nhiều yếu tố, diễn biến rất phức tạp, rất khó có thể ước lượng hết. Để đánh giá, so sánh các phương pháp dự báo một cách định lượng, các chỉ số đánh giá mô hình dự báo được sử dụng. Dưới đây là một số chỉ số đánh giá thông dụng nhất:

(i) Sai số quân phương *MSE* (*Mean Square Error*):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Q_i - \hat{Q}_i\right)^2 \qquad (2.1)$$

(ii) Sai số căn quân phương *RMSE* (*Root Mean Square Error*):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_i - \hat{Q}_i\right)^2} \qquad (2.2)$$

(iii) Sai số tuyệt đối *MAE* (*Mean Absolute Error*):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Q_i - \hat{Q}_i\right| \quad (2.3)$$

Trong đó:

n: Số lượng các điểm dữ liệu trong bộ dữ liệu kiểm thử.

$\hat{Q}_i$: Giá trị tính toán tại điểm dữ liệu thứ i trong bộ dữ liệu kiểm thử.

Q_i: Giá trị thực đo tại điểm dữ liệu thứ i trong bộ dữ liệu kiểm thử.

Giá trị các chỉ số đánh giá này của một phương pháp dự báo càng nhỏ thì chứng tỏ rằng phương pháp dự báo đó càng tốt.

Các chỉ số MSE, RMSE, MAE trực quan và dễ dàng tính toán, song trong nhiều trường hợp khi dung lượng dữ liệu lớn hay dữ liệu có độ biến động cao thì các chỉ số này trở nên quá thô sơ. Trong một số trường hợp, người ta còn sử dụng Chỉ số hiệu quả - E và Chỉ số xác định - R^2. Các chỉ số này tuy có độ phức tạp tính toán cao hơn song có thể khắc phục được hạn chế về tính thô sơ của các chỉ số MSE, RMSE, MAE. Dưới đây là công thức tính các chỉ số E và R^2:

(iv) Chỉ số hiệu quả - E (*Coefficient of Efficiency*)

$$E = 1 - \frac{\sum_{i=1}^{n}\left(Q_i - \hat{Q}_i\right)^2}{\sum_{i=1}^{n}\left(Q_i - \overline{Q}\right)^2} \quad (2.4)$$

(v) Chỉ số xác định - R^2 (*Coefficient of Determination*)

$$R^2 = \frac{\sum_{i=1}^{n}\left(Q_i - \overline{Q}\right)\left(\hat{Q}_i - \overline{\hat{Q}}\right)}{\sqrt{\sum_{i=1}^{n}\left(Q_i - \overline{Q}\right)^2 \sum_{i=1}^{n}\left(\hat{Q}_i - \overline{\hat{Q}}\right)^2}} \quad (2.5)$$

Các chỉ số E và R^2 có thể được dùng theo cách kết hợp hoặc riêng rẽ. Phương pháp dự báo tốt là phương pháp cho giá trị của các chỉ số này cao.

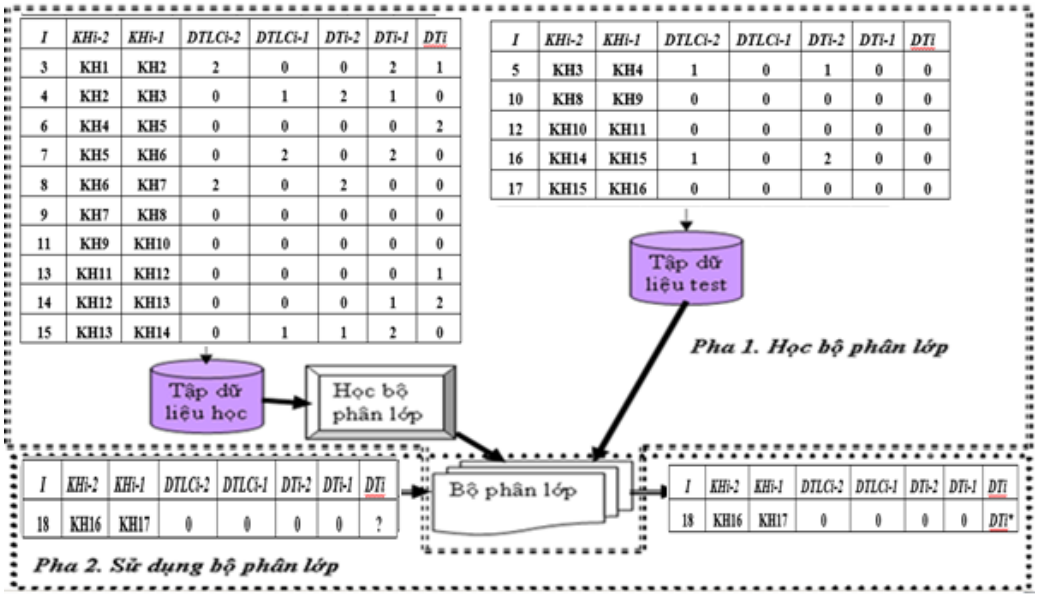
2.2.2 Dự báo với kỹ thuật phân lớp

Phân lớp là một kỹ thuật khai phá dữ liệu, bản chất là dự đoán các nhãn (hay lớp) của các phần tử dữ liệu đầu vào và các nhãn này nhận các giá trị rời rạc. Đầu vào của bài toán phân lớp là một tập các mẫu dữ liệu huấn luyện với một nhãn phân lớp

cho mỗi mẫu dữ liệu. Đầu ra là bộ phân lớp dựa trên tập huấn luyện hoặc những nhãn phân lớp. Kỹ thuật phân lớp dữ liệu gồm hai bước:

Bước 1: xây dựng mô hình từ tập huấn luyện gọi là bước học (learning step, hay *pha học*: learning phase) và tập dữ liệu gán nhãn phục vụ quá trình học này được gọi là *dữ liệu huấn luyện* (training data). Dữ liệu huấn luyện là một tập các *phần tử dữ liệu* có gán nhãn. Một điểm (phần tử) dữ liệu X thường được biểu diễn bằng một vector n chiều $X=(x_1, x_2,\ldots, x_n)$, trong đó mỗi thành phần x_i trong vector chứa một giá trị biểu diễn *thuộc tính* (attribute, còn được gọi là *đặc trưng*: feature) A_i của phần tử dữ liệu đó. Về bản chất trong bước 1 này, các thuật toán phân lớp học ra hàm $y=f(X)$ để từ đó khi có một phần tử X mới nó sẽ dự đoán ra nhãn y tương ứng với X. Theo khía cạnh này thì ta có thể thấy bước 1 là quá trình học ra một hàm có khả năng dự đoán được nhãn lớp dữ liệu.

Bước 2: Sử dụng mô hình – kiểm tra tính đúng đắn của mô hình và dùng nó để phân lớp dữ liệu mới.



Hình 2.3. Quá trình học và sử dụng mô hình (bộ) phân lớp

Tùy vào các thuật toán khác nhau mà hàm $f(X)$ có thể có các dạng khác nhau như ở dạng luật (rule), cây quyết định (decision tree) hay các công thức toán học… Hình 2.3 minh họa quá trình học và sử dụng mô hình phân lớp đối với bài toán dự báo dịch tả nhưng với một điểm khác biệt về giá trị của biến đầu ra. Biến dịch tả chỉ

nhận một trong ba giá trị “0” (không có dịch tả), “1” (“mức tả thấp”), và “2” (“mức tả cao”).

Kiểm thử trong phân lớp

Hiện nay, tồn tại nhiều độ đo để đánh giá các mô hình mà điển hình nhất là bộ độ đo (độ hồi tưởng, độ chính xác, f_1 (f_β)) và bộ độ đo (độ chính xác, hệ số lỗi). So sánh các mô hình có thể sử dụng một hoặc một vài độ đo cũng như thực hiện trên một bộ các tập dữ liệu liên quan tới bài toán phân lớp đang nghiên cứu. Trong phương án kiểm thử theo bộ độ đo (độ hồi tưởng, độ chính xác, f_1 (f_β)), lớp đang quan tâm được gọi là lớp dương (positives), và lớp còn lại được gọi là lớp âm (negatives). Mỗi điểm dữ liệu trong tập dữ liệu kiểm thử sẽ thuộc vào một trong bốn tình huống sau đây:

- Gọi TP là số lượng các điểm dữ liệu thuộc D_{test} rơi vào tình huống mà giá trị thực sự và giá trị dự báo đều là P.

- Gọi TN là số lượng các điểm dữ liệu thuộc D_{test} rơi vào tình huống mà giá trị thực sự và giá trị dự báo đều là N.

- Gọi FP là số lượng các điểm dữ liệu thuộc D_{test} rơi vào tình huống mà giá trị thực sự là P và giá trị dự báo là N.

- Gọi FN là số lượng các điểm dữ liệu thuộc D_{test} rơi vào tình huống giá trị thực sự là N và giá trị dự báo là P.

Ma trận nhầm lẫn là tổng hợp các kết quả trên đây thể hiện trong bảng 2.3.

Bảng 2.3: Ma trận nhầm lẫn.

Lớp dự báo / Lớp thực sự	Lớp = P	Lớp = N
Lớp = P	TP	FN
Lớp = N	FP	TN

Khi đó, độ hồi tưởng (recall) ρ, độ chính xác (precision) π, và độ đo f_β kết hợp độ hồi tưởng và độ chính xác được xác định theo các công thức sau đây:

$$\pi = \frac{TP}{TP + FP}, \quad \rho = \frac{TP}{TP + FN}, \quad f_\beta = \frac{(\beta^2 + 1)\rho\pi}{\beta^2\pi + \rho} \qquad (2.6)$$

Độ đo f_1 (trường hợp β=1) được sử dụng rất phổ biến và thường được viết là f.

2.2.3. Dự báo bệnh tả dựa trên học máy hồi qui và 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 mô tả trong Chương 1 của luận án theo hướng hạn chế phạm vi các chiều không gian, thời gian như sau: Về chiều thời gian, do các giai đoạn 2001-2006 và 2011-2012 hoặc không có số liệu về ca dịch tả cho nên mô hình dự báo được tập trung vào giai đoạn các năm 2007-2010. Mô hình dự báo dịch tả tại khu vực Hà Nội thuộc loại bài toán dự báo dữ liệu chuỗi thời gian, vì vậy, tập dữ liệu được dùng để học mô hình là tập dữ liệu "quá khứ" (từ tháng 01/2007 đến tháng 06/2010) và tập dữ liệu kiểm tra mô hình là tập dữ liệu "tương lai" (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 xây dựng mô hình dự báo bùng phát dịch tả được diễn giải như sau:

Coi đơn vị thời gian là tháng: chỉ số thời gian dữ liệu nhận các giá trị 0, 1, 2,.., t, t+1,....

Biến ra y là trạng thái dịch tả cần dự báo tại thời điểm t+k, trong đó t là thời điểm dự báo và k là khoảng cách dự báo (dự báo trước k tháng). Giá trị biến ra hoặc là liên tục

(số bệnh nhân mắc dịch tả) tương ứng với mô hình hồi quy, hoặc là rời rạc {0, 1,.., N} hoặc {Có dịch tả, Không có dịch tả} tương ứng với mô hình phân lớp.

Các số liệu đã có về giá trị của biến về dịch tả, về môi trường và khí hậu sẽ được tập hợp thành tập dữ liệu ví dụ $D_{example}$. Như vậy với khoảng cách dự báo k = 2 thì bài toán được phát biểu như sau:

***Đầu vào*:** Tập dữ liệu ví dụ $D_{example}$ bao gồm các phần tử dữ liệu d có dạng:

$d = (KH_{t-2}, KH_{t-1}, DTLC_{t-2}, DTLC_{t-1}, DT_{t-2}, DT_{t-1}, DT_t)$

Trong đó, KH_{t-2}, KH_{t-1} lần lượt là giá trị khí hậu vào thời điểm t-2, t-1 tại quận/huyện đang xét, là danh sách các biến khí hậu – thủy văn trong thực tế. $DTLC_{t-2}$, $DTLC_{t-1}$ lần lượt là giá trị dịch tả vào thời điểm t-2, t-1 tại quận/huyện lân cận với quận/huyện đang xét. DT_{t-2}, DT_{t-1}, DT_t lần lượt là giá trị dịch tả vào thời điểm t-2, t-1, t tại quận/huyện đang xét. Như vậy, DT_t là biến mục tiêu, tập { KH_{t-2}, KH_{t-1}, $DTLC_{t-2}$, $DTLC_{t-1}$, DT_{t-2}, DT_{t-1}} là tập biến đầu vào.

***Đầu ra*:** Mô hình dự báo thường được viết dưới dạng $y=f(x_1, x_2,..., x_n) + \varepsilon$ (trong trường hợp mô hình hồi quy) hoặc một mô hình tương ứng theo một thuật toán phân lớp.

Từ tập dữ liệu đầu vào, xây dựng mô hình dự báo đầu ra, thực nghiệm áp dụng các kĩ thuật hồi quy, phân lớp. Áp dụng các bộ công cụ phân tích dữ liệu có các thành phần thực thi các mô hình hồi quy, phân lớp điển hình.

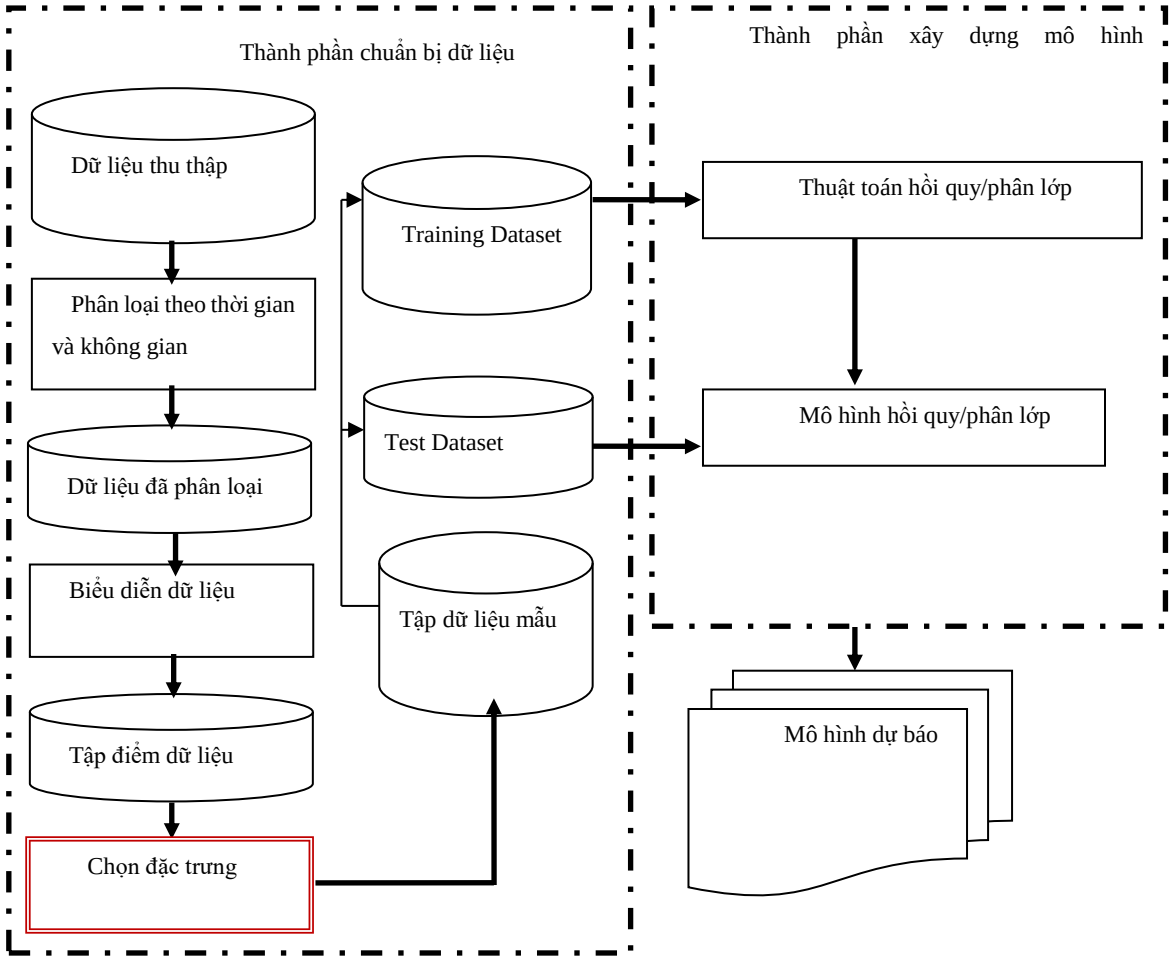
Mô hình cục bộ cho một quận huyện, mỗi điểm dữ liệu thể hiện cho một vector ($KH_{i,t-2}$, $KH_{i,t-1}$, $KHLC_{i,t-1}$,$KHLC_{i,t-2}$,$DTLC_{i,t-1}$, $DTLC_{i,t-2}$,$DT_{i,t}$) trong đó $DT_{i,t}$ là biến mục tiêu và những biến khác là biến điều kiện. Ở mô hình dự báo toàn cục, mỗi điểm dữ liệu thể hiện cho một vector (KH_{t-2}, KH_{t-1}, DT_{t-1},DT_{t-2},DT_t) trong đó DT_t là vetor mục tiêu và những biến khác là biến điều kiện.

Ở mô hình biểu diễn dữ liệu cục bộ, dự báo tình trạng dịch tả cho quận-huyện QH_i tại thời điểm t dựa trên thông tin về tình trạng dịch tả và khí hậu ở quận-huyện QH_i và các quận huyện lân cận tại thời điểm t-1 và t-2. Các tham số khí hậu được xác định dựa trên dữ liệu tại các trạm khí tượng, thủy văn gần nhất với quận huyện đang dự báo.

Ở mô hình biểu diễn dữ liệu toàn cục, biến mục tiêu là tình trạng dịch tả tại 29

quận/huyện ở thời điểm t. Các biến điều kiện là trạng thái dịch tả ở tất cả các quận/huyện trong thời điểm t-1 và t-2 và tham số khí hậu được lấy ở tất cả các trạm trong thời điểm t-1 và t-2.

Nghiên cứu áp dụng việc lựa chọn đặc trưng, một vài các đặc trưng yếu sẽ được loại bỏ. Tập mẫu nhận được sau bước Chọn đặc trưng được phân thành hai tập dữ liệu độc lập.Việc phân chia để tập dữ liệu học và tập dữ liệu kiểm thử rời rạc nhau nhằm đảm bảo tính độc lập giữa việc huấn luyện với việc đánh giá mô hình dự báo, do đó việc đánh giá mô hình dự báo đảm bảo tính khách quan.



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

Thực nghiệm được thực hiện sử dụng chức năng *Feature Selection* từ bộ công cụ STATISTICA[2] xác định hệ số tương quan (Correlation Coefficient) của các biến điều kiện với (các) biến mục tiêu và chỉ có các biến điều kiện có hệ số tương quan với (các) biến mục tiêu được giữ lại.

Để tiến hành xây dựng mô hình, các thuật toán khai phá dữ liệu đã được áp dụng bao gồm: hồi qui tuyến tính, RandomForest,, Naive Bayes, SVM. Tập dữ liệu học sẽ sử dụng cho đào tạo mô hình và tập dữ liệu kiểm thử sẽ được dùng để đánh giá mô hình.

Để đánh giá hiệu quả của việc áp dụng giải pháp lựa chọn đặc trưng, hai trường hợp đầu vào là dữ liệu gốc và dữ liệu đã chọn đặc trưng đều được tiến hành. Cả hai trường hợp biểu diễn dữ liệu cục bộ và toàn cục được tiến hành để so sánh, xác định mối quan hệ giữa các yếu tố khí hậu và dịch tả, nghiên cứu thực hiện với trường hợp biến điều kiện chỉ là các yếu tố khí hậu và trường hợp kết hợp cả khí hậu và dịch tả với các giá trị phân 2 lớp {0,1} và phân 3 lớp {0,1,2}; cuối cùng là thực hiện với trường hợp biến điều kiện chỉ là yếu tố trạng thái dịch.

Việc xử lý dữ liệu được tiến hành trên bộ dữ liệu đã thu thập của luận án thông qua các bước sau:

- Thứ nhất, dữ liệu dịch tả tại các năm 2007 đến 2010 được thống kê theo từng tháng, trong mỗi tháng lại thống kê theo từng quận/huyện, theo độ tuổi, theo giới tính.

- Thứ hai, tiến hành chia 29 quận/huyện vào các trạm khí hậu dựa trên quan sát bản đồ. Sau đó lọc lấy các giá trị sau trong các năm 2007-2010: Nhiệt độ trung bình ngày trung bình theo tháng, nhiệt độ cao nhất ngày trung bình theo tháng, nhiệt độ thấp nhất ngày trung bình theo tháng, tổng lượng mưa tháng, độ ẩm trung bình ngày trung bình theo tháng, độ ẩm cao nhất ngày trung bình theo tháng, độ ẩm thấp nhất ngày trung bình theo tháng, tổng số giờ nắng của tháng, vận tốc gió trung bình ngày trung bình theo tháng.

---

[2] *Công cụ thống kê STATISTICA* http://www.statsoft.com/Products/STATISTICA/Product-Index

- Thứ ba, chia 29 quận/huyện vào ba trạm thủy văn dựa trên quan sát bản đồ. Sau đó lọc lấy giá trị mực nước bình quân từng tháng trong các năm 2007-2010.

