ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN ĐỨC LINH

NGHIÊN CỨU VÀ XÂY DỰNG QUI TRÌNH CHUẨN HÓA DỮ LIỆU QUAN TRẮC MÔI TRƯỜNG Ở VIỆT NAM

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

HÀ NỘI - 2016

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN ĐỨC LINH

NGHIÊN CỨU VÀ XÂY DỰNG QUI TRÌNH CHUẨN HÓA DỮ LIỆU QUAN TRẮC MÔI TRƯỜNG Ở VIỆT NAM

Ngành: Công nghệ thông tin

Chuyên ngành: Kỹ thuật phần mềm

Mã số: 60480103

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. NGUYỄN THỊ NHẬT THANH.

TS. BÙI QUANG HƯNG

HÀ NỘI - 2016

LỜI CAM ĐOAN

Tôi xin cam đoan kết quả đạt được trong luận văn là sản phẩm nghiên cứu, tìm hiểu của riêng cá nhân tôi. Trong toàn bộ nội dung của luận văn, những điều được trình bày hoặc là của cá nhân tôi hoặc là được tổng hợp từ nhiều nguồn tài liệu. Tất cả các tài liệu tham khảo đều có xuất xứ rõ ràng và được trích dẫn hợp pháp.

Tôi xin hoàn toàn chịu trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của mình.

Hà Nội, Tháng 03 – Năm 2016

Người cam đoan

Nguyễn Đức Linh.

LỜI CẢM ƠN

Đề tài luận văn cao học của tôi được hoàn thành tại Đại học Công Nghệ - Đại học Quốc gia Hà Nội. Để có thể hoàn thành được đề tài luận văn này, tôi xin bày tỏ lòng biết ơn chân thành và sâu sắc tới Trung tâm Công nghệ tích hợp liên ngành giám sát hiện trường (FIMO), Đại học Công Nghệ, ĐHQG Hà Nội, đặc biệt là TS. Nguyễn Thị Nhật Thanh và TS. Bùi Quang Hưng đã trực tiếp hướng dẫn, dìu dắt, tận tình giúp đỡ tôi về cả chuyên môn, nghiên cứu và định hướng phát triển trong suốt quá trình triển khai, nghiên cứu, hoàn thành đề tài *"Nghiên cứu và xây dựng qui trình chuẩn hóa dữ liệu quan trắc môi trường ở Việt Nam"*.

Xin chân thành cảm ơn các Thầy, Cô giáo của khoa Công nghệ thông tin đã truyền dạy cho tôi những kiến thức khoa học bổ ích, truyền cho tôi ngọn lửa yêu nghề, lòng nhiệt huyết và tình yêu công việc.

Xin chân thành cảm ơn toàn thể các thành viên đã và đang hoạt động tại trung tâm FIMO đã động viên, khích lệ, tạo điều kiện và giúp đỡ tôi trong suốt quá trình thực hiện và hoàn thành luận văn này

Cuối cùng, với gia đình, tôi xin gửi lời biết ơn sâu sắc vì gia đình đã luôn ở bên cạnh tôi, mang lại cho tôi nguồn động viên tinh thần to lớn và tạo mọi điều kiện thuận lợi cho tôi học tập, nghiên cứu để hoàn thành luận văn này.

Mặc dù đã có nhiều cố gắng để thực hiện đề tài một cách hoàn chỉnh nhất. Song với kinh nghiệm còn non trẻ trong công việc nghiên cứu khoa học cũng như hạn chế về kiến thức nên không thể tránh khỏi những thiếu sót nhất định mà chính bản thân cũng chưa nhận thấy được. Qua bản luận văn này tôi rất mong nhận được sự góp ý của quý Thầy, Cô giáo và các bạn đồng nghiệp để luận văn được hoàn chỉnh hơn.

Tôi xin chân thành cảm ơn!