Cuối cùng tổng hợp các dữ liệu thống kê được tạo 29 file dạng.csv ứng với 29 quận/huyện. Trong đó, mỗi file sẽ chứa 46 điểm dữ liệu (từ tháng 3-2007 đến tháng 12-2010). Mỗi điểm dữ liệu sẽ chứa các thuộc tính ứng với điểm dữ liệu đã xác định ở phần phát biểu bài toán: **d = (KH_{t-2}, KH_{t-1}, $DTLC_{t-2}$, $DTLC_{t-1}$, DT_{t-2}, DT_{t-1}, DT_t).**

Sử dụng một số độ đo đánh giá mô hình dự báo, điển hình là các độ đo 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) [45]. Các công thức tính toán sau đây được áp dụng cho các độ đo tương ứng:

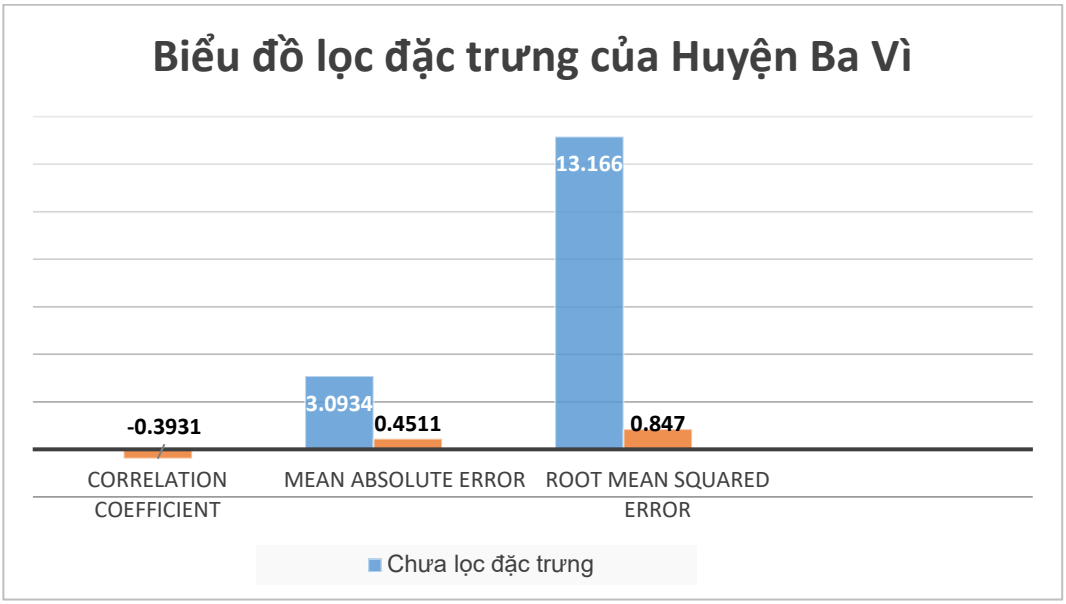
$$MAE = \frac{\sum_{i=1}^{n}|p_i - a_i|}{n}, \qquad RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}},$$

$$CC = \frac{S_{PA}}{\sqrt{S_P S_A}}, \quad \text{trong đó} \quad S_{PA} = \frac{(p_i - \bar{p})(a_i - \bar{a})}{n-1}, \qquad S_P = \frac{\sum_{i=1}^{n}(p_i - \bar{p})^2}{n-1}, \qquad (2.7)$$

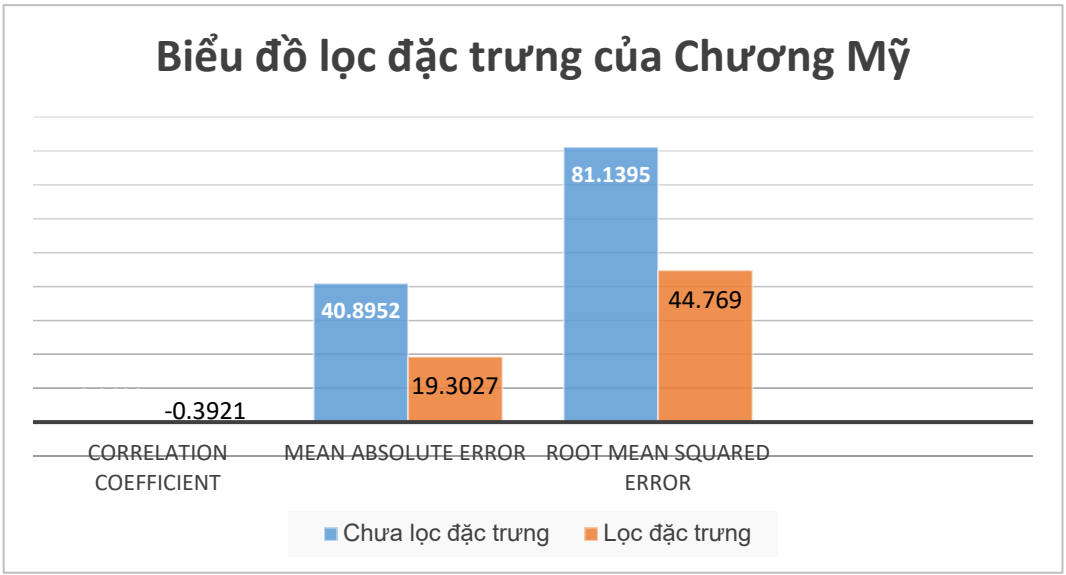
$$S_A = \frac{\sum_{i=1}^{n}(a_i - \bar{a})^2}{n-1}, \qquad \bar{p} = \frac{\sum_{i=1}^{n} p_i}{n}, \text{ và} \qquad \bar{a} = \frac{\sum_{i=1}^{n} a_i}{n}$$

2.2.4.Kết quả thử nghiệm

Áp dụng tính năng lựa chọn đặc trưng trong bộ công cụ STATISTICA với điều kiện lọc là giá trị p-value<=0.09 ứng với độ tin cậy 91%. Sau khi áp dụng hồi quy tuyến tính với mô hình của 29 quận riêng biệt cho kết quả: Sau khi lọc đặc trưng hệ số tương quan (Correlation coefficient) có tốt hơn (càng gần 1 hoặc -1), sai số tuyệt đối (Mean absolute error) và sai số căn quân phương (Root mean squared error) giảm đáng kể. Biểu đồ 2.1 và 2.2 dưới đây là kết quả tiêu biểu cho mô hình dự báo của hai huyện Ba Vì và Chương Mỹ



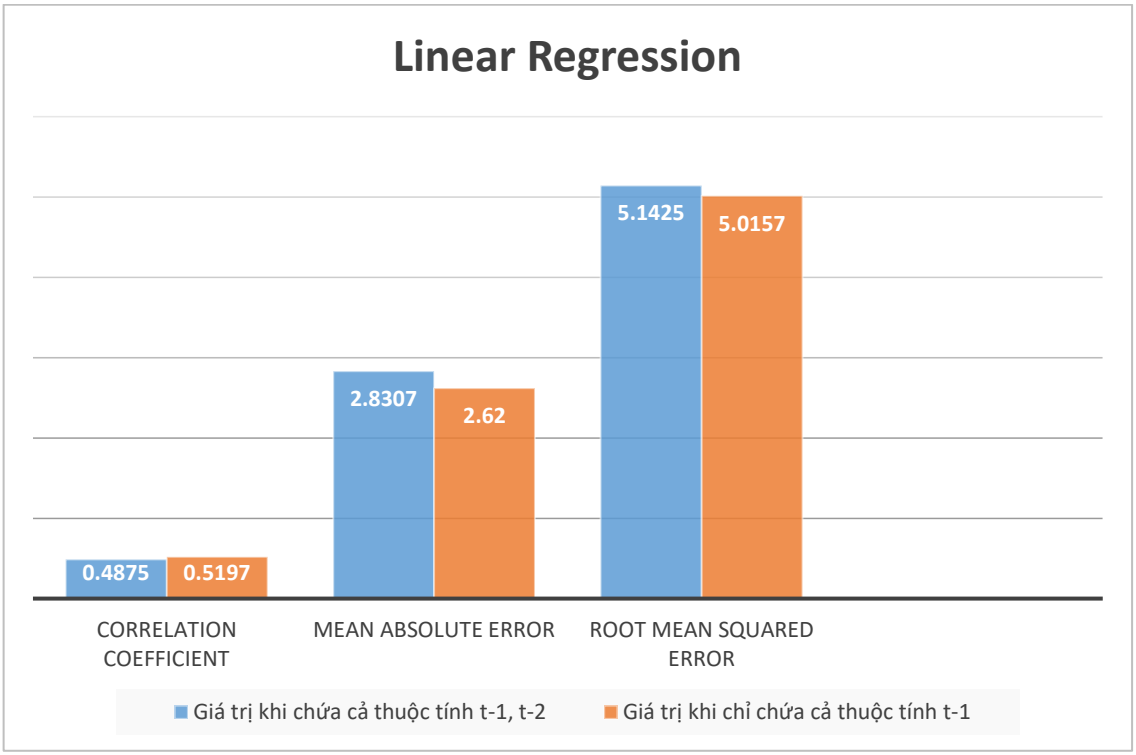
Biểu đồ 2.1: Kết quả so sánh lọc đặc trưng cho mô hình huyện Ba Vì



Biểu đồ 2.2: Kết quả so sánh lọc đặc trưng cho mô hình huyện Chương Mỹ

Từ kết quả lọc đặc trưng cho thấy, các thuộc tính dịch tả lân cận và dịch tả tại quận huyện xem xét ở tháng t-1, một số biến khí hậu cho giá trị p-value nhỏ hơn.

Áp dụng thuật toán hồi quy tuyến tính cho mô hình gộp 29 quận huyện khi chứa cả các thuộc tính t-1, t-2 và khi chỉ có thuộc tính t-1. Kết quả thể hiện trong biểu đồ 2.3



Biểu đồ 2.3: Kết quả đánh giá mô hình áp dụng hồi quy tuyến tính

Biểu đồ 2.3 cho thấy hệ số tương quan của mô hình đạt được tương đối. Đồng thời, sau khi bỏ các thuộc tính t-2 kết quả mô hình đạt được cao hơn: Hệ sô tương quan dương tăng, các độ đo lỗi giảm

2.2.4.1.Kết quả mô hình cục bộ

Mô hình dự báo cho 29 quận /huyện ở Hà nội. Kết quả thực nghiệm cho 2 quận điển hình được thể hiện ở bảng sau:

Bảng 2.4: Kết quả mô hình cho hai quận điển hình Đống Đa và Hoàng Mai

Quận/Huyện	Các độ đo	Linear Regression	NaiveBayes	LibSVM	RandomForest
Đống Đa	CC	**-0.0713**			
	MAE	22.8332	0.2504	0.2222	0.333
	RMSE	26.5469	0.4741	0.4714	0.5774
	Precision		0.583	0.444	0.7220
	Recall		0.667	0.667	0.5000
	F-Measure		**0.611**	0.533	0.5280
Hoàng Mai	CC	**0.5317**			
	MAE	12.7367	0.2227	0.2222	0.222
	RMSE	13.8483	0.453	0.4714	0.4714
	Precision		0.444	0.444	0.5830
	Recall		0.667	0.667	0.6670
	F-Measure		0.533	0.533	**0.6110**

Độ đo đánh giá mô hình kết quả cho các quận-huyện nằm trong vùng dịch tả là khá thấp trong khoảng từ 0.6 và 0.758. Giá trị hệ số tương quan dường như bị tách biệt. Trong một số trường hợp, giá trị tuyệt đối là rất nhỏ, cho biết không có sự tương quan giữa biến mục tiêu và biến điều kiện. Nhưng cũng có một số trường hợp có giá trị tuyệt đối cao và có sự tương quan giữa biến mục tiêu và biến điều kiện (Xem chi tiết phụ lục 2)

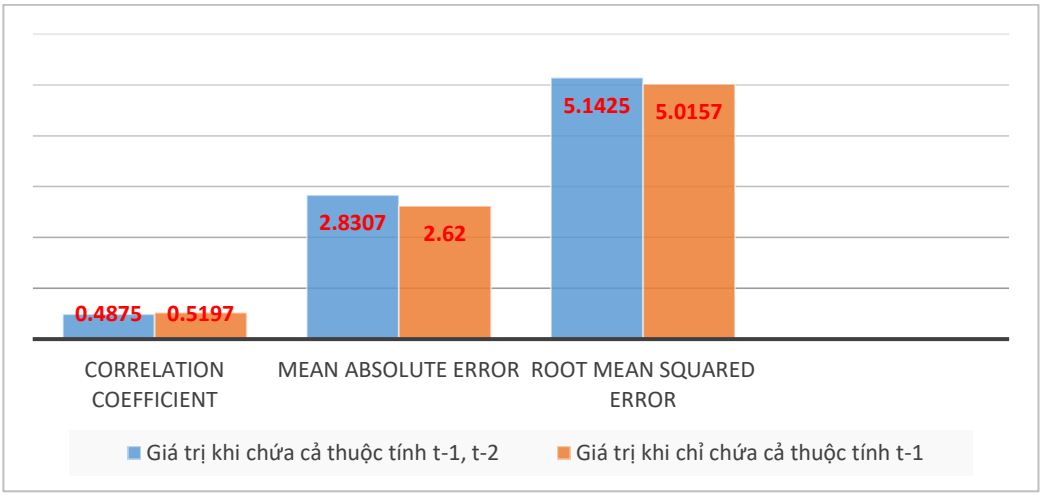
2.2.4.2.Kết quả mô hình toàn cục

Với mô hình toàn cục có ba thực nghiệm đã được tiến hành.

- Thực nghiệm thứ nhất kết hợp trạng thái khí hậu và dịch tả để làm các biến điều kiện sử dụng cho mô hình.
- Thực nghiệm thứ hai chỉ có biến khí hậu được sử dụng cho mô hình.
- Thực nghiệm cuối cùng chỉ có biến trạng thái dịch tả được sử dụng cho mô hình.

Trong mỗi thực nghiệm, thì biến số thời gian "t-12" chỉ dẫn kết quả khi lấy dữ liệu trong quá khứ của hai tháng t-1 và t-2, trong khi "t-1" chỉ dẫn kết quả khi lấy dữ liệu trong quá khứ trong một tháng t-1;

Kết quả thực nghiệm trong trường hợp kết hợp các biến điều kiện khí hậu và dịch tả với mô hình hồi qui được thể hiện trong bảng và viểu đồ sau:



Biểu đồ 2.4 Kết quả hồi qui trong trường hợp kết hợp các biến điều kiện

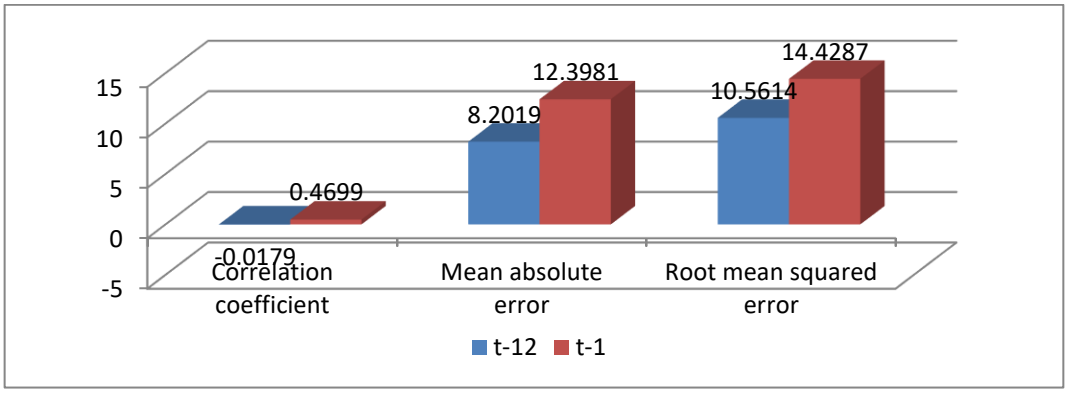
Và bảng 2.5 thể hiện với các bộ phân lớp Naïve Bayes, SVM, RandomForest.

Bảng 2.5 Kết quả mô hình với các bộ phân lớp

Số lớp	Các độ đo	NaiveBayes		LibSVM		RandomForest	
		t-12	t-1	t-12	t-1	t-12	t-1
Hai lớp {0,1}	MAE	0.1406	**0.0958**	**0.0958**	**0.0958**	0.1145	0.1860
	RMSE	0.2913	0.3095	0.3095	0.3095	0.3082	0.3935
	Precision	0.7190	0.7330	0.7330	0.7330	0.7330	0.7190
	Recall	0.7010	0.8560	0.8560	0.8560	0.8560	0.7010
	F-Measure	0.7100	**0.7900**	**0.7900**	**0.7900**	**0.7900**	0.7100
Ba lớp {0,1,2}	MAE	0.1558	**0.1437**	**0.1437**	**0.1437**	0.1718	0.2354
	RMSE	0.3791	0.3790	0.3790	0.3790	0.3775	0.4223
	Precision	0.7330	0.7330	0.7330	0.7330	0.7330	**0.8890**
	Recall	0.8560	0.8560	0.8560	0.8560	0.8560	0.7760
	F-Measure	0.7900	0.7900	0.7900	0.7900	0.7900	**0.8060**

Cả hai trường hợp, hệ số tương quan trong khoảng 0.5 và độ đo đánh giá mô hình trong khoảng 0.8 và MAE từ 0.1 tới 0.2. Thuật toán RandomTree là thuật toán tốt nhất trong thực nghiệm phân ba lớp {0,1,2}.

Kết quả thực nghiêm hồi qui khi kết hợp với biến điều kiện chỉ là khí hậu thể hiện trong hình sau:



Biểu đồ 2.5: Kết quả hồi qui trong trường hợp biến điều kiện chỉ là khí hậu

Kết quả thực nghiệp phân lớp với biến điều kiện chỉ là khí hậu thể hiện trong bảng 2.6

Bảng 2.6 Kết quả mô hình phân lớp khi biến điều kiện chỉ là khí hậu

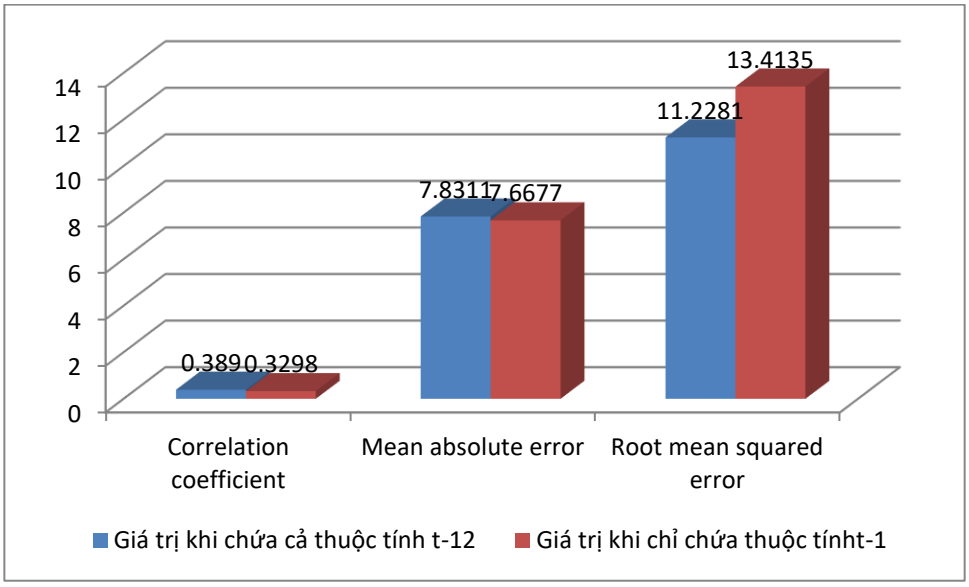
Số lớp	Các độ đo	NaiveBayes		LibSVM		RandomForest	
		t-12	t-1	t-12	t-1	t-12	t-1
Hai lớp {0,1}	MAE	**0.0958**	**0.0975**	**0.0958**	**0.0958**	0.1315	0.1900
	RMSE	0.3095	0.3093	0.3095	0.3095	0.3261	0.3973
	Precision	0.7330	0.7330	0.7330	0.7330	0.7330	0.7190
	Recall	0.8560	0.8560	0.8560	0.8560	0.8560	0.7010
	F-Measure	**0.7900**	**0.7900**	**0.7900**	**0.7900**	**0.7900**	0.7100
Ba lớp {0,1,2}	MAE	0.1437	0.1437	0.1437	0.1437	0.3363	0.5660
	RMSE	0.3790	0.3790	0.3790	0.3790	0.5322	0.7178
	Precision	0.7330	0.7330	0.7330	0.7330	0.7200	0.6330
	Recall	0.8560	0.8560	0.8560	0.8560	0.7010	0.3790
	F-Measure	**0.7900**	**0.7900**	**0.7900**	**0.7900**	0.7100	0.4700

Hệ số tương quan trong trường hợp t-2 là -0.0179 (không tương quan) và hệ số tương quan trong trường hợp t-1 là 0.4699 (tương quan trung bình). Các độ đo đánh giá mô hình có giá trị trong khoảng gần 0.8. Thuật toán RandomForest là thuật toán kém hiệu quả hơn trong mô hình ba lớp {0,1,2}

Kết quả thực nghiệm với biến điều kiện chỉ là trạng thái dịch tả

Bảng 2.7 Kết quả phân lớp khi biến điều kiện chỉ là trạng thái dịch tả

Số lớp	Các độ đo	NaiveBayes		LibSVM		RandomForest	
		t-12	t-1	t-12	t-1	t-12	t-1
Hai lớp {0,1}	MAE	0.5225	0.1393	**0.0958**	**0.0958**	0.1138	0.2041
	RMSE	0.7178	0.3336	0.3095	0.3095	0.2471	0.3765
	Precision	0.8400	0.8260	0.7330	0.7330	0.8760	0.7260
	Recall	0.2070	0.8280	0.8560	0.8560	0.8790	0.7070
	F-Measure	0.2860	**0.8180**	0.7900	0.7900	**0.8750**	0.7170
Ba lớp {0,1,2}	MAE	0.6515	0.1759	**0.1437**	**0.1437**	0.1853	0.3127
	RMSE	0.7825	0.3182	0.3790	0.3790	0.2941	0.4026
	Precision	0.8510	0.7330	0.7330	0.7330	0.9150	0.8520
	Recall	0.2990	0.8560	0.8560	0.8560	0.9080	0.5980
	F-Measure	0.3100	0.7900	0.7900	0.7900	**0.9110**	0.6560



Biểu đồ 2.6 Kết quả hồi qui khi biến điều kiện chỉ là trạng thái dịch tả

Kết luận:

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 sau đây:

- Nghiên cứu cho kết quả biểu diễn dữ liệu toàn cục tốt hơn cục bộ.
- Tồn tại sự tương quan giữa các biến điều kiện khí hậu với biến mục tiêu trạng thái dịch tả trong nhiều trường hợp (hệ số tương quan trên 0.3, thậm chí có trường hợp giá trị này xấp xỉ 1.0). Khi xem xét các biến điều kiện chỉ bao gồm các yếu tố khí hậu của tháng hiện thời (Bảng 2.6) thì hệ số tương quan cũng xấp xỉ 0.47.
- 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 [25], [32],[105] 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 với độ chính xác cao.

2.2.5 Hiệu chỉnh mô hình dự báo với dữ liệu không cân bằng

Đặc điểm của dữ liệu ca bệnh tả tại Hà Nội là không cân bằng, số lượng các ca bệnh tả chỉ chiếm một phần nhỏ trong toàn bộ dân số. Bài toán phân lớp dữ liệu không cân bằng là một trong những vấn đề khó đang được cộng đồng nghiên cứu học máy và khai phá dữ liệu quan tâm [78]. Vấn đề không cân bằng lớp thường xảy ra với bài toán phân lớp mà ở đó lớp được quan tâm chiếm tỉ lệ rất nhỏ so với lớp còn lại. Trong thực tế, sự không cân bằng lớp ảnh hưởng lớn đến hiệu quả của các mô hình phân loại. Với các tập dữ liệu của các bài toán phân lớp như vậy sẽ làm cho các mô hình học phân lớp gặp nhiều khó khăn trong dự báo cho dữ liệu lớp thiểu số. Hầu hết giải thuật học như cây quyết định C4.5[51], CART [56], SVM [93] đều được thiết kế để

cho độ chính xác tổng thể, không quan tâm đến bất kỳ lớp nào. Chính vì lý do này, các giải thuật phân lớp cho tập dữ liệu không cân bằng gặp phải vấn đề dự báo đó là làm mất lớp thiểu số mặc dù chúng cho độ chính xác phân lớp tổng thể rất cao.

Nhiều giải pháp đã được đề xuất để giải quyết vấn đề trên trong đó có những giải thuật học cây quyết định nhằm cải thiện hiệu quả dự báo lớp thiểu số nhưng không làm giảm hiệu quả dự báo lớp đa số. Có thể liệt kê các giải pháp theo hướng này bao gồm: các phương pháp thay đổi phân bố dữ liệu, phương pháp lấy mẫu tăng thêm cho lớp thiểu số, lấy mẫu giảm cho lớp đa số đã được đề xuất [47], [50], [76], [101] hoặc chiến lược can thiệp trực tiếp giải thuật cây quyết định, đề xuất thay đổi hàm phân hoạch dữ liệu nhằm cải thiện dự báo lớp thiểu số nhưng không làm mất nhiều dự báo lớp đa số [75] hay đề xuất gán giá phải trả cho dự báo sai của các lớp khác nhau (giá của lớp thiểu số lớn hơn giá của lớp đa số)[41], [74]. Ngoài ra, cũng có những phương pháp đề xuất điều chỉnh ước lượng xác suất tại nút lá của cây nhằm cải thiện dự báo lớp thiểu số [74].

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 2.8. 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. Kết quả trên Bảng 2.13 cũng cho thấy thuật toán Random Forest 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

Chương này đã mô tả việc thiết lập mô hình dự báo dịch bệnh bằng các phương pháp khai phá luật kết hợp, phương pháp học máy với các kỹ thuật hồi qui và phân lớp thông qua hai hình thức biểu diễn cục bộ và toàn cục.

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, mô hình hồi qui tuyến tính cho hệ số tương quan thấp trong hầu hết các trường hợp vì vậy cần tìm kiếm một giải pháp hồi qui phù hợp hơn. 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. Mặc dù có sự khác biệt trong kết quả ở biểu diễn toàn cục và biểu diễn cục bộ, các

kết quả thực nghiệm nói chung cho thấy các mô hình dự báo đề xuất đều cho kết quả phù hợp và có khả năng dự báo được.

Các kết quả nghiên cứu trong chương này đã được đăng trong Kỷ yếu của hội nghị quốc tế 8th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2016) tại Đà Nẵng- Việt Nam, Tạp chí khoa học công nghệ Đại học Đà nẵng và Kỷ yếu hội thảo quốc gia 2015 về điện tử, truyền thông và công nghệ thông tin (ECIT2015).

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

Chương này đề xuất mô hình dự báo ngắn hạn có xem xét toàn diện mức độ ảnh hưởng của các yếu tố khí hậu và địa lý đến số ca mắc tả ở Hà nội dựa trên kỹ thuật hồi qui Random Forest. Cụ thể, chương này thực hiện phân rã dữ liệu đầu vào không cân bằng theo phương pháp cửa sổ trượt để dự báo và đánh giá mức độ ảnh hưởng của các yếu tố khí hậu, không gian địa lý và thời gian lên mô hình dự báo.

3.1 Xây dựng mô hình dự báo dịch tả ngắn hạn

Kết quả của nhiều nghiên cứu về dịch tả đã khẳng định nguyên nhân bùng phát dịch tả phụ thuộc vào một số yếu tố chính, bao gồm: vị trí địa lý, nhiệt độ, độ ẩm, lượng mưa, mức nước sông, mức nước biển, nhiệt độ bề mặt nước biển và chỉ số dao động phía Nam..., [14], [18], [21], [30], [39],[60], [63], [98]. Trong lĩnh vực quản lý y tế và dự phòng dịch tả , việc dự báo dịch trong ngắn hạn (theo ngày) trong giai đoạn chớm bùng phát dịch là rất cần thiết và hữu ích cho việc bố trí bệnh viện, thuốc và các phương tiện điều trị khác [6]. Trong những năm gần đây, sự sẵn có và ngày càng tăng nguồn dữ liệu khí hậu từ các cảm biến từ xa và những dữ liệu phân tích lại, cũng như sự phát triển trong việc dự báo về biến đổi khí hậu, đã mang lại cơ hội mới cho phân tích và dự báo dịch bệnh.

Dựa trên kết quả khả quan của mô hình dự báo sử dụng phương pháp học máy hồi qui đã trình bày ở chương 2 của luận án, nghiên cứu tập trung vào việc xây dựng mô hình dự báo ngắn hạn với các biến đầu vào là các tham số khí hậu, thời tiết và biến đầu ra là số ca bệnh tả tại từng quận huyện trên địa bàn Hà Nội sử dụng phương pháp học máy hồi qui Random Forest. Bên cạnh đó nghiên cứu cũng phân tích mức độ ảnh hưởng của các yếu tố thời tiết và hệ số giao động phía Nam (SOI) lên số ca mắc tả trong giai đoạn 2001-2012, cũng như đánh giá độ quan trọng của các yếu tố khí hậu và không gian địa lý trong mô hình dự báo.

Các dữ liệu sử dụng cho thử nghiệm này là các tập dữ liệu đã được mô tả ở Chương 1 của luận án. Do dữ liệu các ca tả phân bố không đồng đều và chỉ xuất hiện trong 5 năm, nghiên cứu đã 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.

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	**Nhóm biến lân cận**
- Độ ẩm trung bình ngày	Các biến liên quan số ca mắc tả của
- Nhiệt độ trung bình ngày	các quận/huyện lân cận: Số ca mắc
- Lượng mưa ngày	tả của các quận/huyện lân cận trong
- Số giờ nắng ngày	0, 1, 2, …, *d* ngày trước đó
- Tốc độ gió theo ngày	
- Chỉ số dao động phía Nam SOI (theo ngày)	

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

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.

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 FS theo phương pháp cửa sổ trượt. Đây là phương pháp đã được chứng minh là phù hợp với các bài toán chuỗi thời gian [84]. Theo phương pháp này, ban đầu sẽ tạo ra một cửa sổ s_1 tương ứng với tập dữ liệu huấn luyện ban đầu. Với tập dữ liệu kiểm thử lựa chọn cửa sổ s_2. Chú ý rằng ở mỗi điểm dữ liệu trong tập huấn luyện bao gồm tất cả các biến đầu vào và đầu ra, còn mỗi tập dữ liệu kiểm thử sẽ chỉ bao gồm các biến dự báo. Khung cửa sổ sẽ được trượt dọc theo trục thời gian cho đến khi không còn dữ liệu.

Bảng 3.1: Mô tả mô hình dự báo với các nhóm biến đầy đủ, độc lập với khí hậu, độc lập với địa lý

Nhóm dự báo	Mô hình		
	DD	DLKH	DLDL
Dữ liệu về khí hậu	Nhiệt độ trung bình Độ ẩm trung bình Lượng mưa Chỉ số SOI Số giờ nắng Tốc độ gió		Nhiệt độ trung bình Độ ẩm trung bình Lượng mưa Chỉ số SOI Số giờ nắng Tốc độ gió
Dữ liệu địa lý không gian lân cận	Số lượng ca bệnh tả trong quận *D* Số lượng ca bệnh tả của các quận lân cận quận *D*	Số lượng ca bệnh tả trong quận *D*. Số lượng ca bệnh tả của các quận lân cận quận *D*	

Mô hình được xây dựng trong sự chuyển dịch và cải thiện dọc theo trục thời gian. Chọn kích thước các cửa sổ trượt $s_1=s_2=l$ cho tất cả các mô hình. Độ trễ thời gian của mô hình được lựa chọn là d=3, 7, 14 hoặc 30 ngày, trong đó cửa sổ trượt có kích cỡ cố định ban đầu là d=3, 7, 14, 30. Chuỗi thời gian được sử dụng để kiểm thử tương ứng là n=3, 7, 14, 30.

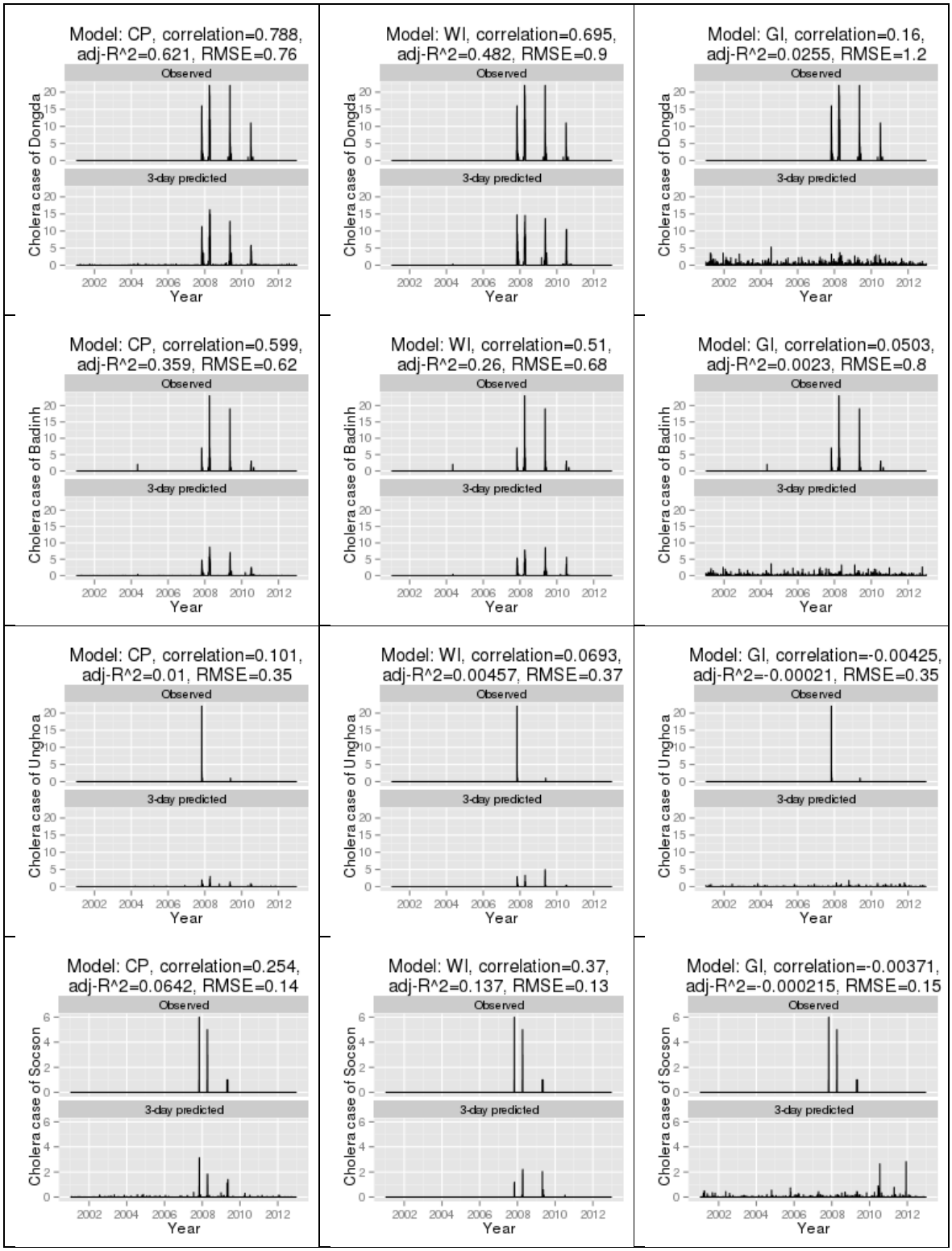
<table>
<tr><td>m1</td><td>m2</td><td>m3</td><td>m4</td><td>m5</td><td>m6</td><td>m7</td><td>m8</td></tr>
<tr><td>c4</td><td>c5</td><td>c6</td><td>c7</td><td>c8</td><td>c9</td><td>c10</td><td>c11</td></tr>
<tr><td>n1</td><td>n2</td><td>n3</td><td>n4</td><td>n5</td><td>n6</td><td>n7</td><td>n8</td></tr>
<tr><td colspan="3">Dữ liệu huấn luyện 1</td><td colspan="3">Dữ liệu kiểm thử 1 (dự báo)</td><td></td><td></td></tr>
<tr><td></td><td colspan="3">Dữ liệu huấn luyện 2</td><td colspan="3">Dữ liệu kiểm thử 2 (dự báo)</td><td></td></tr>
<tr><td></td><td></td><td colspan="3">Dữ liệu huấn luyện 3</td><td colspan="3">Dữ liệu kiểm thử 3 (dự báo)</td></tr>
</table>

Hình 3.1. Minh họa việc huấn luyện mô hình hồi qui RF theo phương pháp cửa sổ trượt có độ trễ thời gian

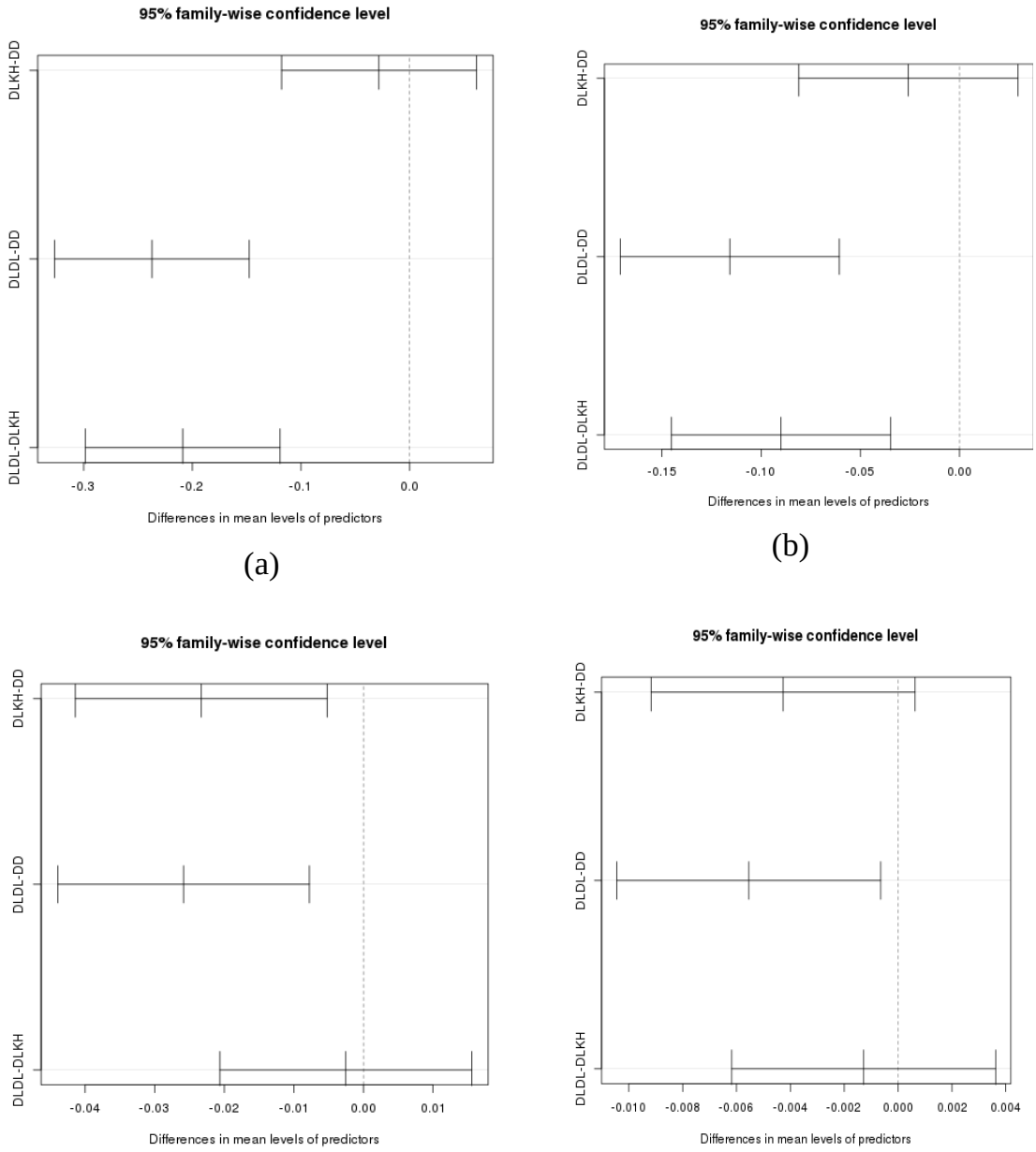
Hình 3.1 minh họa việc huấn luyện mô hình hồi qui RF theo phương pháp cửa sổ trượt với độ trễ thời gian là 3 ngày, kích cỡ cửa sổ trượt là 3 ngày và số ngày dự báo trước là 3 ngày. Giá trị các tham số: *n=3, d=3*. Các ô *m1, m2,..., m8* là các biến khí hậu từ ngày 1 đến ngày 8; các ô *c4, c5,..., c11* là các biến ghi nhận các ca bệnh mắc của quận *C* ở các ngày 4, 5,..., 11; và *n1, n2,..., n8* là số các ca mắc tả ở các quận lân cận của các ngày 1 đến 8. Thời điểm bắt đầu huấn luyện mô hình là ngày 6. Dữ liệu huấn luyện là tập {*m1, m2, m3, n1, n2, n3, c4, c5, c6*}. Dữ liệu kiểm thử là tập {*m4, m5, m6, n4, n5, n6*}. Kết quả kiểm thử (dự báo) là tập {*c7, c8, c9*}. Quá trình này lặp lại cho các ngày 7, 8,...Với dữ liệu 4383 ngày (12 năm) trong giai đoạn nghiên cứu, số lần lặp trong quá trình huấn luyện và kiểm thử là 4377.

3.2 Thực nghiệm và đánh giá mô hình

Nghiên cứu đã xây dựng 29x3 mô hình hồi qui RF cho 29 quận/huyện sử dụng tập dữ liệu FS như mô tả trong mục 3.1. Để đánh giá các mô hình hồi qui, nghiên cứu 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 [4] với 4 khoảng dự báo 3, 7, 14 và 30 ngày. Minh họa kết quả so sánh được biểu diễn trên Hình 3.2.



Hình 3.2. Minh họa so sánh độ chính xác dự báo của ba mô hình với khoảng dự báo là 3 ngày ở các quận Đống Đa,Bai Đình, Ứng Hòa, Sóc Sơn.



Hình 3.3. So sánh ảnh hưởng của nhóm biến khí hậu và nhóm biến lân cận đến độ chính xác của mô hình với độ đo R^2: (a),(b),(c),(d) lần lượt ứng với khoảng dự báo trước là 3,7,14 và 30 ngày.

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.

Tuy nhiên, các kết quả về độ đo RMSE của các mô hình không có sự khác biệt đáng kể. Hơn nữa, việc so sánh độ đo RMSE không cho phép chỉ ra mô hình nào tốt hơn. Do vậy, nghiên cứu chỉ sử dụng độ đo R^2 để so sánh các mô hình.

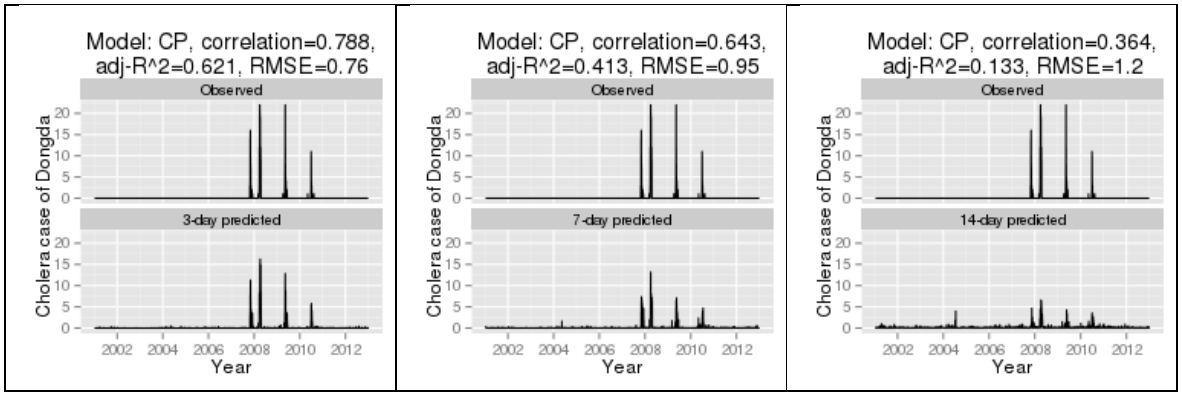
3.3. Mối quan hệ giữa độ chính xác và khoảng thời gian dự báo

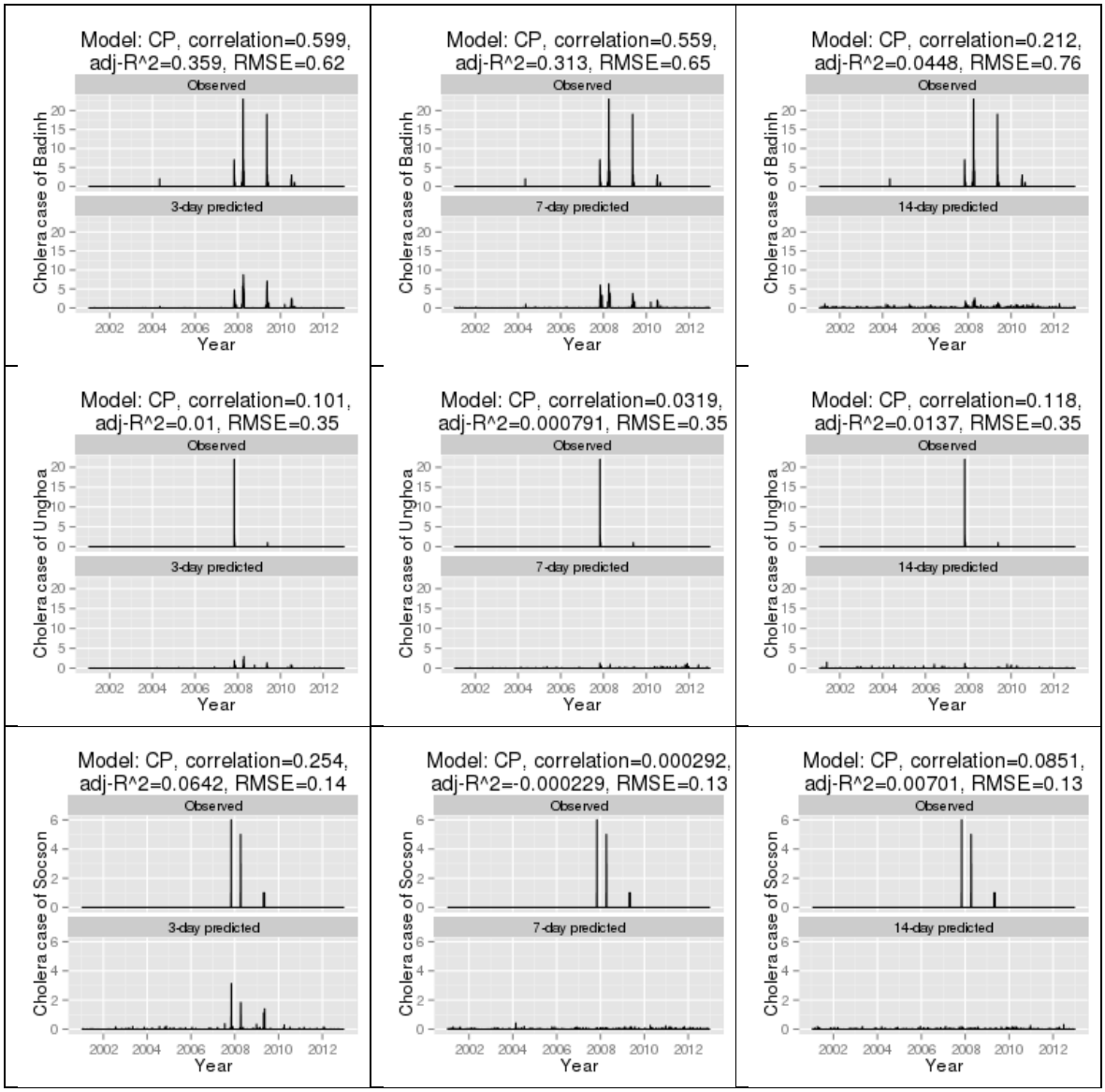
Căn cứ vào kết quả so sánh các mô hình dự báo ở Mục 3.2, có thể kết luận mô hình đầy đủ (DD) là tốt nhất. Trong mục này, nghiên cứu sử dụng mô hình đầy đủ để dự báo với khoảng dự báo là 3,7,14 và 30 ngày để xem xét mối quan hệ giữa độ chính xác và khoảng thời gian dự báo. Cụ thể, kết quả số ca mắc tả 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 5 của luận án.



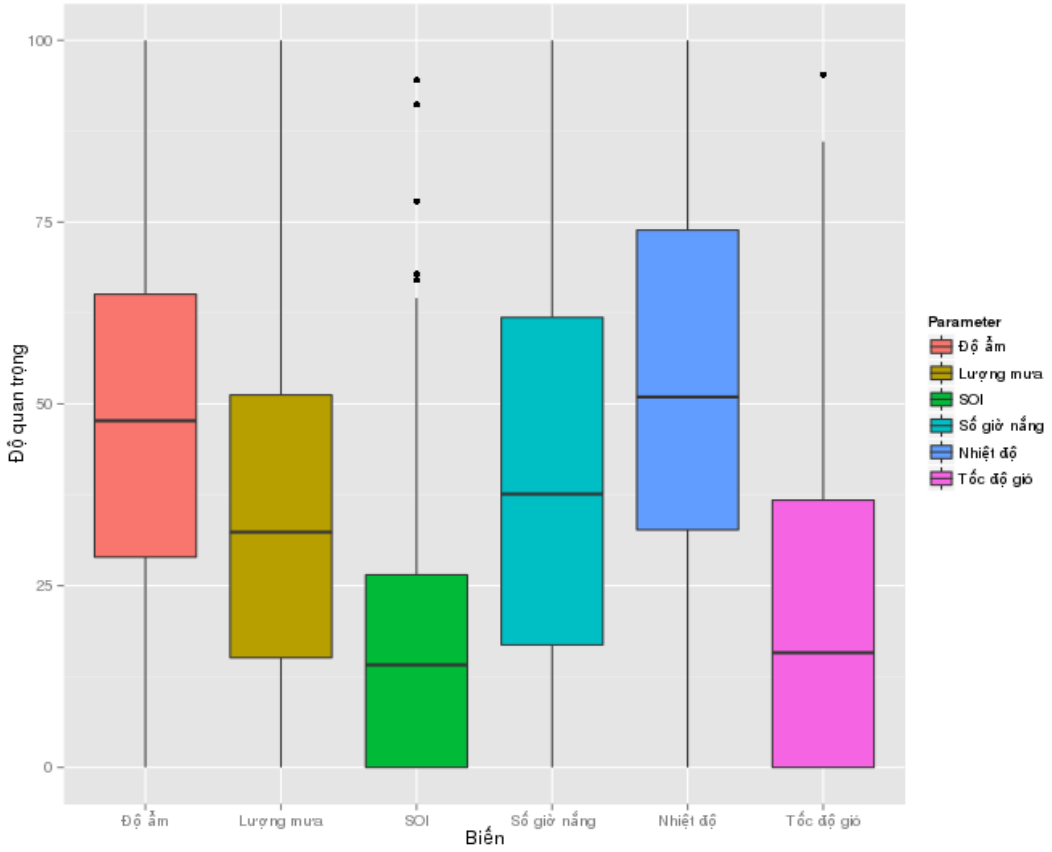


Hình 3.4. So sánh tính chính xác của mô hình Đầy đủ với độ dài dự đoán khác nhau

3.4 Mức độ quan trọng của các biến khí hậu

Từ kết quả xây dựng mô hình dựa báo dựa trên kỹ thuật hồi qui RF, luận án xem xét mức độ quan trọng của các biến khí hậu trong mô hình với từng quận 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. Có thể thấy rằng, các biến nhiệt độ trung bình ngày, độ ẩm trung bình ngày có ảnh hưởng cao nhất (~50%). Biến số giờ nắng có ảnh hưởng thấp hơn hai biến trên (~35%). Các biến lượng mưa, tốc độ gió và SOI có ảnh hưởng thấp (<20%) mặc dù có một vài ngoại lệ ảnh hưởng cao hơn 50% của SOI và tốc độ gió. Kết quả chi tiết

đánh giá độ quan trọng của các biến khí hậu trong mô hình kết hợp đầy đủ các yếu tố khí hậu và địa lý ở mỗi quận với các độ dài khoảng dự báo khác nhau được trình bày ở Phụ lục 3.