Hà Nội, Tháng 03 – Năm 2016

Nguyễn Đức Linh

MỤC LỤC

BẢNG CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

Ký hiệu	Tiếng Anh	Tiếng Việt
SARS	Severe Acute Respiratory syndrome	Hội chứng hô hấp cấp tính nặng
MERS	Middle East Respiratory Ryndrome	Bệnh viêm đường hô hấp cấp do vi-rút
PM	Particulate matter	Bụi
WHO	World Health Organization	Tổ chức Y tế Thế giới
CEM	Centre for Environmental Monitoring	Trung tâm Quan trắc môi trường
TSP	Total Suspended Pảticles	Tổng bụi lơ lửng
AQI	Air Quality Index	Chỉ số chất lượng không khí
VOCs	Volatile Organic Compounds	Hợp chất hữu cơ dễ bay hơi
IQR	Interquartile Range	Khoảng tứ phân vị
TSS	Total sum of squares	Tổng bình phương toàn phần
ESS	Explained sum of squares	Tổng bình phương hồi quy
RSS	Residual sum of square	Tổng bình phương phần dư

CSV	Comma Separated Values	Định dạng văn bản phân tách nhau bởi dấu phẩy
NRI	Nocturnal Radiation Inversion	Nghịch nhiệt do bức xạ về đêm
RMSE	Root Mean Squared Error	Sai số bình phương trung bình
MAPE	Mean Absolute Percent Error	Sai số phần trăm tuyệt đối trung bình
JSON	JavaScript Object Notation	Một định dạng dữ liệu
SAS	Statistical Analysis System	Hệ thống phân tích thống kê
SPSS	Statistical Package for the Social Sciences	Giải pháp thống kê cho các ngành khao học xã hội
pH		Một chỉ số xác định tính chất hoá học của nước
H5N1		Một phân nhóm có khả năng gây nhiễm cao của virus cúm gia cầm
TNMT		Tài nguyên môi trường
TP.HCM		Thành phố Hồ Chí Minh
KCN		Khu công nghiệp
CSDL		Cơ sở dữ liệu
QCVN		Quy chuẩn Việt Nam

DANH MỤC CÁC HÌNH VẼ

DANH MỤC CÁC BẢNG BIỂU

MỞ ĐẦU

1. Đặt vấn đề, định hướng nghiên cứu

Vấn đề đảm bảo vệ sinh môi trường đang là vấn đề được nhiều cơ quan chức năng, đơn vị, cộng đồng quan tâm. Không chỉ ở riêng Việt Nam mà ngay cả cộng đồng thế giới đặc biệt chú ý. Môi trường bao gồm các yếu tố tự nhiên và yếu tố vật chất nhân tạo quan hệ mật thiết với nhau, bao quanh con người, có ảnh hưởng tới đời sống, sản xuất, sự tồn tại, phát triển của con người và thiên nhiên [1], vì vậy khi môi trường có sự thay đổi dù ít hay nhiều thì cũng đều kéo theo những hệ lụy vô cùng lớn mà khó có thể giải quyết được. Thực trạng môi trường hiện nay đang là vấn đề nan giải, nhiều đại dịch lớn như dịch SARS, MERS, H5N1 ... những căn bệnh liên quan đến môi trường. Môi trường là chiếc khiên vững chắc bảo vệ loài người từ mọi phía, song dưới sự phát triển của kinh tế, của khoa học công nghệ, đời sống xã hội… con người đã hoàn toàn quên mất rằng điều cần thiết là phải bảo vệ chính bầu không khí mà họ đang hít thở hàng ngày.

Một trong những dẫn chứng tiêu biểu có thể nói tới là Trung Quốc, với lượng dân số tăng đột biến cùng với quá trình công nghiệp hóa từ những năm 1950 nên quốc gia này đã phải đối mặt với vấn nạn ô nhiễm môi trường trong nhiều thập kỷ liên tiếp [13,30,31]. Những năm gần đây tình trạng ô nhiễm không khí ở Bắc Kinh đã đạt tới mức báo động, cụ thể tháng 12/2015 chính phủ Trung Quốc phải đưa ra cảnh báo Đỏ [27,28,29]. Cảnh báo Đỏ là mức cao nhất trong thang cảnh báo ô nhiễm không khí và khói bụi gồm 4 bậc tại Trung Quốc. Theo dữ liệu của Cơ quan Bảo vệ Môi trường Mỹ, nồng độ hạt bụi phân tử PM2.5[1] trong không khí ở mức 300 $\mu g/m^3$ sẽ được coi là rất nguy hiểm. Trong khi đó, nồng độ được ghi nhận ở đây có lúc đạt tới ngưỡng 500 $\mu g/m^3$, cao hơn cả chục lần so với khuyến nghị của tổ chức Y tế Thế giới - WHO [29,32]. Với mức cảnh báo này, tầm nhìn ở những khu vực chịu ảnh hưởng nặng nề nhất sẽ giảm xuống dưới một kilomet, người dân được khuyến cáo ở trong nhà không nên đi ra ngoài, các hoạt động công cộng, các trường học không được phép vui chơi

[1] *PM2,5 là bụi có đường kính động học ≤ 2,5μm, PM10 là các hạt bụi có đường kính động học ≤ 10μm*

ngoài trời và hoạt động ở nhiều công trường xây dựng cũng như nhà máy bị hạn chế theo yêu cầu của nhà chức trách… [27].

Hình 1. Hiện trạng ô nhiễm không khí tại Bắc Kinh, Trung Quốc.

Ảnh hưởng cụ thể nhất của ô nhiễm không khí là đối với sức khỏe con người, tuy chưa có một nghiên cứu trực tiếp nào nhưng một báo cáo mới đây của WHO cho thấy, Trung Quốc là nước phát hiện nhiều trường hợp nhiễm bệnh ung thư và có số ca tử vong nhiều nhất, trong đó điển hình là 4 loại ung thư gan, ung thư thực quản, ung thư dạ dày và ung thư phổi. Theo WHO, ung thư phổi vẫn là căn bệnh phổ biến nhất và gây tử vong nhiều nhất trên thế giới, với khoảng 1,8 triệu ca nhiễm mới và 1,59 triệu ca tử vong trong năm 2012, trong đó hơn 1/3 số trường hợp này xảy ra ở Trung Quốc. Theo giới chuyên gia, hút thuốc lá, ô nhiễm không khí kéo dài và tiếp xúc với các chất gây ung thư là những yếu tố chính làm gia tăng nguy cơ mắc ung thư phổi [2]. Theo một thống kê khác thì trung bình mỗi năm ở Trung Quốc mỗi ngày có khoảng 4400 người chết vì ô nhiễm không khí, và mỗi năm số lượng này có thể lên tới hơn 1.6 triệu người nếu tình trạng ô nhiễm ngày càng diễn biến phức tạp như hiện nay [33]

Ở Việt Nam hiện trạng ô nhiễm không khí hiện nay cũng đang ở mức nguy hiểm bởi nhiều nguyên nhân như ô nhiễm bởi số lượng phương tiện giao thông quá lớn, quy hoạch các khu công nghiệp bừa bãi, các hoạt động sản xuất từ các làng nghề thủ công… đã tạo ra một phần không nhỏ ô nhiễm không khí ảnh hưởng trực tiếp tới con người. Theo số liệu thống kê của Bộ Y tế trong những năm gần đây các bệnh về đường

hô hấp có tỷ lệ mắc cao nhất trên toàn quốc và một trong các nguyên nhân gây bệnh chính là ô nhiễm không khí [2].