Hình 3.5. Mức độ quan trọng của các biến khí hậu trong các mô hình hồi qui RF

3.5. Nhận xét

Chương 3 đã xây dựng 29x3 mô hình hồi qui RF cho dự báo dịch tả ở từng quận/huyện của thành phố Hà Nội trong giai đoạn 2001-2012. Kết quả so sánh, phân tích cho thấy các mô hình đầy đủ cho kết quả dự báo chính xác nhất trong ngắn hạn do có xem xét đến tất cả các yếu tố khí hậu và địa lý. 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, trong đó nhiệt độ và độ ẩm trung bình ngày có mức ảnh hưởng lớn nhất, trong khi tốc độ gió và chỉ số SOI có mức ảnh hưởng thấp nhất. 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.

3.6. Kết luận

Chương 3 đã đề xuất mô hình dự báo ngắn hạn để dự báo nguy cơ bùng phát dịch tả trong giai đoạn 2001-2012 tại thành phố Hà Nội dựa trên kỹ thuật hồi qui RF sử dụng phương pháp cửa sổ trượt và đánh giá mức độ ảnh hưởng của các yếu tố khí hậu và lân cận địa lý đến mô hình. Các kết quả so sánh, phân tích khẳng định vai trò của lân cận không gian trong mô hình dự báo, giúp gia tăng độ chính xác dự báo. Kết quả phân tích mức độ ảnh hưởng của các yếu tố khí hậu giúp đánh giá chính mức độ ảnh hưởng của từng yếu tố khí hậu. Điều này hỗ trợ công tác quản lý, dự phòng bệnh dịch, theo đó tập trung vào các yếu tố khí hậu có độ nhạy cao, giúp tiết kiệm được thời gian và chi phí trong quá trình thu thập thông tin cũng như ra quyết định. Các kết quả nghiên cứu đã được công bố tại Kỷ yếu hội nghị quốc tế khoa học máy tính và ứng dụng toán -ICCSAMA2016 tổ chức tại Cộng hòa Áo năm 2016.

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

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.

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

Luận án xây dựng mô hình dự báo dịch tả trên địa bàn thành phố Hà Nội, có xem xét đến ảnh hưởng của một số biến như khí hậu, diện tích mặt nước, dân số đến số ca bệnh tả trên cơ sở ứng dụng kỹ thuật phân tích hồi qui không gian trong công nghệ GIS. Dữ liệu sử dụng trong chương 4 được thể hiện trong bảng 4.1.

Bảng 4.1 Mô tả các dữ liệu sử dụng trong thực nghiệm

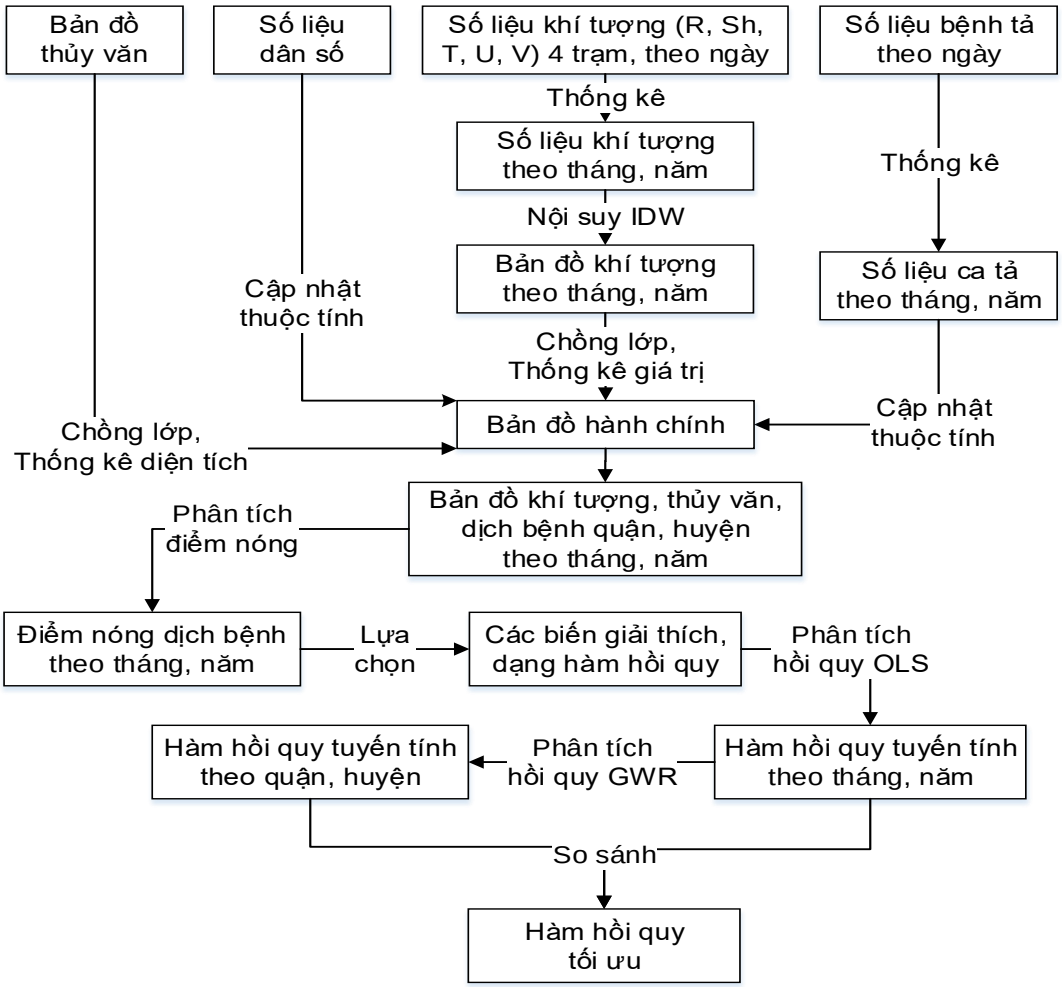
Dữ liệu	Mô tả
Bản đồ hành chính	Thể hiện ranh giới 29 đơn vị quận huyện của Tp. Hà Nội.
Bản đồ thủy văn	Thể hiện mạng lưới sông suối, ao hồ trên địa bàn Tp. Hà Nội.
Số liệu dân số	Giai đoạn 2007 - 2010 trên địa bàn Tp. Hà Nội.
	Thể hiện quy mô dân số các quận huyện theo năm.
Số liệu khí tượng	Giai đoạn 2001 - 2011 tại các trạm đo: Ba Vì, Hà Đông, Hoài Đức, Láng, Sơn Tây trên địa bàn Tp. Hà Nội.
	Thể hiện số liệu đo lượng mưa, nhiệt độ không khí, độ ẩm không khí tương đối, số giờ nắng, tốc độ gió theo ngày.
Số liệu bệnh tả	Giai đoạn 2001 - 2011 trên địa bàn Tp. Hà Nội.
	Thể hiện số liệu lưu trữ thông tin về các ca mắc bệnh tả theo ngày.

Mô hình dự báo đề xuất dựa trên phân tích không gian mô tả trên Hình 4.1 bao gồm các bước xử lý sau:

- **Bước 1:** Tiến hành thu thập dữ liệu đầu vào bao gồm bản đồ hành chính, thủy văn, số liệu dân số, số liệu khí tượng theo ngày (R- lượng mưa; Sh- số giờ nắng; T- nhiệt độ không khí; U- độ ẩm không khí tương đối; V- tốc độ gió) và số liệu ca mắc bệnh tả theo ngày giai đoạn 2001 - 2011.

- **Bước 2:** Từ bản đồ thủy văn, tiến hành chồng lớp với bản đồ hành chính các quận, huyện và thống kê diện tích mặt nước (sông ngòi, ao hồ) cho từng quận, huyện.
- **Bước 3:** Từ số liệu dân số, tiến hành cập nhật vào bản đồ hành chính các quận, huyện theo từng năm.
- **Bước 4:** Dựa trên số liệu khí tượng thu thập, tiến hành thống kê theo tháng (R, Sh, T, U, V lấy giá trị trung bình từng tháng), theo năm (R, Sh lấy giá trị tổng cho từng năm; T, U, V lấy giá trị trung bình cho từng năm) cho từng trạm.
- **Bước 5:** Ứng dụng phương pháp nội suy IDW, thành lập bản đồ khí tượng cho toàn Tp. Hà Nội theo từng tháng, năm. Sau đó, chồng lớp với bản đồ hành chính, thống kê giá trị trung bình tháng, năm từng yếu tố khí tượng cho mỗi quận, huyện.
- **Bước 6:** Dựa trên số liệu ca mắc bệnh tả, thống kê tổng số ca mắc theo từng tháng, năm cho các quận, huyện trong bản đồ hành chính.
- **Bước 7:** Kết quả tổng hợp sau khi thực hiện bước 2, 3, 5, 6 cho ra bản đồ hành chính quận, huyện chứa đựng các thuộc tính: diện tích mặt nước, dân số, các yếu tố khí tượng, số ca bệnh tả theo tháng, năm.
- **Bước 8:** Phân tích điểm nóng dịch bệnh theo tháng, năm sử dụng thống kê Getis-Ord Gi* trên toàn địa bàn Tp. Hà Nội nhằm xác định khu vực thường xuyên xuất hiện ca mắc bệnh. Từ đó, tạo tiền đề cho việc lựa chọn biến giải thích trong mô hình hồi qui dịch bệnh.
- **Bước 9:** Lựa chọn các biến giải thích (diện tích mặt nước, dân số, R, Sh, T, U, V) và dạng hàm hồi qui (logarit), chuẩn bị cho khâu phân tích hồi qui tuyến tính (Ordinary Least Square- OLS).
- **Bước 10:** Tiến hành phân tích hồi qui OLS, thiết lập hàm mô phỏng, dự báo ca bệnh tả theo tháng, năm.
- **Bước 11:** Dựa trên kết quả phân tích hồi qui OLS, tiến hành phân tích hồi qui trọng số không gian (Geographically Weighted Regression- GWR) nhằm thiết lập hàm tuyến tính phù hợp cho từng quận, huyện.

- **Bước 12:** So sánh kết quả từ hai mô hình OLS, GWR sử dụng chỉ số AICc, hệ số xác định hiệu chỉnh R^2. Từ đó lựa chọn mô hình tối ưu.



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

Theo mô hình dự báo dịch tả đề xuất, các thử nghiệm sau sẽ được thực hiện:

- Phân tích điểm nóng dịch tả : Mục tiêu của thực nghiệm này là tìm ra các điểm nóng (Hot Spot) bùng phát dịch tả và mối quan hệ giữa sự bùng phát dịch tả với các yếu tố không gian bao gồm: khí hậu, thủy văn (mặt nước) và mật độ dân số;

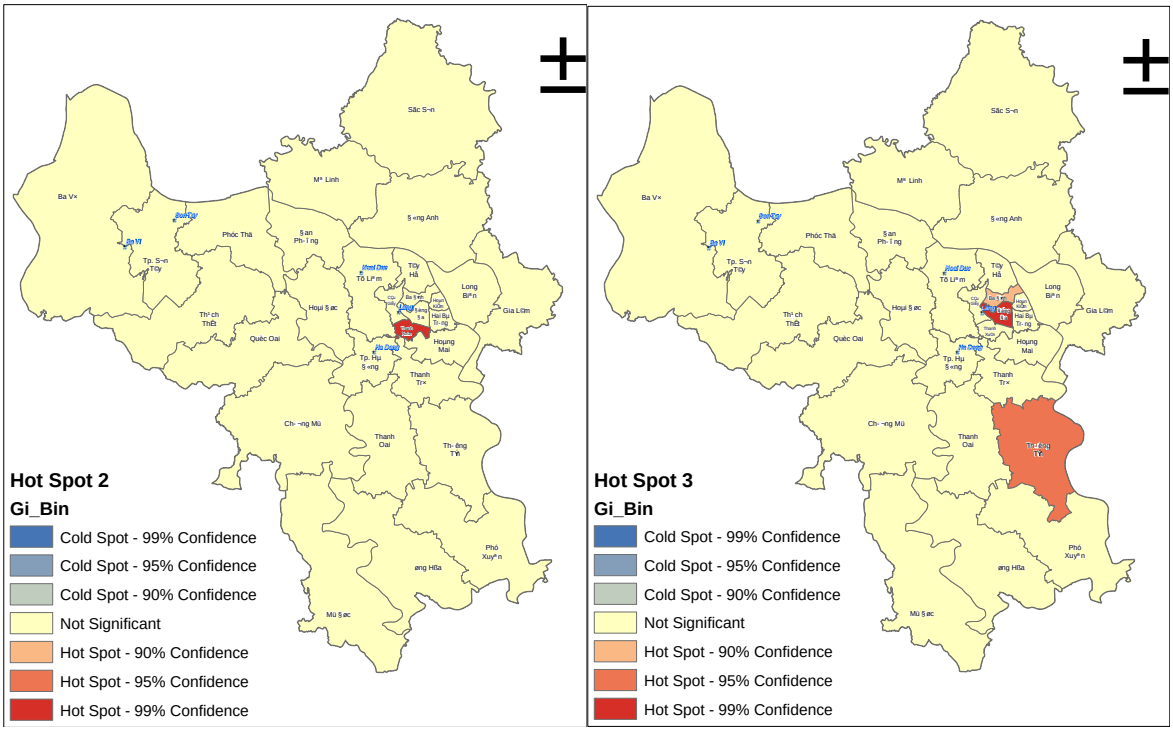
- Thử nghiệm các mô hình hồi qui đa biến cho dự báo dịch tả. Thử nghiệm này được thực hiện trên cơ sở kết quả của bước phân tích điểm nóng dịch tả và bao gồm 3 bước: (i) Lựa chọn biến giải thích phát sinh ca bệnh tả, (ii) Phân tích hồi qui tuyến tính OLS và (iii) Phân tích hồi qui trọng số không gian GWR.

4.2. Kết quả thực nghiệm

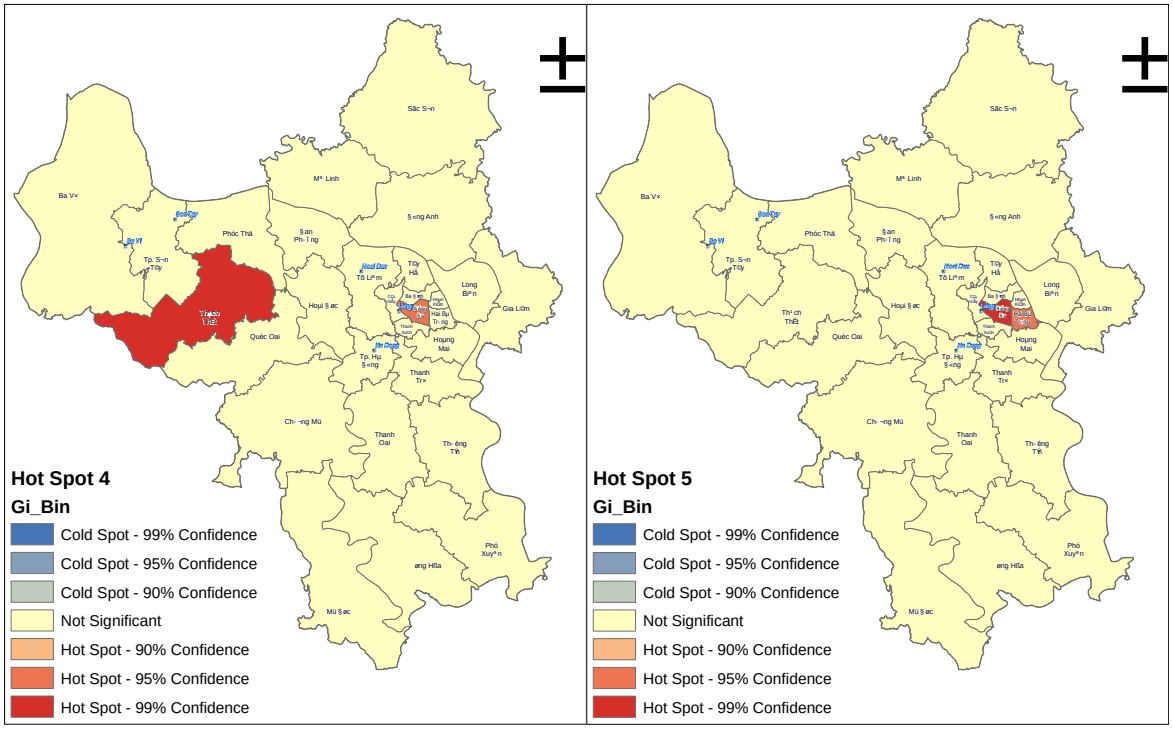
4.2.1. Phân tích điểm nóng dịch tả

4.2.1.1. Theo tháng

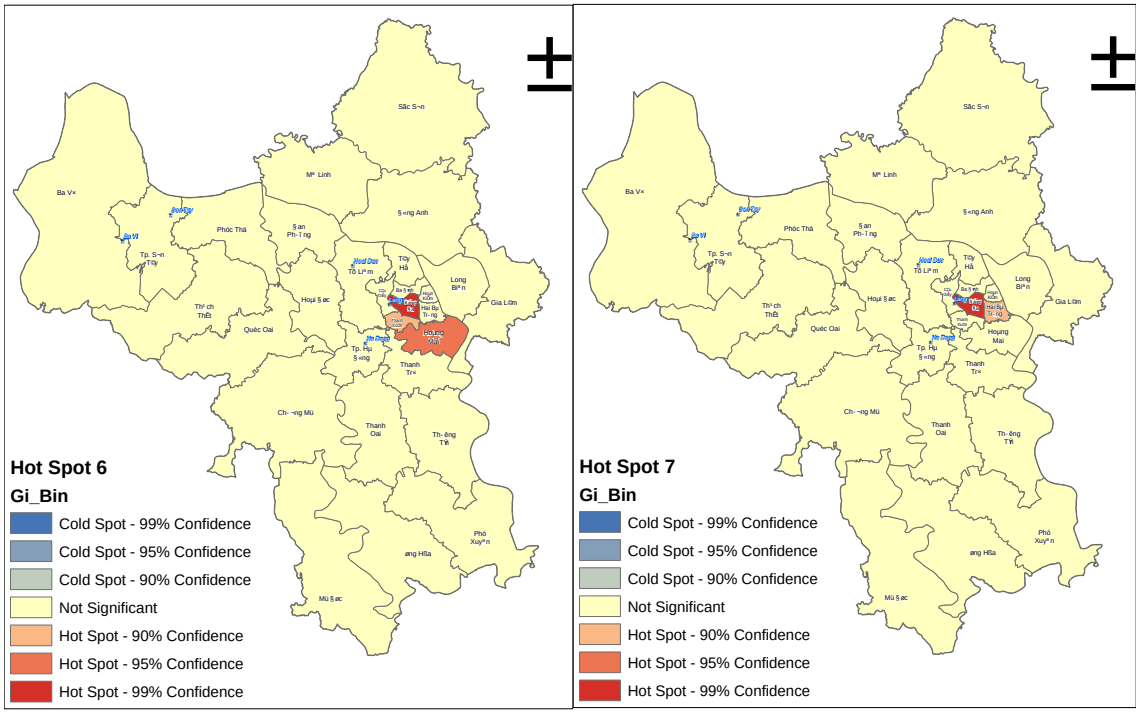
Dựa trên số liệu thống kê số ca bệnh tả theo tháng trong giai đoạn 2004 - 2010, có thể thấy cao điểm của dịch tả xảy ra vào hai khoảng thời gian: tháng 3, 4, 5, 7 (tháng mưa ít) và tháng 10, 11 (tháng mưa nhiều). Tháng có ít số ca bệnh nhất là tháng 1, 2, 8 và 9. Thế nhưng về mặt không gian, câu hỏi đặt ra là các ca bệnh thường xuất hiện ở khu vực nào? các ca bệnh phân bố tập trung thành cụm hay phân tán rải rác toàn vùng? Để trả lời cho hai câu hỏi này, nghiên cứu tiến hành phân tích điểm nóng theo từng tháng với kết quả được thể hiện từ Hình 4.2 đến Hình 4.6. Theo đó, có thể thấy các điểm nóng về số ca bệnh tả thay đổi theo từng tháng, 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.



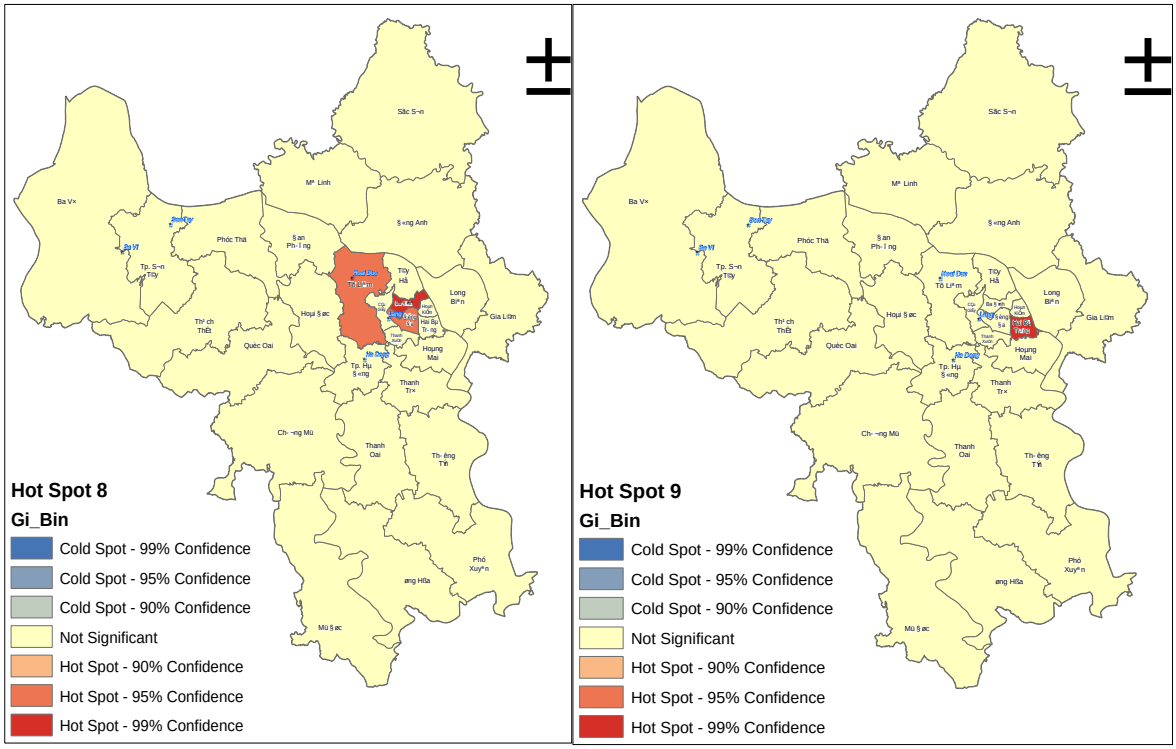
Hình 4.2. Phân tích điểm nóng số ca bệnh tả tháng 2, 3



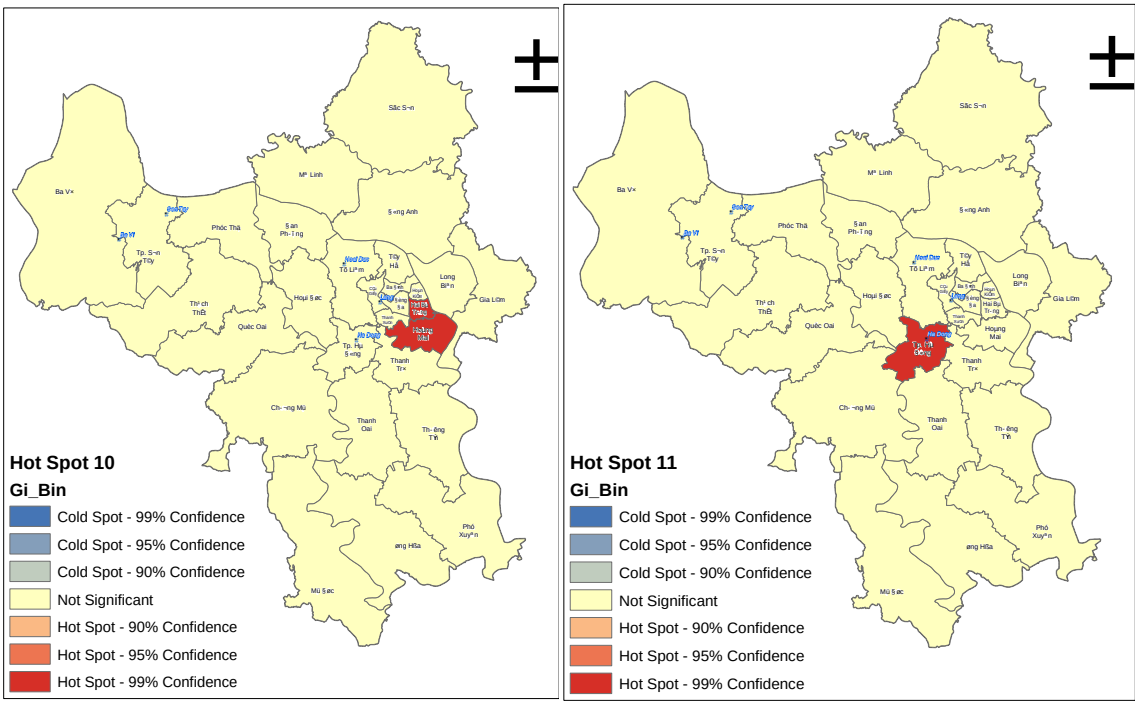
Hình 4.3. Phân tích điểm nóng số ca bệnh tả tháng 4, 5



Hình 4.4. Phân tích điểm nóng số ca bệnh tả tháng 6, 7



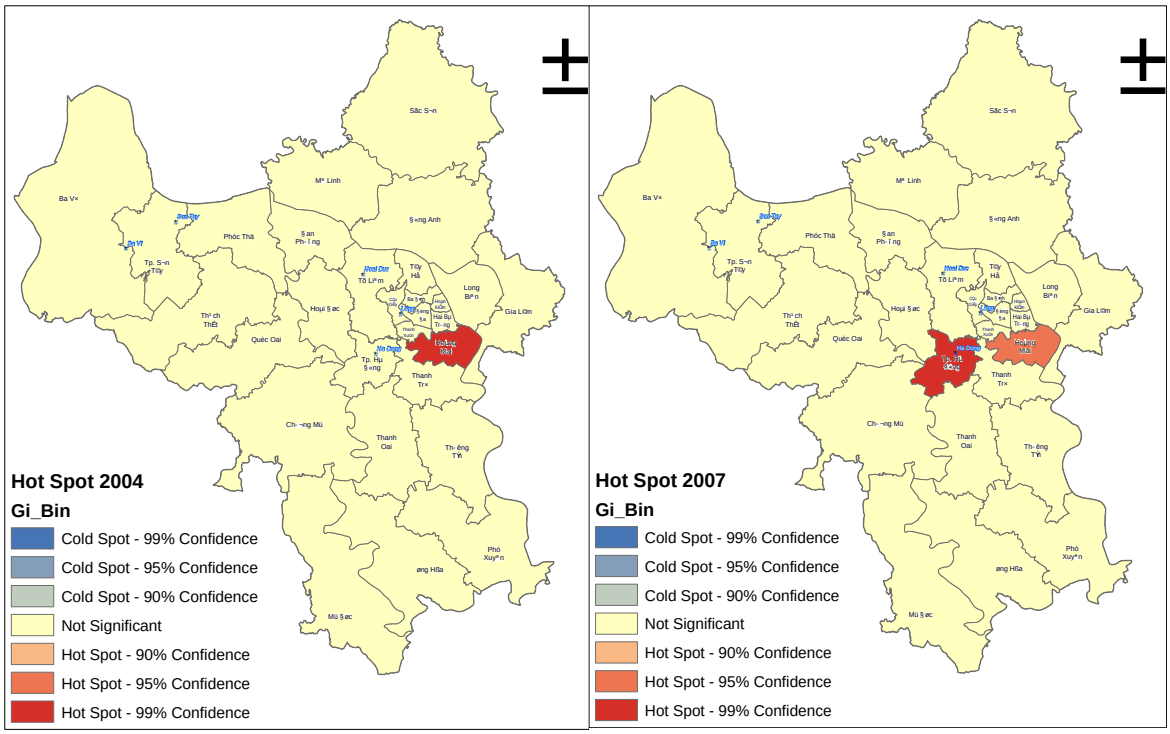
Hình 4.5. Phân tích điểm nóng số ca bệnh tả tháng 8, 9



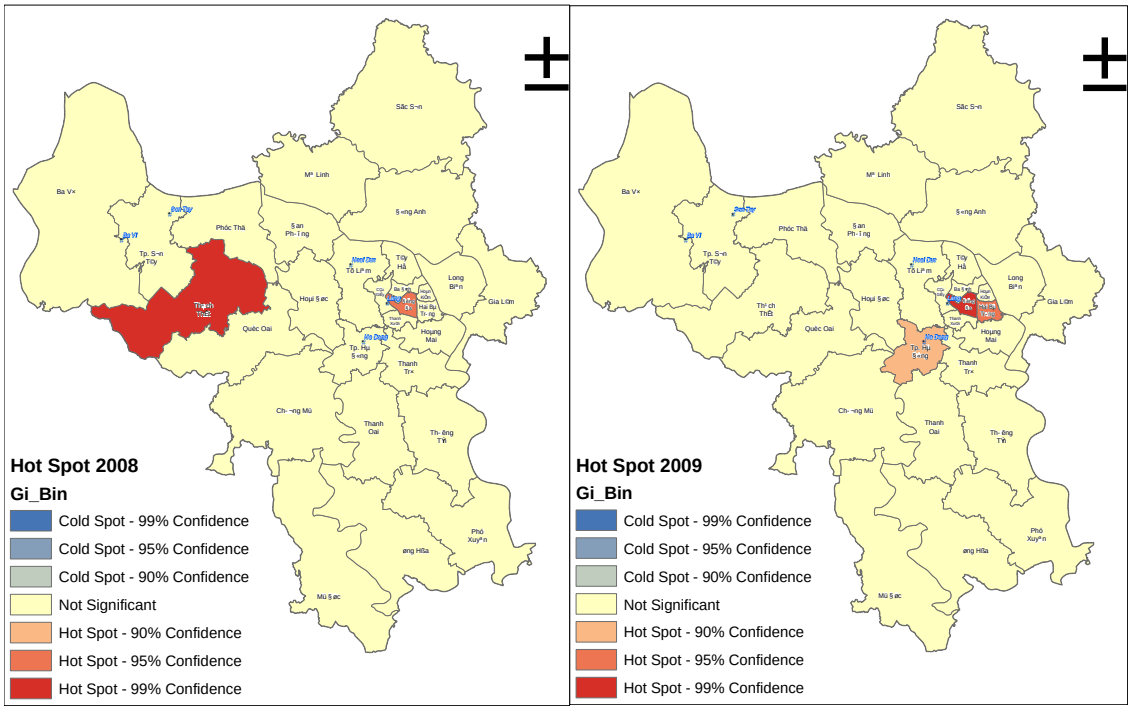
Hình 4.6. Phân tích điểm nóng số ca bệnh tả tháng 10, 11

4.2.1.2. Theo năm

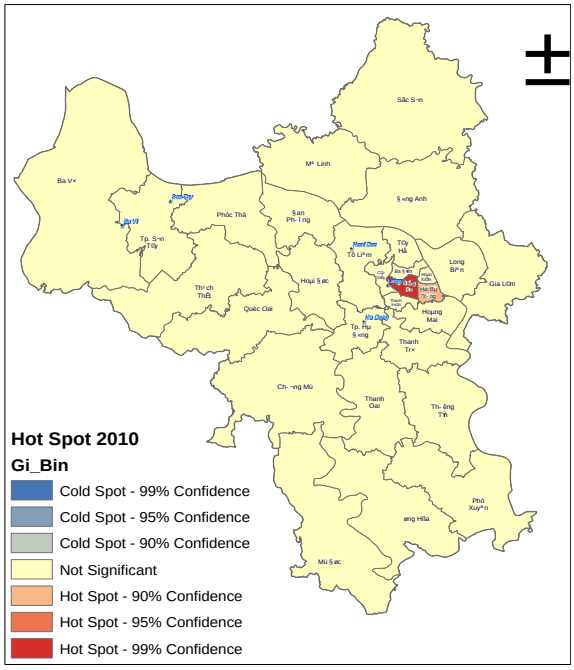
Theo số liệu thống kê số ca bệnh tả xuất hiện trong giai đoạn 2004 - 2010, có thể thấy năm 2004 bắt đầu ghi nhận các ca mắc bệnh tả tại Tp. Hà Nội với số lượng chỉ 25 ca. Sau đó, dịch tả bùng phát liên tục từ năm 2007 đến 2010, đỉnh điểm là 2008. Thế nhưng về mặt không gian, tương tự như trường hợp theo tháng, câu hỏi đặt ra là các ca bệnh thường xuất hiện ở khu vực nào? các ca bệnh phân bố tập trung thành cụm hay phân tán rải rác toàn vùng? Để trả lời cho hai câu hỏi này, nghiên cứu tiến hành phân tích điểm nóng theo từng năm với kết quả được thể hiện từ Hình 4.7 đến Hình 4.9. Theo đó, có thể thấy các điểm nóng về số ca bệnh tả thay đổi theo từng năm. Tuy nhiên, tương tự như tháng, các điểm nóng thường tập trung quanh khu vực nội ô bao gồm các quận Ba Đình, Hoàn Kiếm, Hai Bà Trưng, Thanh Xuân, Đống Đa, Cầu Giấy. Đây là khu vực tập trung dân cư đông đúc, tiếp giáp với sông Hồng về phía Bắc.



Hình 4.7. Phân tích điểm nóng số ca bệnh tả năm 2004, 2007



Hình 4.8. Phân tích điểm nóng số ca bệnh tả năm 2008, 2009



Hình 4.9. Phân tích điểm nóng số ca bệnh tả năm 2010

4.2.2.Xây dựng mô hình hồi qui đa biến dự báo dịch tả trên địa bàn Tp. Hà Nội

4.2.2.1 Lựa chọn biến giải thích phát sinh dịch tả

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 + ε (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 + ε (sai số ngẫu nhiên)

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

4.2.2.2. Mô hình OLS

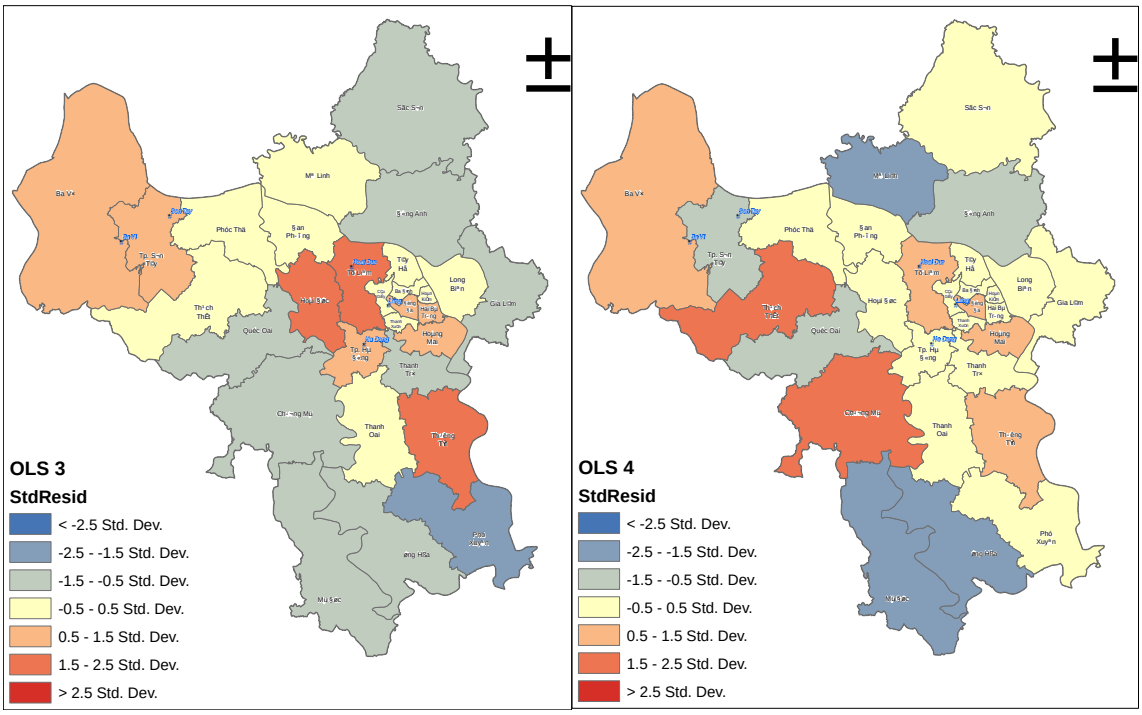
a, Theo tháng: Kết quả phân tích hồi qui OLS cho các tháng 3, 4, 5, 6, 7, 10, 11, 12 được thể hiện trong Bảng 4.2. Độ lệch chuẩn của phần dư (số ca bệnh thực tế - số ca bệnh mô phỏng) được lần lượt tính toán cho các tháng. Đối với các tháng 2, 8, 9, do số ca bệnh rất ít nên nghiên cứu không tìm ra được hàm hồi qui có ý nghĩa thống kê để giải thích sự xuất hiện ca bệnh. Tháng 1 không có ca bệnh nào nên không thiết lập hàm hồi quy. Bảng tổng hợp kết quả phân tích hồi qui OLS cho tháng được thể hiện ở bảng 4.2 và từ Hình 4.10 đến Hình 4.13. Từ các kết quả phân tích hồi qui OLS, có thể rút ra một số nhận xét như sau: (1) Trong các tháng 3, 11, 12, yếu tố khí hậu giải thích được 55%, 38%, 32% số ca bệnh trên toàn khu vực; (2) Trong tháng 4 cao điểm về dịch bệnh, yếu tố mặt nước giải thích được 25% số ca bệnh trên toàn khu vực; và (3) Trong các tháng còn lại bao gồm các tháng 5, 6, 7, 10, sự kết hợp của yếu

tố khí hậu và mặt nước giải thích được 72%, 41%, 57%, 55% số ca bệnh trên toàn khu vực.

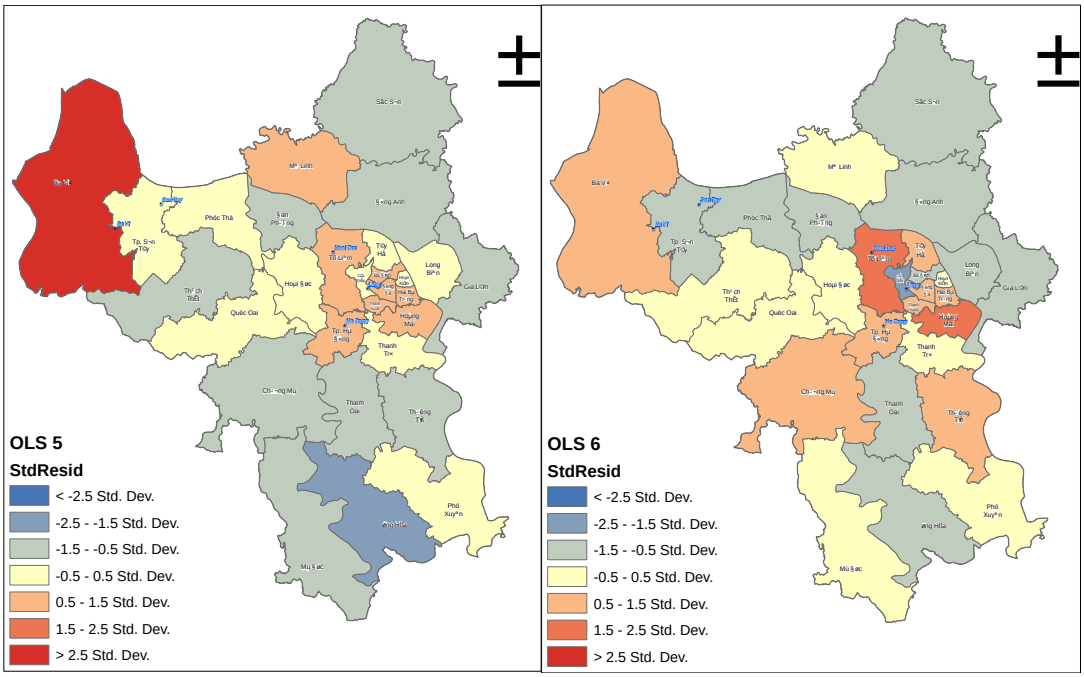
Bảng 4.2. Tổng hợp kết quả phân tích hồi qui OLS theo tháng khu vực Hà Nội

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

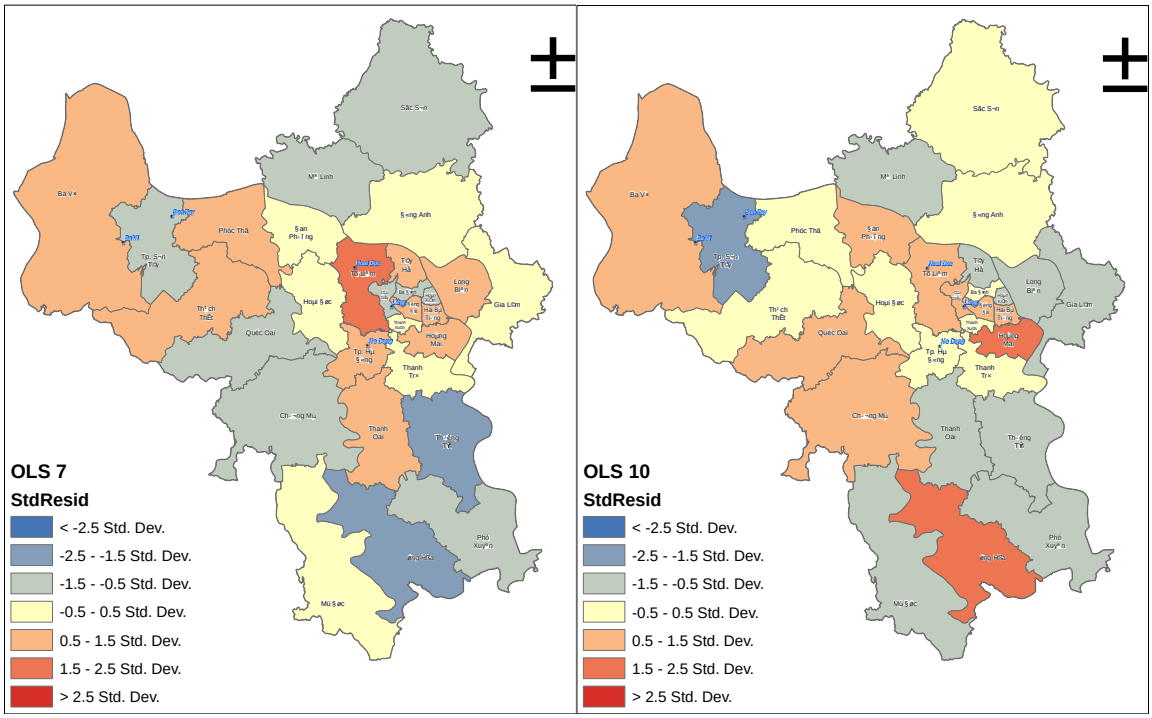
**p_value: giá trị thống kê*



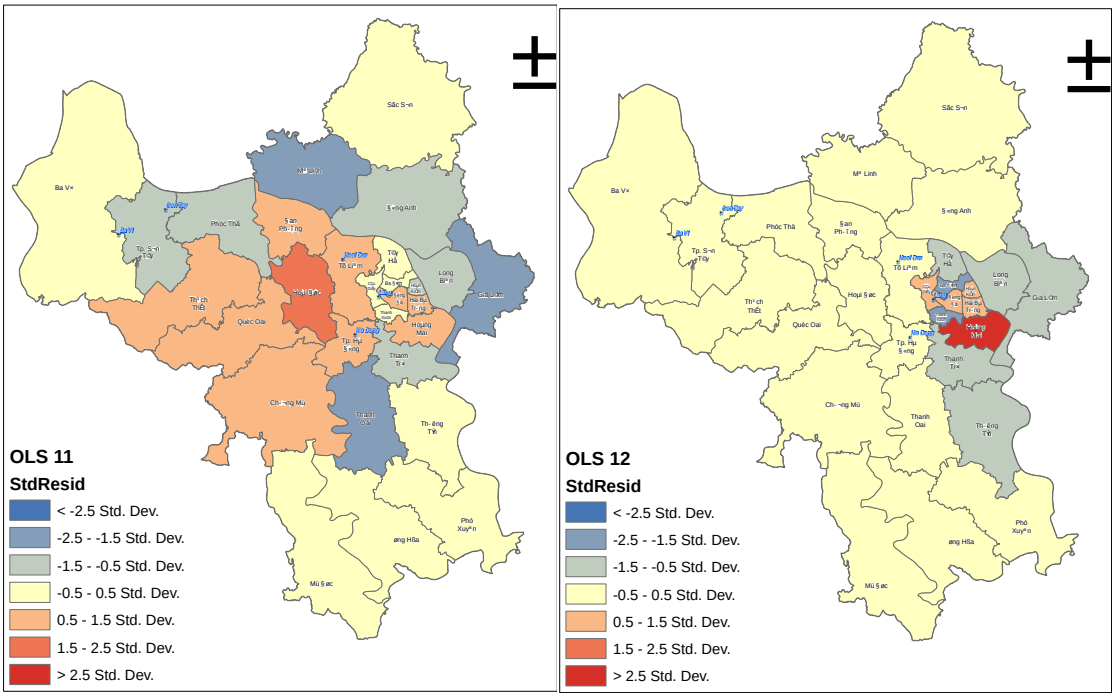
Hình 4.10. Độ lệch chuẩn của phần dư (số ca thực tế - số ca mô phỏng) tháng 3, 4



Hình 4.11. Độ lệch chuẩn của phần dư (số ca thực tế - số ca mô phỏng) tháng 5, 6



Hình 4.12. Độ lệch chuẩn của phần dư (số ca thực tế - số ca mô phỏng) tháng 7, 10



Hình 4.13. Độ lệch chuẩn của phần dư (số ca thực tế - số ca mô phỏng) tháng 11, 12

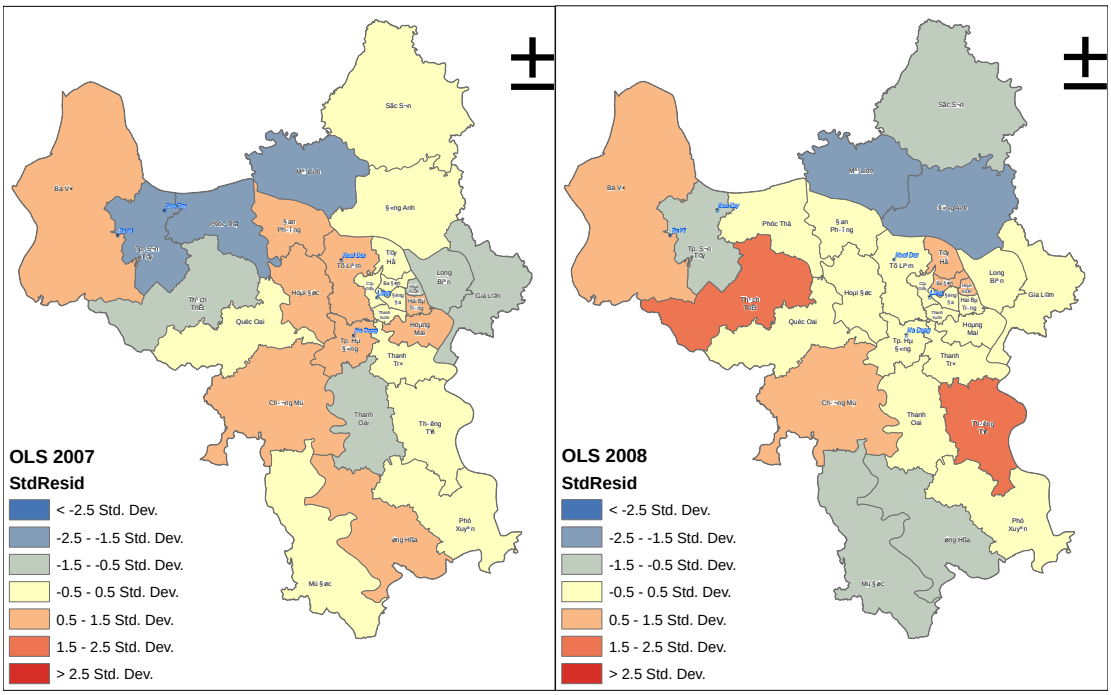
b, Theo năm: Kết quả phân tích hồi qui OLS cho các năm 2007, 2008, 2009, 2010 được thể hiện lần lượt trên bảng 4.3. Độ lệch chuẩn của phần dư (số ca bệnh thực tế - số ca bệnh mô phỏng) cho các tháng trên được thể hiện trên các Hình 4.14 và 4.15. Năm 2004 do số ca bệnh rất ít nên nghiên cứu không tìm ra được hàm hồi qui có ý nghĩa thống kê để giải thích sự xuất hiện ca bệnh. Đối với các năm 2005, 2006 do không có ca bệnh nào nên không thiết lập hàm hồi quy.