Bảng 1. Bảng thống kê tỉ lệ người mắc bệnh có liên quan đến đường hô hấp tại Việt Nam năm 2010-2011

TT	Bệnh	Năm 2010		Năm 2011	
		Số người (trên 100.000 dân)	Tỷ lệ (%)	Số người (trên 100.000 dân)	Tỷ lệ (%)
1	Viêm phổi	420.49	4.2	419.05	4.2
2	Viêm họng, viêm amidan cấp	685.17	6.9	349.89	3.5
3	Viên phế quản và viêm tiểu phế quản	354.46	3.5	272.98	2.7

Gần đây nhất là tháng 3/2016, Trung tâm quan trắc môi trường, Tổng cục môi trường đã tiến hành đo tại Hà Nội, cho thấy giá trị PM10 trung bình ngày quan trắc được đạt giá trị rất cao là 160 µg/m^3 vào ngày 29/2, vượt quy chuẩn cho phép là 150 µg/m^3 [3]. Còn với PM2.5 đều vượt giới hạn cho phép ở tất cả các ngày, trong đó thời điểm cao nhất cũng rơi vào 29/2 với giá trị là 89 µg/m^3, vượt gần 2 lần quy chuẩn cho phép. Hai loại bụi trên thường tăng cao vào giờ cao điểm khi có mật độ phương tiện giao thông đi lại lớn. Tại các đô thị, nguyên nhân chủ yếu gây ô nhiễm bụi là các hoạt động giao thông, hoặc từ các hoạt động sản xuất công nghiệp tại các khu công nghiệp, khu chế xuất xung quanh. Liên hệ với các thành phố lớn như Hà Nội hay TP.HCM có thể thấy được chất lượng không đã bị ô nhiễm nghiêm trọng và ảnh hưởng trực tiếp tới sức khỏe người dân [34].

Một thống kê sức khỏe cho người lao động tại các mỏ than lộ thiên ở Quảng Ninh bởi Trung tâm Y tế Lao động, Tập đoàn Công nghiệp Than khoáng sản Việt Nam, 12/2009 [2] là một dẫn chứng cho thấy ảnh hưởng của ô nhiễm không khí khi con người hít phải sẽ tiềm tàng gây ra những căn bệnh về phổi. Cụ thể:

- Kết quả chụp X quang tim phổi 372 người lao động tại mỏ than Hà Tu, Quảng Ninh cho thấy có 115 người bị nghi bụi phổi, 10 người có biểu hiện nhiều vết mờ ở giữa phổi, hai bên phổi hoặc hạ đòn phổi do xơ hóa tổn thương phổi cũ, 23 người bị viêm phế quản.
- Kết quả chụp X quang tim phổi của 367 người lao động ở Công ty than Đèo Nai thì có 128 người bị nghi bụi phổi, 19 người có biểu hiện nhiều vết mờ ở giữa phổi, hai bên phổi hoặc hạ đòn phổi do xơ hóa tổn thương phổi cũ, 2 người bị quai động mạch chủ giãn.

Các tác nhân do ô nhiễm không khí đều ảnh hưởng trực tiếp tới con người thông qua quá trình hô hấp. Theo cơ quan quốc tế chuyên nghiên cứu về bệnh ung thư thuộc Tổ chức Y tế Thế giới (WHO), đã xếp ô nhiễm không khí là một trong những nguyên nhân hàng đầu gây các căn bệnh ung thư ở người mà ô nhiễm không khí là tác nhân gây ung thư trong môi trường nguy hiểm nhất. [2]

Chính bởi nguyên nhân đó mà hiện nay hầu hết các quốc gia đều chú trọng nghiên cứu, đánh giá tính hình môi trường hiện tại. Qua đó đưa ra những đề xuất, giải pháp phù hợp để hạn chế tình trạng ô nhiễm môi trường. Nhận thức được điều này, ở Việt Nam hiện nay đã và đang tiến hành xây dựng các trạm quan trắc môi trường để đo đạc, phân tích, đánh giá, dự báo ảnh hưởng của môi trường tới cuộc sống con người. Dựa vào dữ liệu quan trắc môi trường, ngành Y tế có thể đưa ra các phân tích, đánh giá ảnh hưởng của các bệnh về da liễu, hô hấp hay các dịch bệnh… từ đó khoanh vùng phạm vi để xứ lý. Hay như thông qua các chỉ số, biểu đồ ô nhiễm các nhà Quản lý có thể quy hoạch các khu dân cư, khu công nghiệp, giao thông… theo chiều hướng có lợi nhất với đời sống con người. Ngoài ra, dựa vào những số liệu này sẽ giúp các nhà Quản lý sẽ đưa ra những đánh giá và giải pháp phù hợp, kịp thời để hạn chế tình trạng ô nhiễm không khí ngày càng gia tăng như hiện nay.