Từ các kết quả phân tích hồi quy OLS thể hiện trên Bảng 4.3 và các Hình 4.14 và 4.15, có thể rút ra một số nhận xét như sau: (1) Trong năm 2007, sự kết hợp của các yếu tố khí hậu và mặt nước giải thích được 26% số ca bệnh trên toàn khu vực; (2) Trong năm 2008, sự kết hợp của các yếu tố dân số và mặt nước giải thích được 42% số ca bệnh trên toàn khu vực; và (3) Trong các năm 2009, 2010, sự kết hợp của các yếu tố khí hậu, dân số và mặt nước giải thích được lần lượt 70%, 64% số ca bệnh trên toàn khu vực.

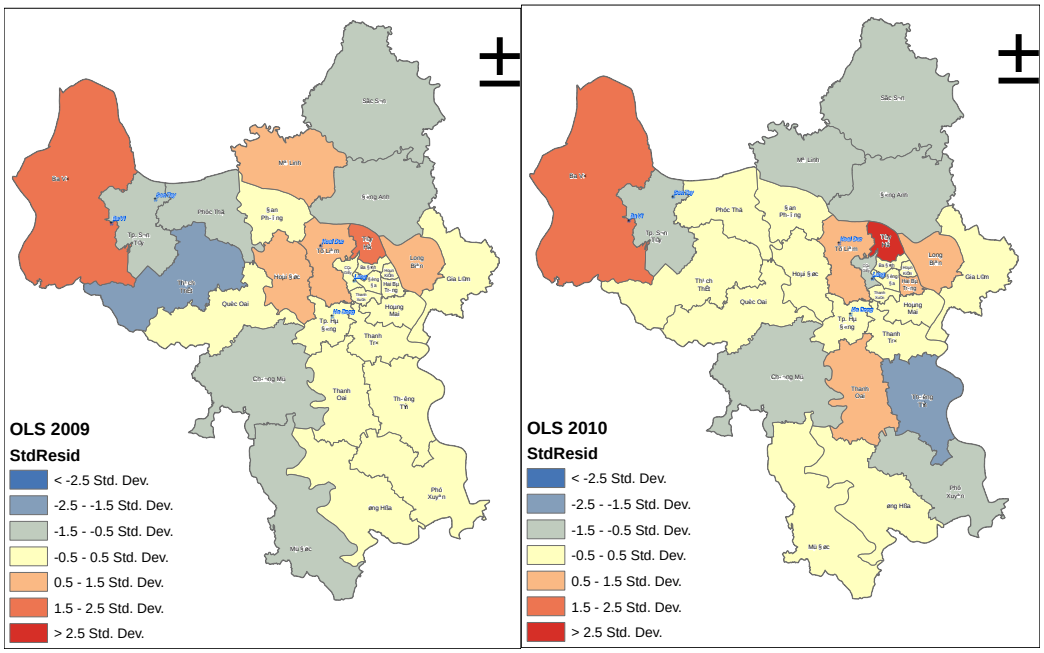
Bảng 4.3. Tổng hợp kết quả phân tích hồi qui OLS theo năm trong khu vực Hà Nội

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*



Hình 4.14. Độ lệch chuẩn của phần dư (số ca thực tế - số ca mô phỏng) năm 2007,2008



Hình 4.15. Độ lệch chuẩn của phần dư (số ca thực tế - số ca mô phỏng) năm 2009, 2010

4.2.2.3. Mô hình GWR

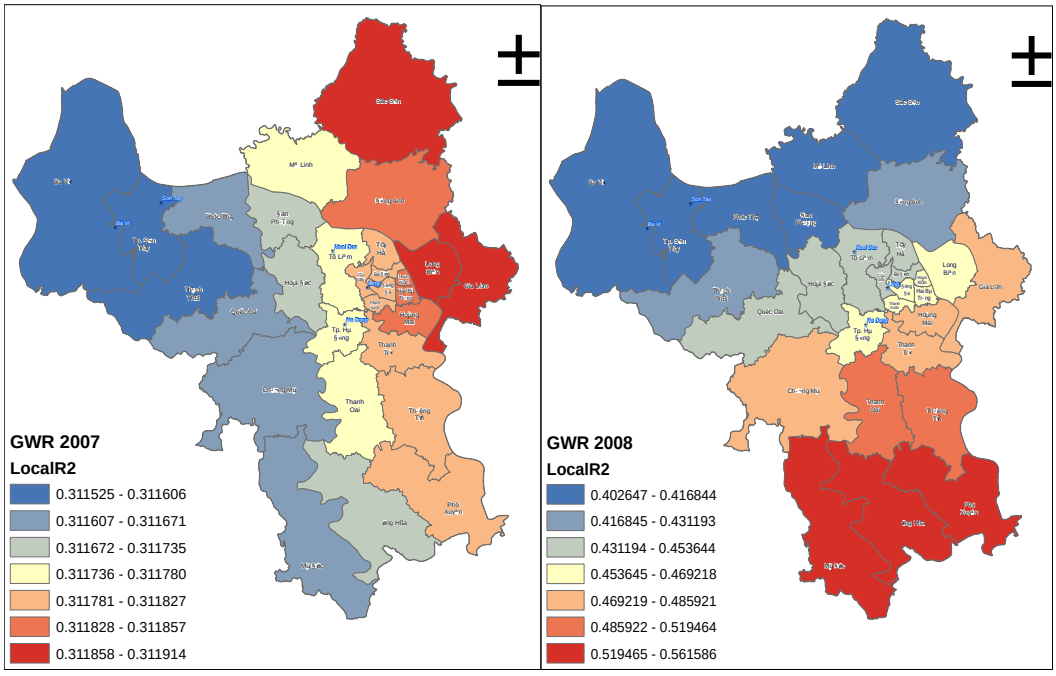
Kỹ thuật thống kê toàn cục OLS giả định tính đồng nhất theo không gian của các mối quan hệ giữa biến phụ thuộc và biến giải thích. Giả thiết này có thể có thể dẫn đến kết quả sai lệch khi OLS được sử dụng cho bộ dữ liệu với tính không đồng nhất của các mối quan hệ theo không gian. Để khắc phục điểm yếu trên, phương pháp thống kê cục bộ hồi qui trọng số không gian (Geographically Weighted Regression-GWR) đã ra đời. Phương pháp này xem xét tính không đồng nhất của các mối quan hệ theo không gian. Nói cách khác, nó mô hình hóa các mối quan hệ thay đổi theo các vị trí không gian khác nhau. 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. Mô hình nào có giá trị AIC thấp sẽ chính xác hơn mô hình có giá trị AIC cao. Kết quả so sánh chỉ số AIC, R^2 hiệu chỉnh giữa OLS và GWR theo từng năm được thể hiện trong Bảng

4.4. Theo đó, ngoại trừ năm 2007, ba năm còn lại mô hình GWR đều cho kết quả tốt hơn mô hình OLS.

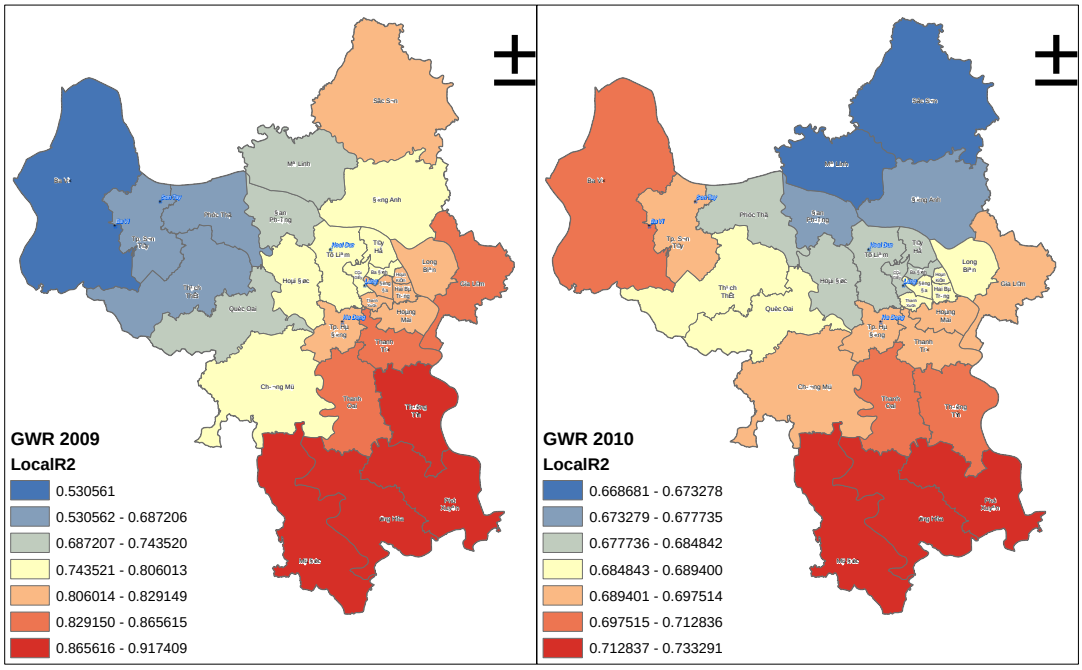
Bảng 4.4. So sánh hiệu quả giữa hai mô hình OLS và GWR theo năm

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

Hình 4.16 và Hình 4.17 thể hiện giá trị các giá trị R^2 cục bộ thay đổi theo từng quận huyện của mô hình GWR. Qua đó cho thấy sự biến động không gian về mối quan hệ giữa các biến giải thích và số ca bệnh tả trong năm. Cụ thể, có thể chia R^2 thành hai nhóm giá trị thấp và cao. Theo đó, đối với năm 2007, có sự gia tăng giá trị R^2 từ Tây sang Đông. Các năm tiếp theo, sự gia tăng giá trị R^2 chuyển thành hướng từ Bắc xuống Nam.Chi tiết kết quả của thực nghiệm mô hình GWR cho các năm từ 2007-2010 được trình bày trong Phụ lục 4.



Hình 4.16. Hệ số R^2 cục bộ của mô hình GWR cho năm 2007, 2008



Hình 4.17. Hệ số R^2 cục bộ của mô hình GWR cho năm 2009, 2010

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

Chương này đã tiến hành nghiên cứu xác định các điểm nóng về dịch tả , xây dựng các mô hình hồi qui OLS, GWR cho dự báo dịch tả trên địa bàn Tp. Hà Nội theo tháng và năm dựa trên các biến khí hậu (nhiệt độ không khí, lượng mưa, độ ẩm, số giờ nắng, tốc độ gió), dân số, diện tích mặt nước trong giai đoạn 2001 - 2012. Các kết quả đạt được 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. Điều đó góp phần hỗ trợ cho công tác quản lý dịch bệnh theo không gian và 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.

Việc xây dựng các mô hình và kết quả dự báo trong chương này đã được công bố trong tạp chí Khoa học Công nghệ thông tin và truyền thông của Học Viện Công nghệ Bưu Chính Viễn thông số 1 năm 2016 và Tạp chí Khoa học công nghệ Đại học Thái Nguyên số 5 năm 2017.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Kết luận

Dự báo dịch bệnh nói chung và dự báo dịch tả nói riêng là một chủ đề nghiên cứu có vai trò quan trọng trong ngành y tế và đặc biệt quan trọng trong công tác y tế dự phòng. Các kết quả dự báo bệnh dịch là một đầu vào quan trọng cho công tác lập kế hoạch và chuẩn bị các nguồn lực cho công tác phòng chống bệnh dịch một cách hiệu quả. 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 (Random Forest, Naïve Bayes, SVM,) và hai bộ hồi qui (tuyến tính và Random Forest). Các kết quả quả thực nghiệm khẳng định phương pháp cửa sổ trượt là phù hợp và kỹ thuật Random Forest cho kết quả dự báo tốt nhất trong số các kỹ thuật được sử dụng để xây dựng mô hình.

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 RandomForest để xây dựng các mô hình dự báo 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 ước lượng bình phương nhỏ nhất 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.

Về ý nghĩa khoa học:

- 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 nghiên cứu thích hợp trong dự báo dịch tả với đặc thù của Hà nội.
- Kết quả thực nghiệm trong luận án giúp nhận định được các chỉ số chính của dịch tả chịu ảnh hưởng của khí hậu, tạo nên sự dễ dàng trong việc ra quyết định, giúp hiểu biết sâu sắc các yếu tố trong mô hình, nhận thức được tầm ảnh hưởng, tác động, vai trò của các yếu tố giúp xác định tập trung vào các yếu tố có độ nhạy cao.
- Kết quả thực nghiệm trong luận án giúp nhận định được các chỉ số chính của dịch tả chịu ảnh hưởng của khí hậu, hỗ trợ quá trình ra quyết định, giúp nhận thức được tầm ảnh hưởng, tác động, vai trò của từng yếu tố. Trên cơ sở đó có thể tập trung xử lý các yếu tố ảnh hưởng có độ nhạy cao.
- Luận án nghiên cứu xây dựng mô hình và lựa chọn kỹ thuật phù hợp trong việc giải quyết từng nội dung bài toán dự báo dịch bệnh góp phần hoàn thiện phương pháp dự báo theo hướng hiện đại và hiệu quả. Trong đó, luận án đã bước đầu tích hợp thời gian, không gian và khí hậu để giải quyết bài toán dự báo dịch bệnh. Như vậy, luận án đã nghiên cứu giải quyết bài toán dự báo dịch bệnh hướng tới tính toàn diện không chỉ về lý thuyết mà còn cả về công nghệ.

Về ý nghĩa thực tiễn:

- Kết quả nghiên cứu của luận án có thể dự báo được số ca mắc tả nhanh chóng giúp cho quá trình ra quyết định, xây dựng chính sách dự phòng y tế, quy hoạch nguồn lực y tế tối ưu để đối phó với khả năng bùng phát dịch bệnh. Trong tương lai có thể nhân rộng mô hình này trong cả nước, hoặc cho các dịch bệnh truyền nhiễm khác có liên quan đến khí hậu.
- Kết quả nghiên cứu trên địa bàn thành phố Hà nội (bao gồm tài liệu, số liệu, bản đồ) là cơ sở dữ liệu hữu ích cho công tác dự phòng y tế cũng như công tác quản lý giám sát dịch bệnh trên địa bàn tỉnh Hà nội.

- Các mô hình dự báo đề xuất 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 hạn chế của luận án

Do hạn chế về số liệu, các mô hình dự báo trong luận án còn có giới hạn theo cả thời gian lẫn không gian, vì vậy có thể phần nào ảnh hưởng tới hiệu quả dự báo của mô hình. Hơn nữa, dữ liệu thu nhận cho nghiên cứu không chi tiết được đến từng địa chỉ các ca bệnh nên việc ứng dụng kỹ thuật phân tích dữ liệu không gian trong GIS còn hạn chế. Các thực nghiệm nghiên cứu dừng lại ở phân tích dịch bệnh theo cấp độ quận/huyện nên độ chính xác về không gian còn hạn chế. Do đó, cần tiến hành thêm phân tích ở cấp độ phường/xã. Ngoài ra, khi thu thập số liệu ca bệnh tả cần ghi nhận chi tiết đến địa chỉ, tọa độ GPS để cung cấp đầu vào chi tiết hơn cho quá trình phân tích trong GIS.

Các mô hình dự báo đề xuất trong luận án chưa giải thích được các yếu tố liên quan đến sự bùng phát dịch tả , như lây truyền bệnh qua đường nước, qua di dân, vệ sinh ăn uống, v.v.. Các mô hình cũng chưa giải thích được số ca mắc tả bị hạn chế do sự can thiệp của ngành y tế (vacxin, truyền thông). Đối với khí hậu, dữ liệu theo dõi khá đầy đủ, trong khi với dữ liệu nước, chỉ có dữ liệu diện tích mặt nước. Vì vậy, cần thu thập thêm số liệu về chất lượng nước mặt trên địa bàn, đặc biệt tại các con sông trong khu vực, để có thể phân tích sâu hơn, toàn diện hơn diễn biến dịch bệnh. Thời gian theo dõi số ca bệnh tả còn ngắn, các ca bệnh chỉ xuất hiện trong 5 năm. Chính vì vậy, để thấy rõ hơn tác động của biến đổi khí hậu đến dịch tả , cần tiếp tục theo dõi tình hình dịch tả , trong một chu kỳ dữ liệu dài hơn.

Hướng nghiên cứu tiếp theo

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

Vấn đề thứ nhất: Nghiên cứu nâng cấp các mô hình đã được xây dựng trong luận án thành hệ hỗ trợ ra quyết định hoàn chỉnh phục vụ cho dự báo dịch bệnh trong ngành y tế. Hệ hỗ trợ ra quyết định bao gồm 5 thành phần: Hệ thống máy tính, cơ sở dữ liệu, quản lý mô hình, quản lý cơ sở tri thức, hệ thống giao tiếp với người dùng.

Trong đó, quản lý mô hình và cơ sở dữ liệu là các thành phần đã được luận án nghiên cứu xây dựng.

Vấn đề thứ hai: Tiếp tục bổ sung dữ liệu với khoảng thời gian lớn hơn để tăng độ chính xác và hoàn thiện mô hình. Nghiên cứu 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.

TÀI LIỆU THAM KHẢO

1. Lê Thị Ngọc Anh, Nguyễn Thị Lan Hương, Nguyễn Hoàng Long và cộng sự (2012). Thiết lập mô hình cảnh báo với độ chễ thời gian cho dịch sốt xuất huyết Dengue tại Hà Nội. *Tạp chí nghiên cứu y học,* 83 (3), 186-192.
2. Lê Thị Ngọc Anh và Nguyễn Minh Sơn (2009). Ứng dụng hệ thống thông tin địa lý (GIS) để xây dung hệ thống bản đồ dịch tễ học về tình trạng lây nhiễm HIV và sử dụng ma túy trong các quận huyện của thành phố Hà Nội. *Tạp chí nghiên cứu y học,* 4 (3), 134-141.
3. Nguyễn Đào (2015). Tăng cường triển khai ứng dụng Hệ thống thông tin địa lý (GIS) chuyên ngành y tế, <https://syt.thuathienhue.gov.vn/?gd=27&cn=90&tc=2870>,
4. Nguyễn Trần Hiển (2012). *Giáo trình dịch tễ học,* NXB Y học,
5. Nguyễn Văn Hiếu (1984). *Đặc điểm dịch tễ học về các vụ dịch tả ở Hải Phòng năm 1976-1981 tại Hải Phòng*, Đại học Y Hà Nội
6. Công ty IBM (2011). Báo cáo phân tích dự báo trong chăm sóc y tế. *Tạp chí Công nghệ thông tin - Truyền thông,* 4 (17), 23-26.
7. Nguyễn Kim Lợi, Lê Cảnh Định và Trần Thống Nhất (2009). *Hệ thống thông tin địa lý nâng cao*, NXB Nông nghiệp,
8. Thành phố Hà nội (2011). *Báo cáo tổng thể hiện trạng môi trường thành phố Hà Nội giai đoạn 5 năm 2006-2010*, Ủy ban nhân dân thành phố Hà Nội,
9. Nguyễn Đình Sơn, Nguyễn Thái Hòa và Dương Quang Minh (2005). Một số đặc điểm dịch tế học bệnh tả tại tỉnh Thừa Thiên Huế,. *Tạp chí y học dự phòng,,* 29740, 194-197.
10. Cục y tế dự phòng và Môi trường - Bộ Y tế (2012). Giám sát và kiểm soát các bệnh truyền nhiễm ở các thành phố lớn tại Việt Nam. *Hội nghị ANMC21*, Hà nội
11. Đỗ Thanh Toàn, Nguyễn Thanh Bình và Lưu Ngọc Hoạt (2012). Tác động của các yếu tố thời tiết lến sự lan truyền của bệnh sốt dengue/sốt xuất huyết dengue tại Hà Nội từ năm 1998-2009. *Tạp chí Nghiên cứu Y Học,* 2.1, 72-78.

12. Getis. A, Ord. J. K (1992). The analysis of Spatial Association by use of Distance Statistics Geographic Analysis,. 24 (3), 189-206.
13. Gray A, Greenhalgh D, Hu L et al (2011). A Stochastic Differential Equation SIS Epidemic Model. *SIAM Journal of Applied Mathematics* 71 (3), 876-902.
14. Huq A, Sack RB (2005). Critical factors in uencing the occurrence of Vibrio cholerae in the environment of Bangladesh. *Applied and Environmental Microbiology*, 71 (8), 4645-4654.
15. Rakesh Agrawal, Tomasz Imielinski, Arun Swami (1993). Mining association rules between sets of items in large databases. *In Proc. of theACM SIGMOD Conference on Management of Data,*, 207-216.
16. Rakesh Agrawal, Ramakrishnan Srikant (1994). Fast Algorithms for Mining Association Rules. *In Proc. of the 20th International Conference on Very Large Databases.*,
17. Agrawal.R, Mannila.H, Srikant.R et al (1996). Fast discovery of association rules. *Advances in knowledge discovery and data mining*, American Association for Artificial Intelligence, Menlo Park, CA, USA, 307-328.
18. Gil AI, Louis VR, Rivera ING (2004). Occurrence and distribution of Vibrio cholerae in the coastal environment of Peru. *Environmental Microbiology.,* 6 (7), 699-706.
19. Dang Duc Anh, Anna Lena Lopez, Vu Dinh Thiem et al (2011). Use of oral cholera vaccines in an outbreak in Vietnam: a case control study. *PLoS Neglected Tropical Diseases*, 5 (1),
20. Rasam ARA, Ghazali R, Noor AMM et al (2014). Spatial epidemiological techniques in cholera mapping and analysis towards a local scale predictive modelling. *IOP Conference Series: Earth and Environmental Science.*, IOP Publishing.
21. Lobitz B, Beck L, Huq A et al (2000). Climate and infectious disease: use of remote sensing for detection of Vibrio cholerae by indirect measurement. *Proceedings of the National Academy of Sciences,* 97 1438–1443.

22. Osei Frank B, Alfred A Duker, Alfred Stein (2012). *Cholera and Spatial Epidemiology*, INTECH Open Access Publisher,
23. Osei Frank B, Alfred A. Duker. (2008). Spatial dependency of V. cholera prevalence on open space refuse dumps in Kumasi, Ghana: a spatial statistical modelling. *International Journal of Health Geographics* 7(1),
24. Fred Brauer, Pauline Van den Driessche, Jianhong Wu (2008). Mathematical Epidemiology. *Springer.*,
25. Breiman.L (2001). Random forests. *Machine Learning*, 45 (1), 5–32,.
26. Butler và Colin D. (2014). *Climate Change and Global Health.*,
27. Martin Charlton, Stewart Fotheringham, Chris Brunadon (2005). *Geographically Weighted Regression*, ESRC National Centre for Research Methods
28. Keya Chaudhuri, Chatterjee S.N (2009). Cholera Toxins.. *Springer*.
29. Colin Childs (2004). *Interpolating Surfaces in ArcGIS: spatial analyst*, ESRI Education.
30. Acosta CJ, Galindo CM, Kimario J và cộng sự (2001). Cholera outbreak in southern Tanzania: Risk factors and patterns of transmission, Emerging Infectious Diseases, <http://www.cdc.gov/ncidod/eid/vol7no3_supp/acosta.htm.>.
31. Rita R Colwell (1996). Global climate and infectious disease: the cholera paradigm. *Science*, 274 (5295), 2025–2031.
32. Adele Cutler. (2015). Random Forests: Statistical Methods for Prediction and Understanding., < http://www.math.usu.edu/~adele/RandomForests/index.htm. http://www.math.usu.edu/~adele/RandomForests/Ovronnaz.pdf>.
33. M.Hurtado- Diaz (2007). Short communication: impact ofclimate variability on the incidence off dengue in Mexico *Tropical medicine & international health*, 12 (11), 1327-1337.
34. Thomas G. Dietteri (2000). An Experimental Comparison of Three Methods for

Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning,* 40 (2), 139–157.

35. David Easley, Jon Kleinberg (2010). Network, Crowds and Market: Reasoning about a hightly connected world. *Cambridge University Press*,
36. Michael Emch, Caryl Feldacker, Mohammad Yunus et al (2008). Local Environmental Predictors of Cholera in Bangladesh and Vietnam.. *Am. J. Trop. Med. Hyg,,* 78(5), 823–832.
37. Yoichi Enatsu, Eleonora Messina, Yoshiaki Muroya et al (2012). Stability analysis of delayed SIR epidemic models with a class of nonlinear incidence rates. *Applied Mathematics and Computation* 218 (9), 5327-5336.
38. Lipp Erin, Huq Anwar, Colwell R (2002). Effects of global climate on infectious disease: the cholera model. *Clinical microbiology reviews*, 15 (4), 757-770.
39. Van den Bergh F, Holloway J P, Pienaar M et al (2008). A comparison of various modelling approaches applied to Cholera case data. *The Journal of ORSSA,* 24 (1), 17-36.
40. Charlie Frye (2011). Choosing an appropriate cell size when interprolating raster data, ArcGIS Resource:,
41. Weiss G.M, Provost (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research* 19, 315– 354.
42. Aarti Garg, Dinesh Gupta (2008). VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatic,* 62 (9).
43. Leckebusch GC, Abdussalam AF (2015). Climate and socioeconomic influences on interannual variability of cholera in Nigeria. *Health & Place*, 34 ([online]), 107–117.
44. Attila Gyenesei (2000). *A Fuzzy Approach for Mining Quantitative Association Rules*, Turku Centre for Computer Science
45. Lan H., Witten, Eibe Frank et al (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd edition). *Morgan Kaufmann*,

46. H.Halide, P.Ridd (2008). A predictive model for Dengue Hemorrhagic Fever epidemics. *International Journal of Environmental Health Research,* 18 (4), 253-265.
47. He Haibo, Edwardo A Garcia (2009). *Learning from Imbalanced Data* IEEE Transactions on knowledge and data engineering, 1263-1284.
48. Christopher Hamlin (2009). Cholera: The biography (Biographies of Diseases). *Oxford University Press*,
49. Tiberiu Harko, Francisco SN, Lobo M K Mak (2014). Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates. *Applied Mathematics and Computation*, 236, 184-194.
50. Hido, Shohei, Hisashi Kashima et al (2008). Roughly balanced bagging for imbalanced data. *Statistical Analysis and Data Mining,* 2 (5.6), 412-426.
51. Quinlan JR (2014). *C4.5: Programs for Machine Learning*, Elsevier,
52. M.N Karim (2012). Climatic factors influencing dengue cases in Ddhaka city, a model fordegue prediction. *The Indian journal of medicine ressearch,* 136 (1), 32-39.
53. Ramandeep Kaur, Er.Prabhsharn Kaur (2016). A review -Heart disease forecasting pattern using various data mining techniques. *International Journal of Computer Science and Moblie Computing,* 5 (6), 350-354.
54. Kelly-Hope (2008). Temporal Trends and Climatic Factors Associated with Bacterial Enteric Diseases in Vietnam, 1991–2001. *Environmental Health Perspectives*, p.7-12.
55. Charles L.Briggs, Clara Mantini Briggs (2002). Stories in the Time of Cholera Racial Profiling during a Medical Nightmare. *University of California Press.*
56. Olshen LBJFR và Stone C (1984). *Classification and Regression Trees.*
57. Hakizimana Leopord, Wilson K Cheruiyot, Stephen Kimani (2016). A Survey and Analysis on Classification and Regression Data Mining Techniques for

Diseases Outbreak Prediction in Datasets *International Journal Of Engineering and Science,* 5 (9), 01-11.

58. Ali M, Lopez AL, You Y et al (2012). *The global burden of cholera,* Bulletin of the World Health Organization., 209-218.
59. Houtsma M, Swami A. (1993). *Set oriented mining for association rules*, IBM Research Report RJ9567,
60. Pascual M, Rodo X, Ellner S P et al (2000). Cholera dynamics and El Ni~no-southern oscillation. *Science,* 289 (5485), 1766-1769.
61. M.Hashizume (2008). Rotavirus infections and climate variability in Dhaka, Bangladesh: a time series analysis. *Epidemiology and Infection,* 136 (9), 1281-1289.
62. M.Hashizume (2008). Association between climate variability and hospital visits for non- chorela diarrhoea in Bangladesh: effects and vulnerable groups. *International Journal of Epidemiol,* 36 (5), 1030-1037.
63. De Magny, Guillaume Constantin, Bernard Cazelles et al (2006). Cholera threat to humans in Ghana is in uenced by both global and regional climatic variability,. *EcoHealth,* 3 (4), 223-231.
64. Heikki Mannila, Hannu Toivonen, A. Inkeri Verkamo. (1994). Efficient Algorithms for Discovering Association Rules. *Workshop on Knowledge Discovery in Databases,* In KDD- AAAI 181-192,.
65. M. Marten, M.Michael (2002). Envirment change, climate and health. *Cambridge University Press,*
66. Cluskey MC, Connell C (2010). Global stability for an SIR epidemic model with delay and nonlinear incidence. Nonlinear Analysis. *Real World Applications* 11 (4), 3106-3109.
67. Xu Min, Cao Chun Xiang, Wang Duo Chun et al (2013). District prediction of cholera risk in China based on environmental factors. *Chinese Science Bulletin, August 2013*, Vol.58 (23), 2798 - 2804.
68. Andreas Mueller (1995). *Fast Sequential and Parallel Algorithms for Association Rule Mining: A Comparison.*, University of Maryland-College Park.

69. Balakrish Nair, Yoshifumi Takeda (2014). Cholera Outbreaks. *Springer*
70. Nkeki Felix Ndidi, Animam Beecroft Osirike (2013). GIS-based local spatial statistical model of cholera occurrence: using geographically weighted regression. *Journal of Geographic Information System,* 5, 531–542.
71. Binh Minh Nguyen, Je Hee Lee, Ngo Tuan Cuong et al (2009). Cholera outbreaks caused by an altered Vibrio cholerae O1 El Tor biotype strain producing classical cholera toxin B in Vietnam in 2007 to 2008. *Clinical Microbiol* 47(5), 1568–1571.
72. J. K Ord, A. Getis (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application.. *Geographic Analysis*, 27 (4), 286-306.
73. World Health Organization (2008). *Severe acute watery diarrhoea with cases positive for Vibrio cholerae, Viet Nam.*, World Health Organization. ,
74. Domingos P (1999). Metacost: A general method for making classifiers cost sensitive. *In proc. of Intl Conf. on Knowledge Discovery and Data Mining*, 55–164.
75. Lenca P, Lallich S, Do T-N et al (2008). A comparison of different off-centered entropies to deal with class imbalance for decision trees. *In The Pacific-Asia Conference on Knowledge Discovery and Data Mining*, LNAI 634–643.
76. Tina R. Patil, Sherekar.S. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications,* 6 (2), 256-261.
77. Jonathen A. Patz, Paut R.Epstein, Thomas A.Burle et al (1996). Global climate change and emerging infectious disease. *JAMA*, 275 (3), 217-223.
78. Yang. Q và Wu (2006). 10 Challenging Problems in Data Mining Research. *Intl Journal of Information Technology and Decision Making,* 5 (4), 597-604

79. Chunara R, Andrews JR, Brownstein JS (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American journal of tropical medicine and hygiene* 86 (1), 39–45.

80. Piarroux R, Barrais R, Faucher B et al (2011). Understanding the cholera epidemic, Haiti. *Emerging Infectious Diseases,* 17 (7), 1161-1168.
81. M Nagabhushana Rao, M Muralidhara Rao, Vedavathi P (2013). Disaster Prediction System Using IBM SPSS Data Mining Tool for Cholera. *International Journal of Computer Science And Technology,* 4 (2), 136-140.
82. Reiner RC, King AA, Emch M et al (2012). Highly localized sensitivity to climate forcing drives endemic cholera in a megacity. *Proceedings of the National Academy of Sciences,* 109 (6), 2033–2036.
83. Lorenzo Righetto (2013). *Hydrological, Anthropogenic and Ecological Processes in Cholera Dynamic.*, École Polytechnique Fédérale De Lausanne.
84. Hyndman RJ, AB Koehler (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting,* 22 (4), 679-688.
85. Hamner S, Tripathi A, Mishra R. K et al (2006). The role of water use patterns and sewage pollution in incidence of water-borne/enteric diseases along the Ganges river in Varanasi, India,. *International Journal of Environmental Health Research* 16 (2), 113-132.
86. José Carlos Santos, Sérgio Matos (2014). Analysing Twitter and web queries for flu trend prediction. *Theoretical Biology and Medical Modelling* 11 (Suppl 1:S6.)
87. Savasere.A., Omiecinski.E, Navathe.S. (1995). *An efficient algorithm for mining association in large databases.*, In VLDB,
88. Bhattacharya SK, Bhattacharya MK, Dutta D et al (1994). Vibrio cholerae O139 in Calcutta,. *Archives of disease in childhood,,* 71 (2), 161-162.
89. Vladimir Svetnik, Andy Liaw, Christopher Tong et al (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences,* 43 (6), 1947-1958.
90. Mythili T, Dev Mukherji, Nikita Padalia et al (2013). A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL). *International Journal of Computer Applications,* 68 (16), 11-15.

91. Gaurav Taneja, Ashwini Sethi (2014). Study of classifiers in data mining *International Journal of Computer Science and Mobile Computing,* 3 (9), 263-269.
92. Troy Tassier (2013). The Economics of Epidemiology. *Springer Berlin Heidelberg.*,
93. Vapnik V (2013). The Nature of Statistical Learning Theory. *Springer Science & Business Media.*,
94. V.Racloz (2012). Surveillance of dengue fever virus: a review of epidemiological models and early warning systems. *PLoS Neglected Tropical Diseases,* 6 (5), 1648.
95. Prieto VM, Matos SA, lvarez M et al (2014). Twitter: A Good Place to Detect Health Conditions. *PLoS ONE 9(1),* e86191,
96. Jin Wang, Shu Liao (2012). A generalized cholera model and epidemic – endemic analysis. *Journal of Biological Dynamics,* 6:2, 568-589.
97. James Wu, Stephen Coggeshall (2012). *Foundations of predictive analytics*, CRC Press.,
98. Rodo X, Pascual M, Fuchs G et al (2002). ENSO and Cholera: A nonstationary link related to climate change? *Proceedings of the National Academy of Sciences*, 99, 12901-12906.
99. Wu X, Kumar V (2009). Top 10 Algorithms in Data Mining. *Chapman & Hall/CRC*,
100. Yusheng Xie, Zhengzhang Chen và Alok N Choudhary (2013). Detecting and Tracking Disease Outbreaks by Mining Social Media Data.. *IJCAI 2013*,
101. Liu XY, Zhou ZH (2006). The influence of class imbalance on costsensitive learning: An empirical study. *In Sixth International Conference on Data Mining (ICDM'06),* 970– 974
102. Yujuan Yue, Jianhua Gong, Duochun Wang et al (2014). Influence of climate factors on Vibrio cholerae dynamics in the Pearl River estuary, South China.. *World J Microbiol Biotechnol, ,* DOI 10.1007/s11274-11014-11604-11275.

103. Zheng. Z, Kohavi. R, Mason (2001). Real world performance of association rule algorithms. *In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 401-406.

104. Mohammed J. Zaki, Ching-Jui Hsiao. (1999). *CHARM: An Efficient Algorithm for Closed Association Rules Mining.*, RPI Technical Report 99- 10, 1999.,

105. Zhi-Hua Zhou. (2012). Ensemble methods: foundations and algorithms. *CRC Press*.

106. Martin Gordon Mubangizi, Ernest MwebazeErnest Mwebaze, John A Quinn (2009). Computational Prediction of Cholera Outbreaks. *5th International Conference on Computing and ICT Research (ICCIR'09)*, Uganda

107. A.R.A. Rasam, Ghazali, R., Noor, A.M.M., Mohd, W.M.N.W., Hamid, J.R.A., Bazlan, M.J. and Ahmad, N. (2014). Spatial epidemiological techniques in cholera mapping and analysis towards a local scale predictive modelling. *IOP Conference Series: Earth and Environmental Science*, http://stacks.iop.org/1755-1315/1718/i=1751/a=012095?key=crossref.012018d453478b012090fd012070bf012231feaecdaab012096.

PHỤ LỤC

Phụ lục1. Kết quả tập luật thu nhận được có độ thống kê lớn hơn 1.

Rule #	LHS	RHS	Support	Confidence	Lift
R1	{Đống Đa, Hai Bà Trưng, Hoàng Mai}	{Thanh Xuân}	0.3027	0.8615	2.0972
R2	{Đống Đa, Hoàng Mai}	{Cầu Giấy}	0.3081	0.7308	2.0484
R3	{Hai Bà Trưng, Hoàng Mai}	{Thanh Xuân}	0.3081	0.8261	2.0109
R4	{Đống Đa, Hai Bà Trưng}	{ThanhXuân}	0.3351	0.7949	1.9349
R5	{Đống Đa, Hoàng Mai}	{Ba Đình}	0.3027	0.7467	1.9185
R6	{Đống Đa, Hoàng Mai}	{Ba Đình}	0.3081	0.7308	1.8777
R7	{Đống Đa, Hai Bà Trưng}	{Thanh Xuân}	0.3081	0.7600	1.8500
R8	{Đống Đa, Hai Bà Trưng, Thanh Xuân}	{Hoàng Mai}	0.3027	0.9825	1.7819
R9	{Từ Liêm}	{Thanh Xuân}	0.3027	0.7273	1.7703
R10	{ThanhXuan}	{Từ Liêm}	0.3027	0.7368	1.7703
R11	{Hai Bà Trưng, Hoàng Mai,Thanh Xuân}	{Đống Đa}	0.3027	0.9825	1.7646
R12	{Đống Đa, Thanh Xuân}	{Hoàng Mai}	0.3351	0.9688	1.7570
R13	{Đống Đa, Hoàng Mai}	{Từ Liêm}	0.3081	0.7308	1.7557
R14	{Hoàng Mai, Thanh Xuân}	{Đống Đa}	0.3351	0.9688	1.7400
R15	{Ba Đình,Hoàng Mai}	{Đống Đa}	0.3081	0.9661	1.7352
R16	{Hoàng Mai, Từ Liêm}	{Đống Đa}	0.3081	0.9661	1.7352
R17	{Cầu Giấy,Đống Đa}	{Hoàng Mai}	0.3081	0.9500	1.7230
R18	{Hai Bà Trưng,Thanh Xuân}	{Hoàng Mai}	0.3081	0.9500	1.7230
R19	{Đống Đa,Hoàng Mai,Thanh Xuân}	{Hai Bà Trưng}	0.3027	0.9032	1.7226
R20	{Cầu Giấy, Hoàng Mai}	{Đống Đa}	0.3081	0.9500	1.7063
R21	{Hai Bà Trưng,Thanh Xuân}	{Đống Đa}	0.3081	0.9500	1.7063
R22	{Ba Đình,Hai Bà Trưng}	{Đống Đa}	0.3027	0.9492	1.7048
R23	{Hoàng Mai, Thanh Xuân}	{Hai Bà Trưng}	0.3081	0.8906	1.6986
R24	{Đống Đa, Thanh Xuân}	{Hai Bà Trưng}	0.3081	0.8906	1.6986
R25	{Ba Đình,Đống Đa}	{Hai Bà Trưng}	0.3027	0.8889	1.6953
R26	{Hai Bà Trưng, Hoàng Mai}	{Đống Đa}	0.3514	0.9420	1.6920
R27	{Cầu Giấy}	{Hoàng Mai}	0.3243	0.9091	1.6488
R28	{Ba Đình,Đống Đa}	{Hoàng Mai}	0.3081	0.9048	1.6410
R29	{Cầu Giấy}	{Đống Đa}	0.3243	0.9091	1.6328

R30	{Cầu Giấy}	{Hai Bà Trưng}	0.3027	0.8485	1.6182
R31	{Hoàn Kiếm}	{Đống Đa}	0.3135	0.8923	1.6027
R32	{Đống Đa, Từ Liêm}	{Hoàng Mai}	0.3081	0.8769	1.5905
R33	{Đống Đa, Hoàng Mai}	{Hai Bà Trưng}	0.3514	0.8333	1.5893
R34	{Đống Đa, Hai Bà Trưng}	{Hoàng Mai}	0.3514	0.8667	1.5719
R35	{Ba Đình}	{Đống Đa }	0.3405	0.8750	1.5716
R36	{Ba Đình}	{Hai Bà Trưng}	0.3189	0.8194	1.5629
R37	{Thanh Xuân}	{Hoàng Mai}	0.3459	0.8421	1.5273
R38	{ Từ Liêm}	{ Đống Đa}	0.3514	0.8442	1.5162
R39	{Thanh Xuân}	{Đống Đa}	0.3459	0.8421	1.5125
R40	{Thanh Xuân}	{Hai Bà Trưng}	0.3243	0.7895	1.5057
R41	{Ba Đình}	{Hoàng Mai}	0.3189	0.8194	1.4862
R42	{Từ Liêm}	{Hai Bà Trưng}	0.3081	0.7403	1.4118
R43	{Từ Liêm}	{Hoàng Mai}	0.3189	0.7662	1.3897
R44	{Hà Đông}	{Đống Đa}	0.3135	0.7733	1.3890
R45	{Hai Bà Trưng}	{Đống Đa}	0.4054	0.7732	1.3888
R46	{Đống Đa}	{Hai Bà Trưng}	0.4054	0.7282	1.3888
R47	{Hoàng Mai}	{Đống Đa}	0.4216	0.7647	1.3735
R48	{Đống Đa}	{Hoàng Mai}	0.4216	0.7573	1.3735
R49	{Hà Đông}	{Hoàng Mai}	0.3027	0.7467	1.3542
R50	{Hai Bà Trưng}	{Hoàng Mai}	0.3730	0.7113	1.2902

Phụ lục 2. Kết quả thực nghiệm mô hình dự báo cục bộ với hai thuật toán hồi quy và ba bộ phân lớp cho 29 quận/huyện tại Hà Nội

Quận/Huyện	Các độ đo	Linear Regression	NaiveBayes	LibSVM	Random Forest
Ba Đình	Correlation coefficient	-0.0233			
	Mean absolute error	7.3397	0.2827	0.2222	0.2222
	Root mean squared error	8.8134	0.497	0.4714	0.4714
	Precision		0.4	0.444	0.667
	Recall		0.5	0.667	0.667
	F-Measure		0.444	0.533	0.667
Ba Vì	Correlation coefficient	0			
	Mean absolute error	0.9287	0.0073	0	0.3333
	Root mean squared error	1.1451	0.0176	0	0.5774
	Precision		1	1	1
	Recall		1	1	0.667
	F-Measure		1	1	0.8
Cầu Giấy	Correlation coefficient	-0.0297			
	Mean absolute error	7.71	0.2002	0.1111	0.1111
	Root mean squared error	8.6413	0.4201	0.3333	0.3333
	Precision		0.667	0.694	0.833
	Recall		0.667	0.833	0.833
	F-Measure		0.667	0.758	0.833
Chương Mỹ	Correlation coefficient	-0.1623			
	Mean absolute error	24.1349	0.1126	0.1111	0.1111
	Root mean squared error	30.0617	0.3334	0.3333	0.3333
	Precision		0.694	0.694	0.694
	Recall		0.833	0.833	0.833
	F-Measure		0.758	0.758	0.758
Đan Phượng	Correlation coefficient	0			
	Mean absolute error	2.0773	0.2205	0	0.1111
	Root mean squared error	3.5375	0.4658	0	0.3333
	Precision		1	1	1
	Recall		0.667	1	0.833
	F-Measure		0.8	1	0.909
Đông Anh	Correlation coefficient	-0.0063			

	Mean absolute error	0.7127	0.2313	0.1667	0.3333
	Root mean squared error	0.7554	0.4285	0.4082	0.5774
	Precision		0.917	0.694	0.667
	Recall		0.833	0.833	0.667
	F-Measure		0.852	0.758	0.667
Đông Đa	Correlation coefficient	-0.0713			
	Mean absolute error	22.8332	0.2504	0.2222	0.3333
	Root mean squared error	26.5469	0.4741	0.4714	0.5774
	Precision		0.583	0.444	0.722
	Recall		0.667	0.667	0.5
	F-Measure		0.611	0.533	0.528
Gia Lâm	Correlation coefficient	0.4345			
	Mean absolute error	0.9225	0.2795	0.1667	0.3333
	Root mean squared error	1.1804	0.4861	0.4082	0.5774
	Precision		0.667	0.694	0.667
	Recall		0.667	0.833	0.667
	F-Measure		0.667	0.758	0.667
Hà Đông	Correlation coefficient	0.1117			
	Mean absolute error	11.3664	0.242	0.1111	0.1111
	Root mean squared error	14.4096	0.3898	0.3333	0.3333
	Precision		0.667	0.694	0.694
	Recall		0.667	0.833	0.833
	F-Measure		0.667	0.758	0.758
Hai Bà Trưng	Correlation coefficient	0.2739			
	Mean absolute error	13.9127	0.364	0.3333	0.3333
	Root mean squared error	15.2025	0.5782	0.5774	0.5774
	Precision		0.333	0.25	0.417
	Recall		0.5	0.5	0.5
	F-Measure		0.397	0.333	0.452
Hoài Đức	Correlation coefficient	0.0648			
	Mean absolute error	12.7375	0.2368	0.1111	0.2222
	Root mean squared error	19.4759	0.4734	0.3333	0.4714
	Precision		0.889	0.694	0.667
	Recall		0.667	0.833	0.667
	F-Measure		0.708	0.758	0.667

Hoàng Mai	Correlation coefficient	0.5317			
	Mean absolute error	12.7367	0.2227	0.2222	0.2222
	Root mean squared error	13.8483	0.453	0.4714	0.4714
	Precision		0.444	0.444	0.583
	Recall		0.667	0.667	0.667
	F-Measure		0.533	0.533	0.611
Hoàn Kiếm	Correlation coefficient	0.0642			
	Mean absolute error	6.9152	0.0122	0.1111	0.2222
	Root mean squared error	7.9645	0.0254	0.3333	0.4714
	Precision		1	0.694	0.667
	Recall		1	0.833	0.667
	F-Measure		1	0.758	0.667
Long Biên	Correlation coefficient	0			
	Mean absolute error	7.2068	0.2482	0	0.1111
	Root mean squared error	7.9562	0.4726	0	0.3333
	Precision		1	1	1
	Recall		0.667	1	0.833
	F-Measure		0.8	1	0.909
Mê Linh	Correlation coefficient	0			
	Mean absolute error	0.7311	0	0	0
	Root mean squared error	0.9015	0	0	0
	Precision		1	1	1
	Recall		1	1	1
	F-Measure		1	1	1
Mỹ Đức	Correlation coefficient	0			
	Mean absolute error	0.1772	0	0	0.1667
	Root mean squared error	0.2925	0	0	0.4082
	Precision		1	1	1
	Recall		1	1	0.833
	F-Measure		1	1	0.909
Phúc Thọ	Correlation coefficient	0.8624			
	Mean absolute error	1.3053	0.1667	0.1667	0.1667
	Root mean squared error	1.4983	0.4082	0.4082	0.4082
	Precision		0.694	0.694	0.694
	Recall		0.833	0.833	0.833
	F-Measure		0.758	0.758	0.758

Phú Xuyên	Correlation coefficient	0			
	Mean absolute error	0.878	0	0	0
	Root mean squared error	1.1695	0	0	0
	Precision		1	1	1
	Recall		1	1	1
	F-Measure		1	1	1
Quốc Oai	Correlation coefficient	0.1349			
	Mean absolute error	1.4722	0.5114	0.1667	0.5
	Root mean squared error	2.0455	0.636	0.4082	0.7071
	Precision		0.875	0.694	0.625
	Recall		0.5	0.833	0.5
	F-Measure		0.543	0.758	0.556
Sóc Sơn	Correlation coefficient	0			
	Mean absolute error	1.1202	0	0	0.1667
	Root mean squared error	1.2118	0	0	0.4082
	Precision		1	1	1
	Recall		1	1	0.833
	F-Measure		1	1	0.909
Sơn Tây	Correlation coefficient	0			
	Mean absolute error	0.3072	0	0	0
	Root mean squared error	0.3737	0	0	0
	Precision		1	1	1
	Recall		1	1	1
	F-Measure		1	1	1
Tây Hồ	Correlation coefficient	-0.617			
	Mean absolute error	9.5829	0.2616	0.1111	0.2222
	Root mean squared error	10.23	0.4854	0.3333	0.4714
	Precision		0.889	0.694	0.833
	Recall		0.667	0.833	0.667
	F-Measure		0.708	0.758	0.741
Thạch Thất	Correlation coefficient	0.4328			
	Mean absolute error	51.4789	0.2226	0.1111	0.2222
	Root mean squared error	64.4701	0.4713	0.3333	0.4714
	Precision		0.667	0.694	0.667
	Recall		0.667	0.833	0.667
	F-Measure		0.667	0.758	0.667

Thanh Oai	Correlation coefficient	0.2916			
	Mean absolute error	6.5756	0.1734	0.1111	0.1111
	Root mean squared error	9.048	0.3821	0.3333	0.3333
	Precision		0.667	0.694	0.694
	Recall		0.667	0.833	0.833
	F-Measure		0.667	0.758	0.758
Thanh Trì	Correlation coefficient	0.0078			
	Mean absolute error	4.9363	0.1163	0.1111	0.2222
	Root mean squared error	6.1091	0.3337	0.3333	0.4714
	Precision		0.694	0.694	0.667
	Recall		0.833	0.833	0.667
	F-Measure		0.758	0.758	0.667
Thanh Xuân	Correlation coefficient	0.1829			
	Mean absolute error	13.1576	0.2373	0.1111	0.3333
	Root mean squared error	15.9419	0.467	0.3333	0.5774
	Precision		0.833	0.694	0.625
	Recall		0.667	0.833	0.5
	F-Measure		0.741	0.758	0.556
Thường Tín	Correlation coefficient	0			
	Mean absolute error	21.8024	0	0	0
	Root mean squared error	34.9868	0	0	0
	Precision		1	1	1
	Recall		1	1	1
	F-Measure		1	1	1
Từ Liêm	Correlation coefficient	0.1369			
	Mean absolute error	12.8475	0.5345	0.2222	0.2222
	Root mean squared error	17.0669	0.7164	0.4714	0.4714
	Precision		0.333	0.444	0.722
	Recall		0.167	0.667	0.667
	F-Measure		0.222	0.533	0.655
Ứng Hòa	Correlation coefficient	0			
	Mean absolute error	1.9787	0	0	0
	Root mean squared error	2.3641	0	0	0
	Precision		1	1	1
	Recall		1	1	1
	F-Measure		1	1	1

Phụ lục 3: Kết quả hồi qui và độ quan trọng của các biến khí hậu

Kết quả mô hình hồi qui tuyến tính so sánh sự biến đổi của độ đo R^2 theo độ dài của khoảng thời gian dự báo.