Tại Việt Nam hiện nay có hai hệ thống trạm quan trắc môi trường không khí tự động do Bộ TNMT quản lý đó là mạng lưới quan trắc khí tượng thủy văn và môi trường quốc gia gồm 10 trạm quan trắc và mạng lưới quan trắc môi trường quốc gia do Tổng cục Môi trường quản lý gồm 07 trạm quan trắc. Các trạm quan trắc đa phần là các tram tự động đo đạc các chỉ tiêu quan trắc về khí tượng và ô nhiễm không khí và được đo theo giờ. Dữ liệu sao khi quan trắc được lưu vào bộ nhớ cục bộ và định kì hàng ngày, tuần được nhân viên quan trắc thu thập lại. Hiện nay nguồn dữ liệu quan trắc này khá là kín không được công bố rộng rãi ra bên ngoài chính bởi vậy có nhiều những bất cập trong dữ liệu mà chưa được giải quyết hay nhận được những góp ý, đánh giá của các nhà khoa học chuyên môn.

Trong thực tế với mọi loại dữ liệu nói chung cũng như những dữ liệu quan trắc môi trường nói riêng đều không thể tránh khỏi các vấn đề như dữ liệu không nhất quán, dữ liệu nhiễu và không đầy đủ cụ thể:

- Dữ liệu không nhất quán: Do không sử dụng một chuẩn quy ước khi nhập liệu hay thiết lập thiết bị. Dữ liệu được thiết lập ở những cấu trúc khác nhau, những khác biệt đơn vị đo, về tên cột, về định dạng ngày tháng, thời gian… Gây ra những khó khăn về phân tích và tập hợp dữ liệu.
- Dữ liệu nhiễu: Trường hợp này có thể xảy ra bởi nhiều nguyên nhân như lỗi thiết bị, lỗi truyền dẫn, những dữ liệu mang tính đột ngột nhất thời… Đối với những dữ liệu quan trắc xuất hiện đột ngột rồi biến mất mang tính nhất thời không theo một chiều hướng hay qui luật xác định cần được loại bỏ để đảm bảo chất lượng dữ liệu. Ví dụ, dữ liệu quan trắc nồng độ bụi PM10 cho thấy qui luật hàng ngày là tăng cao vào giờ cao điểm từ 7h-8h và 16h-18h. Nhưng bởi có một đám cháy trong phạm vi trạm quan trắc hoạt động đã phát thải một lượng bụi lớn dẫn tới các giá trị quan trắc đều đạt mức cao trong thời gian từ 6h-15h. Như vậy dữ liệu quan trắc bụi PM10 ngày này chỉ mang tính đột ngột cần được loại bỏ đảm bảo không ảnh hưởng tới toàn bộ dữ liệu.
- Dữ liệu thiếu: có thể xảy ra bởi nhiều lý do khách quan cũng như chủ quan chẳng hạn như những mô đun quan trắc bị hỏng đột xuất, mất điện hoặc do thay đổi vị trí quan trắc….

Với những hiện trang về dữ liệu quan trắc môi trường hiện có thì làm thế nào để có thể hỗ trợ công tác ra quyết định cho nhà quản lý một cách nhanh chóng và chính xác?. Muốn đưa ra một kết quả chính xác thì bộ dữ liệu đầu vào yêu cầu phải được chuẩn hóa, làm sạch và có đủ độ tin cậy. Nhưng với số lượng dữ liệu quan trắc ở Việt Nam là rất lớn với nhiều dữ liệu nhiễu và thiếu. Các trạm quan trắc lại khác nhau về các tham số ô nhiễm, đơn vị đo, cấu trúc dữ liệu khác nhau… nên việc tập hợp dữ liệu rất mất thời gian. Bên cạnh đó việc sử dụng các công cụ phân tích, thống kê hiện tại của các cơ quan quản lý khá là thủ công, chủ yếu là sử dụng công cụ Excel mang nhiều cảm tính nên thời gian xử lí, đánh giá rất chậm và độ chính xác kết quả là không cao.

Chính vì vậy, để giải quyết tình trạng bất cập về dữ liệu quan trắc như trên. Tôi đề xuất xây dựng một qui trình chuẩn hóa nguồn dữ liệu quan trắc môi trường ở Việt Nam. Qui trình này sẽ giúp ích trong việc tổng hợp và làm sạch dữ liệu, giảm thiểu thời gian, công sức phân tích, đánh giá những vấn đề nghiên cứu. Từ quy trình được đề xuất và đánh giá đó tôi tiến hành phân tích thiết kế và xây dựng bộ công cụ hỗ trợ xử

lý dữ liệu quan trắc môi trường theo hướng tự động hóa để giảm tải gánh nặng cho các nhà phân tích, hỗ trợ phân tích dữ liệu một cách tối đa đảm bảo chất lượng cuối cùng của bộ dữ liệu đầu ra.

2. Mục tiêu của luận văn

Trên cơ sở tính cấp thiết và thực tiễn của nguồn dữ liệu quan trắc môi trường, tôi đã tìm hiểu, đề xuất, nghiên cứu chọn ra đề tài "*Nghiên cứu và xây dựng qui trình chuẩn hóa dữ liệu quan trắc môi trường ở Việt Nam*". Đây là một qui trình với nhiều bước thành phần, kết quả mỗi bước đều được đánh giá và phân tích chi tiết. Quy trình được đề xuất và đánh giá thông qua những bộ dữ liệu quan trắc thực tế được cung cấp để kiểm định những phương án được đề xuất. Từ đó đưa ra những hướng nhìn tổng quan nhất về toàn bộ những khía cạnh khác nhau mà dữ liệu có thể gặp phải.

Một mục tiếu nữa có thể nói tới đó chính là "công cụ hỗ trợ xử lý dữ liệu quan trắc môi trường" được xây dựng nhằm thể hiện tính thực tế và ứng dụng của quy trình nghiên cứu đề xuất đối với thực tiến. Với đề tài nghiên cứu này, kết quả mà tôi xây dựng là một hệ thống phần mềm hỗ trợ người dùng xử lý các dữ liệu qun trắc môi trường một cách tự động để tạo ra bộ dữ liệu chuẩn. Từ kết quả này các ngành, các lĩnh vực nghiên cứu khác có thể sử dụng được trực tiếp bộ dữ liệu mà không cần thêm thời gian tổng hợp và tiền xử lý nữa.

Để có thể đạt được mục tiêu nghiên cứu, tôi kết hợp các kiến thức chuyên môn của các ngành môi trường, khí tượng, thống kê với lĩnh vực Công nghê thông tin mà tôi đang theo đuổi. Bên cạnh đó là nghiên cứu, học tập kinh nghiệm của cộng đồng trong nước cũng như quốc tế từ đó áp dụng vào đề tài nghiên cứu mà tôi đã đề xuất.

3. Phạm vi nghiên cứu của luận văn

Đây là một đề tài lớn cần nhiều thời gian và công sức thực hiện bởi vì lượng dữ liệu quan trắc là rất lớn. Đồng thời đặc thù, tính chất cửa từng tham số ô nhiễm lại có những đặc điểm riêng khác nhau, ngoài ra mối tương quan qua lại lẫn nhau cửa các tham số ô nhiễm cũng là một vấn đề cần xem xét, đánh giá.