Quận	Độ dài dự báo			
	30 ngày	14 ngày	7 ngày	3 ngày
Badinh	0.0092	0.0448	0.3126	0.3592
Bavi	-0.0001	0.0015	-0.0002	0.0186
Caugiay	0.0107	0.0623	0.2323	0.4142
Chuongmy	0.0015	0.1117	0.0745	0.2343
Danphuong	0.0003	0.0088	0.0177	0.2331
Donganh	-0.0002	0.0434	0.0085	0.0476
Dongda	0.0491	0.1325	0.4133	0.6214
Gialam	-0.0002	0.0007	0.0509	0.0831
Hadong	0.0106	0.0314	0.2248	0.3998
Haibatrung	0.0085	0.0719	0.2090	0.4544
Hoaiduc	0.0037	0.0386	0.2188	0.5297
Hoankiem	0.0240	0.0739	0.2466	0.4122
Hoangmai	0.0254	0.0458	0.1821	0.3372
Longbien	0.0070	0.0121	0.1152	0.2316
Melinh	-0.0001	-0.0002	0.0666	0.0017
Myduc	-0.0002	0.0171	0.0169	0.3762
phuxuyen	-0.0001	0.0003	0.0335	0.0281
phuctho	0.0005	0.0214	0.0248	0.0130
quocoai	0.0012	0.0161	0.0553	0.0785
socson	0.0020	0.0070	-0.0002	0.0642
sontay	-0.0002	-0.0002	-0.0002	-0.0002
tayho	0.0086	0.0735	0.2706	0.4905
thachthat	0.0005	-0.0002	0.0051	0.1500
thanhoai	0.0094	0.0439	0.1588	0.2535
thanhtri	0.0152	0.0159	0.0386	0.0967
thanhxuan	0.0277	0.0551	0.2870	0.4969
thuongtin	0.0196	0.1813	0.1243	0.3211
tuliem	0.0137	0.0699	0.2345	0.4197
Unghoa	-0.0002	0.0137	0.0008	0.0100

5.2 Độ quan trọng của các biến khí hậu trong mô hình kết hợp đầy đủ các yếu tố khí hậu và địa lý ở mội quận với các độ dài khoảng dự báo

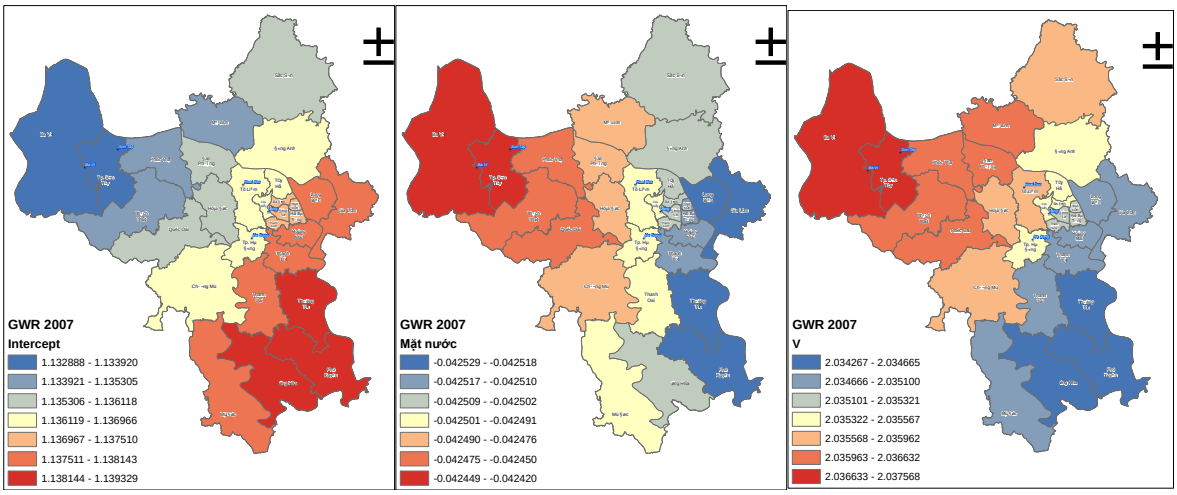
Quận	Độ dài dự bảo (ngày)	Độ ẩm	Lượng mưa	SOI	Số giờ nắng	Nhiệt độ	Tốc độ gió
badinh	3	33.1	11.5	55.2	27.4	60.5	0.0
bavi	3	82.6	47.7	0.0	100.0	64.9	39.6
caugiay	3	8.4	0.0	25.7	7.7	5.9	1.4
chuongmy	3	63.7	13.6	25.9	20.9	15.1	0.0
danphuong	3	2.5	4.7	26.0	0.0	21.1	25.9
donganh	3	53.9	35.9	0.0	28.4	43.6	27.9
dongda	3	58.3	2.0	23.3	12.8	38.8	0.0
gialam	3	37.6	46.5	43.3	9.7	32.0	16.1
hadong	3	19.8	22.6	24.5	16.0	42.1	0.0
haibatrung	3	46.1	15.6	39.9	20.5	21.2	0.0
hoaiduc	3	11.8	10.6	15.8	3.5	28.0	0.0
hoangmai	3	21.3	0.0	13.9	10.6	30.7	30.9
hoankiem	3	68.4	43.8	0.0	3.7	47.1	39.7
longbien	3	34.9	15.2	91.2	46.7	48.5	4.3
melinh	3	100.0	17.8	31.5	96.7	96.2	29.1
myduc	3	49.5	57.7	16.1	31.6	32.7	0.0
phuctho	3	100.0	88.3	0.0	89.2	81.3	5.3
phuxuyen	3	49.3	10.9	10.2	56.1	44.6	0.0
quocoai	3	39.8	0.0	46.5	40.4	59.2	54.6
socson	3	0.0	57.6	39.6	7.1	36.9	15.3
sontay	3	70.6	30.2	0.0	100.0	40.8	6.2
tayho	3	49.8	20.4	30.6	24.4	66.9	14.7
thachthat	3	73.9	83.1	8.4	11.2	42.4	68.9
thanhoai	3	46.7	34.4	32.2	39.5	46.0	0.0
thanhtri	3	50.2	1.8	77.9	15.7	29.2	46.3
thanhxuan	3	53.2	0.0	67.8	2.6	27.9	7.9
Thuongtin	3	20.5	20.5	0.0	18.0	26.6	18.8
Tuliem	3	5.3	0.0	6.6	9.7	9.6	10.3
Unghoa	3	69.6	41.9	14.1	50.0	68.5	68.8
badinh	7	21.7	25.1	15.4	45.4	50.7	0.0
bavi	7	45.9	98.7	0.0	72.4	100.0	41.2
caugiay	7	15.1	16.5	0.0	29.6	55.1	0.0
chuongmy	7	26.6	35.5	56.0	13.7	51.7	0.0

Quận	Độ dài dự bảo (ngày)	Độ ẩm	Lượng mưa	SOI	Số giờ nắng	Nhiệt độ	Tốc độ gió
danphuong	7	21.6	15.5	11.1	12.5	29.7	20.1
donganh	7	23.3	10.9	9.4	36.8	32.6	23.3
dongda	7	29.5	14.9	12.8	51.6	47.4	0.0
gialam	7	43.0	35.1	23.6	30.2	58.3	0.0
hadong	7	22.5	0.0	27.3	17.4	55.0	20.2
haibatrung	7	27.0	1.7	0.0	29.0	45.0	18.9
hoaiduc	7	38.3	0.0	22.4	37.7	45.2	4.2
hoangmai	7	29.0	14.8	0.0	38.2	38.0	10.5
hoankiem	7	38.4	28.6	51.7	52.2	85.6	0.0
longbien	7	73.3	53.0	24.9	57.2	73.6	0.0
melinh	7	28.7	23.8	7.0	67.1	97.8	0.0
myduc	7	66.4	24.0	11.4	51.1	52.2	0.0
phuctho	7	68.8	72.0	51.2	57.4	100.0	35.9
phuxuyen	7	44.8	12.6	5.6	98.5	68.1	0.0
quocoai	7	22.6	36.1	0.0	6.4	27.7	75.1
socson	7	54.7	48.8	24.0	100.0	79.2	21.4
sontay	7	27.4	6.5	0.0	15.9	100.0	33.8
tayho	7	47.2	3.8	23.8	34.1	57.3	0.0
thachthat	7	82.8	100.0	55.0	0.0	44.6	23.5
thanhoai	7	63.5	5.8	0.0	35.8	100.0	14.0
thanhtri	7	24.8	14.6	0.0	31.0	56.6	39.0
thanhxuan	7	12.5	17.1	8.2	10.2	41.0	0.0
thuongtin	7	55.5	57.2	21.6	33.7	28.7	43.7
tuliem	7	32.3	21.3	8.5	27.8	46.4	38.7
unghoa	7	36.8	36.6	20.8	35.9	28.6	0.0
badinh	14	41.4	52.3	5.2	61.0	50.4	0.0
bavi	14	66.4	100.0	39.1	0.0	46.8	37.5
caugiay	14	76.1	63.7	16.9	72.6	100.0	0.0
chuongmy	14	0.0	13.9	34.7	45.2	70.5	34.6
danphuong	14	24.6	0.0	48.1	6.0	68.1	23.8
donganh	14	88.9	85.2	22.1	8.2	100.0	38.9
dongda	14	66.7	38.4	0.0	79.5	100.0	27.1
gialam	14	94.2	57.0	37.6	100.0	92.2	36.6
hadong	14	64.4	32.3	12.5	43.4	57.3	0.0

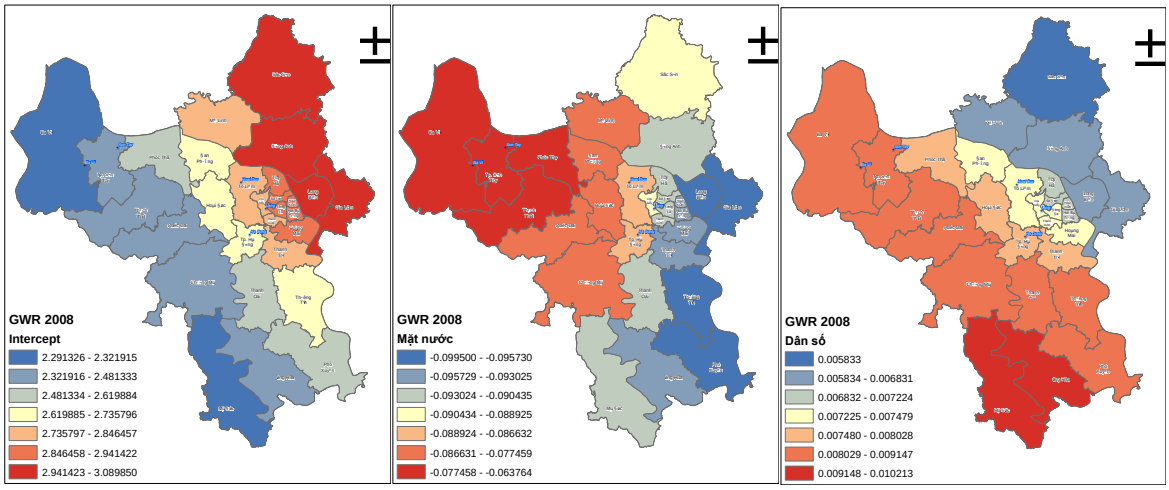
Quận	Độ dài dự bảo (ngày)	Độ ẩm	Lượng mưa	SOI	Số giờ nắng	Nhiệt độ	Tốc độ gió
haibatrung	14	48.2	34.6	0.0	83.5	82.8	14.7
hoaiduc	14	0.0	15.1	8.2	7.2	50.0	14.2
hoangmai	14	64.3	50.9	0.0	56.8	74.8	19.5
hoankiem	14	41.8	32.4	0.0	76.4	80.8	10.0
longbien	14	70.0	36.2	0.0	87.1	82.6	7.0
melinh	14	83.0	42.2	23.2	100.0	85.1	0.0
myduc	14	10.2	0.0	14.8	42.0	100.0	55.7
phuctho	14	48.8	59.0	26.2	93.0	59.9	24.2
phuxuyen	14	100.0	84.3	20.7	57.1	84.4	0.0
quocoai	14	83.9	64.2	20.7	55.6	65.9	29.5
socson	14	86.3	20.9	3.2	96.7	100.0	39.8
sontay	14	33.5	16.8	64.6	84.6	100.0	86.1
tayho	14	49.4	58.0	16.6	56.8	81.4	44.8
thachthat	14	14.4	35.6	16.8	0.0	80.8	95.3
thanhoai	14	62.1	35.9	0.0	47.2	100.0	21.3
thanhtri	14	54.1	34.4	0.0	37.2	40.4	3.1
thanhxuan	14	56.9	39.2	5.7	71.9	66.6	0.0
thuongtin	14	43.6	44.9	17.6	1.4	21.1	4.7
tuliem	14	38.0	32.5	7.8	22.9	53.9	31.9
unghoa	14	18.7	0.0	32.7	33.3	70.5	54.6
badinh	30	57.5	36.6	0.0	61.6	29.9	11.7
bavi	30	59.8	27.4	0.0	23.0	100.0	7.2
caugiay	30	49.6	48.3	0.0	54.6	32.7	14.6
chuongmy	30	32.9	76.0	1.5	0.0	62.6	59.2
danphuong	30	100.0	13.8	16.2	73.9	74.2	33.4
donganh	30	91.6	84.2	38.2	100.0	94.8	20.2
dongda	30	43.0	29.1	0.0	48.7	32.2	4.1
gialam	30	79.4	100.0	0.0	37.5	58.0	57.0
hadong	30	22.0	26.8	0.0	30.0	25.5	10.5
haibatrung	30	58.2	44.9	10.7	62.8	45.3	0.0
hoaiduc	30	39.0	18.6	0.0	17.2	21.7	15.5
hoangmai	30	29.6	41.3	0.0	22.2	34.4	12.5
hoankiem	30	30.7	25.4	0.0	40.5	34.8	19.3
longbien	30	30.6	28.1	0.0	31.3	13.9	14.3

Quận	Độ dài dự bảo (ngày)	Độ ẩm	Lượng mưa	SOI	Số giờ nắng	Nhiệt độ	Tốc độ gió
melinh	30	71.0	57.1	27.6	100.0	67.5	22.9
myduc	30	87.8	0.0	25.8	78.7	100.0	10.2
phuctho	30	7.6	29.1	14.1	0.0	7.4	58.5
phuxuyen	30	63.1	70.1	32.4	100.0	73.8	26.5
quocoai	30	76.3	50.0	38.2	85.9	47.0	71.1
socson	30	57.1	100.0	11.5	21.3	42.5	68.5
sontay	30	64.7	30.8	16.9	87.0	0.0	77.5
tayho	30	76.7	21.6	0.0	87.9	51.2	37.3
thachthat	30	14.8	100.0	67.0	0.0	9.3	49.6
thanhoai	30	46.8	44.3	0.0	79.5	51.7	31.5
thanhtri	30	57.0	19.2	3.6	50.1	0.0	4.5
thanhxuan	30	63.7	52.6	10.4	53.4	54.1	2.6
thuongtin	30	99.5	100.0	51.0	29.9	0.0	84.0
tuliem	30	62.6	63.8	8.6	13.2	49.8	0.0
unghoa	30	46.8	100.0	94.5	38.9	0.0	59.6

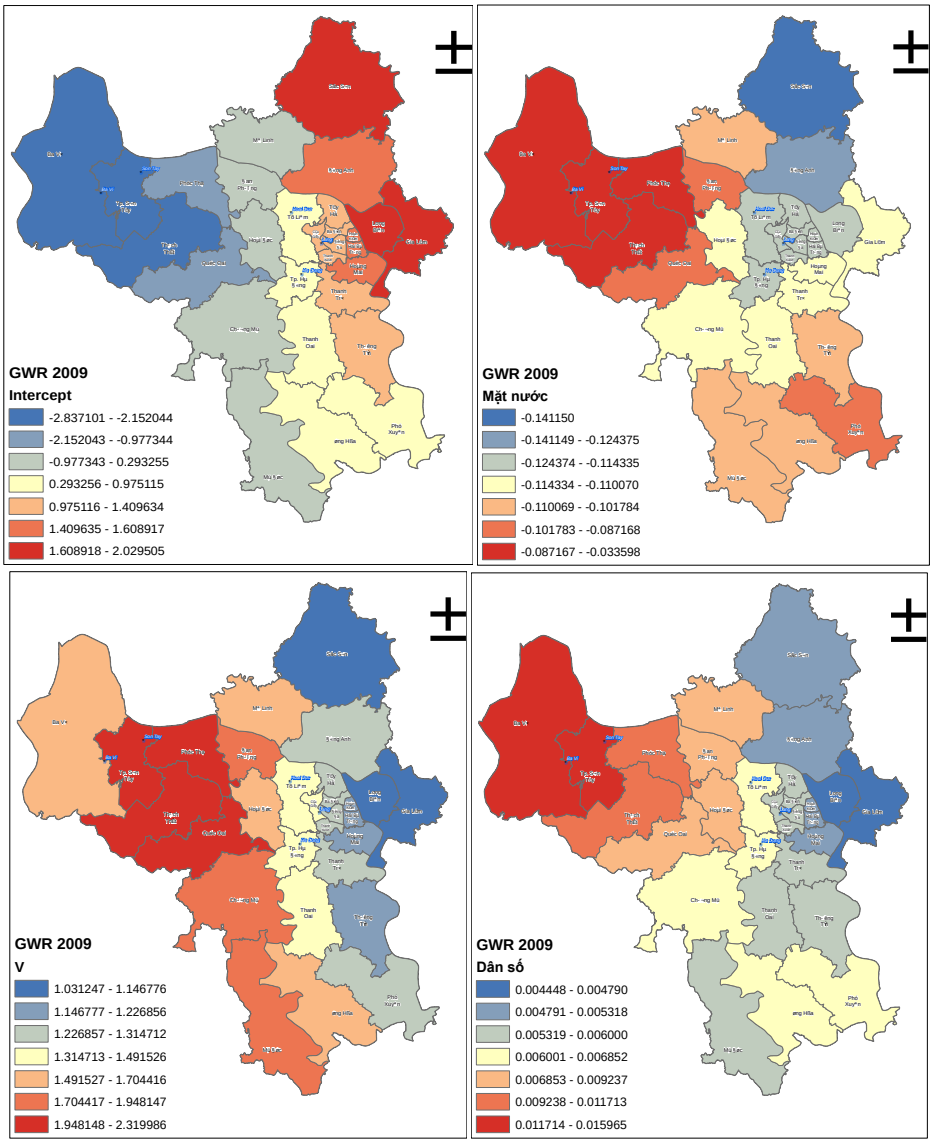
Phụ lục 4. Kết quả thực nghiệm mô hình GWR cho các năm từ 2007-2010



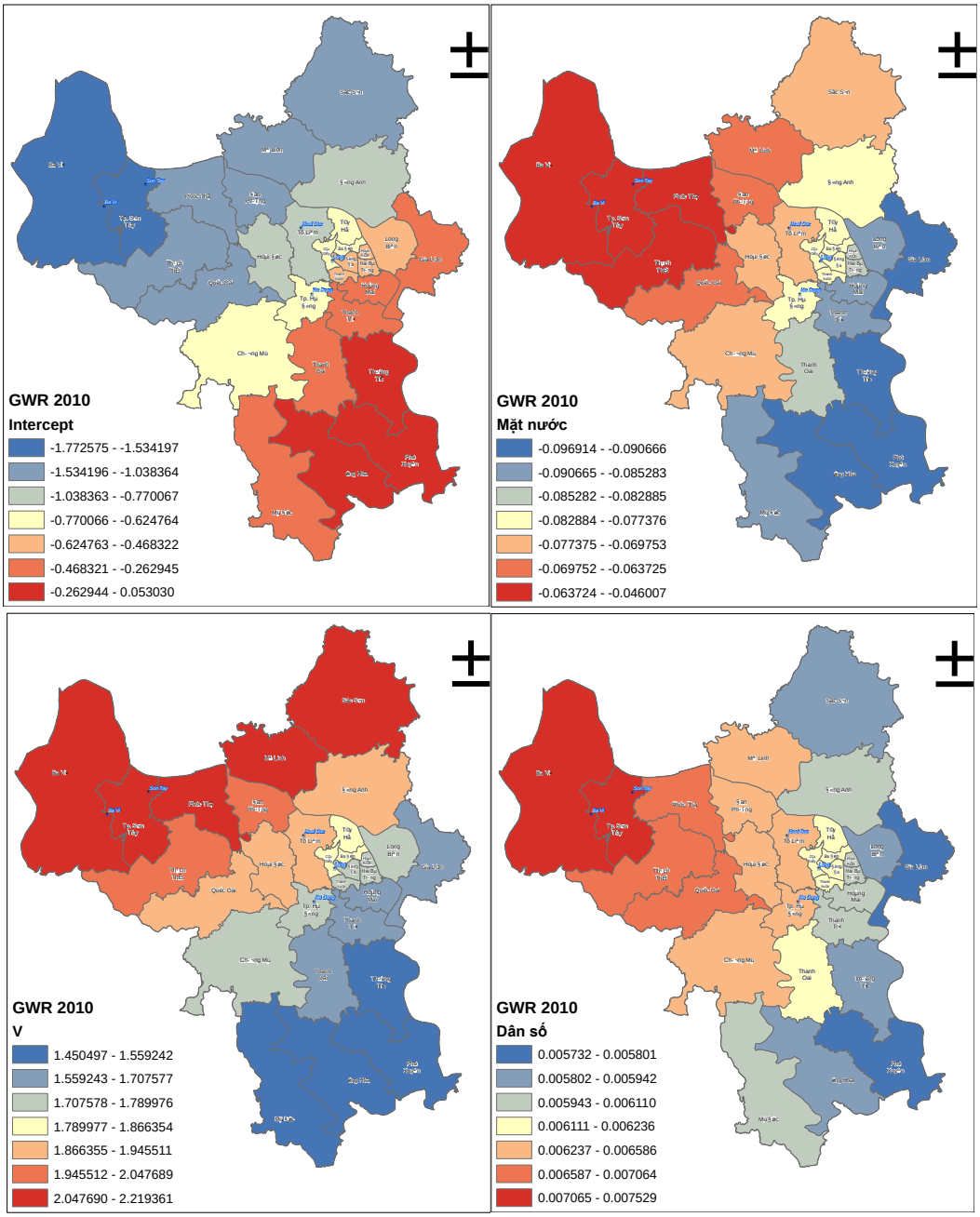
Hình 1: Tham số cục bộ của mô hình GWR cho năm 2007



Hình 2 : Tham số cục bộ của mô hình GWR cho năm 2008



Hình 3: Tham số cục bộ của mô hình GWR cho năm 2009



Hình 4 : Tham số cục bộ của mô hình GWR cho năm 2010