Với kết qui trình chuẩn hóa dữ liệu được xây dựng có thể áp dụng với tất cả các tham số ô nhiễm hiện đang quan trắc ở Việt Nam. Bộ dữ liệu nghiên cứu được cung cấp bởi Trung tâm quan trắc môi trường, Tổng cục môi trường (CEM). Bộ dữ liệu quan trắc trong thời gian tháng 01/2011 và 01/2012 với nhiều các chỉ tiêu quan trắc khác nhau. Với mỗi loại chỉ tiêu lại có những quy trình xử lý riêng biệt khác nhau phụ thuộc vào những đặc trưng của những chỉ tiêu quan trắc đó. Trong luận văn tôi sẽ đề

xuất ra một qui chình chung có thể áp dụng được với mọi chỉ tiêu quan trắc khác nhau. Từ qui trình đề xuất sẽ đánh giá và thực nghiệm với chỉ tiêu quan trắc PM10 trên hai bộ dữ liệu được cung cấp.

Thông qua qui trình đề xuất và những đánh giá thực nghiệm với chỉ tiêu quan trắc PM10. Tôi đề xuất xây dựng công cụ hỗ trợ xử lý dữ liệu quan trắc môi trường tại Việt Nam một cách tự động hóa. Công cụ được xây dựng trên nền tảng web hỗ trợ tối đa cho người dùng về phân tích, thống kê và xử lý dữ liệu nhiễu và thiếu. Các chức năng của thệ thống được ánh xạ từ duy trình đề xuất sang, mỗi chức nawngc hính là một bước nhro trong quy trình. Người dùng hoàn toàn có thể chạy riêng lẻ từng chức năng của hệ thống hoặc chạy theo một vòng tuần hoàn khép kín. Cụ thể, các chức năng chính được xây dựng bao gồm:

- Nhóm chức năng tìm kiếm dữ liệu: Cho phép tìm kiếm dữ liệu ở nhiều nguồn với những chỉ tiêu quan trắc khác nhau và thời gian khác nhau
- Nhóm chức năng thống kê, đánh giá dữ liêu cơ bản: Đưa ra các chỉ số thống kê như Min, Max, Median, Mean, Mode, Quartile, Range, Variance, Standard Deviation. Tính toán thống kê trung bình ngày/tháng/năm. Thống kê dữ liệu thiếu hoặc tìm ngày vượt qui chuẩn. . Từ những số liệu này người dùng có thể nắm bắt tổng quan được về dữ liệu đang phân tích
- Nhóm chức năng xử lý dữ liệu nhiễu: Giúp tìm và loại bỏ những giá trị bất thường theo ý của người sử dụng
- Nhóm chức năng xác định tương quan: Chức năng giúp phân tích và đưa ra các kết quả so sánh tương quan giữ những chỉ tiêu quan trắc trong một trạm hoặc giữa các trạm với nhau
- Nhóm chức năng điền dữ liệu thiếu: Chức năng giúp điền dữ liệu thiếu cho những giá trị không có số liệu quan trắc thông qua các thuật toán, cách thức cụ thể.

Bởi thời gian hạn hẹp nên trong nội dung luận văn này tôi sẽ tiến hành phân tích thiết kế và xây dựng công cụ hỗ trợ xử lý dữ liệu quan trắc môi với hai nhóm chức năng chính đó là:

- Nhóm chức năng xử lý dữ liệu nhiễu
- Nhóm chức năng xử lý dữ liệu thiếu

4. Nội dung của luận văn

Luận văn thực hiện xuyên suốt trong quá trình từ khi hình thành các khái niệm, ý tưởng nghiên cứu, cho đến khi xây dựng được qui trình chuẩn hóa dữ liệu được các nhà chuyên môn đánh giá và cho ý kiến. Nội dung chính bao gồm các phần sau:

- **Mở đầu**: Đặt ra vấn đề, mục tiêu và giải pháp cho bài toán *"Nghiên cứu và xây dựng qui trình chuẩn hóa dữ liệu quan trắc môi trường ở Việt Nam".*
- **Chương 1:** Giới thiệu tổng quan - các khái niệm cơ bản về môi trường, các tham số ô nhiễm, các phương pháp đánh giá chất lượng không khí và thực trạng ô nhiễm không khí ở Việt Nam hiện nay.
- **Chương 2:** Nghiên cứu các kỹ thuật xử lý dữ liệu. Từ đó đưa ra đề xuất qui trình chuẩn hóa dữ liệu quan trắc môi trường ở Việt Nam.
- **Chương 3:** Thực nghiệm và đánh giá qui trình chuẩn hóa dữ liệu đã đề xuất với dữ liệu quan trắc môi trường thực tế**.**
- **Chương 4:** Nghiên cứu và xây dựng hệ thống hỗ trợ xử lý dữ liệu quan trắc môi trường ở Việt Nam (EnvPro).
- **Kết luận và đề xuất:** Tổng kết lại những kiến thức đã tích lũy, kinh nghiệm được áp dụng trong suốt quá trình thực hiện luận văn. Đưa ra các hướng phát triển trong tương lai.

CHƯƠNG I. TỔNG QUAN VỀ Ô NHIỄM KHÔNG KHÍ

1.1. Không khí và ô nhiễm không khí.

1.1.1. Không khí.

Không khí là lớp vật chất tồn tại ở thể khí và bao trùm lên toàn bộ trái đất. Đặc điểm của nó là không màu, không mùi, không vị. Không khí rất cần thiết cho quá trình hô hấp của các loài động vật cũng như quá trình quang hợp của thực vật, là nguồn gốc của sự sống trên trái đất. Không khí bao gồm các thành phần chính cấu thành là N_2, O_2, Ar và một số thành phần không khí khác [14].

1.1.2. Ô nhiễm không khí.

Không khí cung cấp Oxy cho chúng ta hít thở để duy trì sự sống vì vậy bất kì một sự thay đổi vật lý, sinh học hay hóa học đều có thể được gọi là ô nhiễm không khí. Theo tổ chức Y tế Thế giới định nghĩa thì ô nhiễm không khí là sự hiện diện của một số thành phần trong không khí có nguy hại cho con người cũng như môi trường sống [15].

Một cách hiểu khác đơn giản hơn đó là ô nhiễm không khí là sự có mặt một chất lạ hoặc một sự biến đổi quan trọng trong thành phần không khí, làm cho không khí không sạch hoặc gây ra sự tỏa mùi, có mùi khó chịu, giảm tầm nhìn xa (do bụi). Thuật ngữ "vật gây ô nhiễm không khí" thường được sử dụng để chỉ các phần tử bị thải vào không khí do kết quả hoạt động của con người và tự nhiên gây tác hại xấu đến sức khoẻ con người, các hệ sinh thái và các vật liệu khác nhau. Các "vật gây ô nhiễm không khí" có thể ở thể rắn (bụi, mồ hóng, muội than), ở hình thức giọt (sương mù quang hoá) hay thể khí (SO_2, NO_2, CO...) [4].

1.2. Ảnh hưởng, tác động và các nguồn gây nên ô nhiễm không khí.

1.2.1. Ảnh hưởng và tác động của ô nhiễm không khí.

1.2.1.1. Ảnh hưởng tới thời tiết, khí hậu, khí quyển.

a) Hiệu ứng nhà kính.

Với việc không khí ngày càng ô nhiễm dẫn tới khả năng hấp thụ mặt trời của khí quyển tăng lên tạo ra hiện tượng “hiệu ứng nhà kính”. Thuật ngữ “hiệu ứng nhà kính” có thể hiểu như sau. Ở các vùng lạnh với khí hậu ôn đới, để bảo vệ cây trồng người ta dựng các nhà kính để giữ ổn định nhiệt độ không khí giúp cây cối phát triển một cách bình thường. Nhà kính này chỉ có khả năng ngăn cản sự khuếch tán của ánh sáng mà không có khả năng hấp thụ và bức xạ nhiệt giống như khí quyển.

Với trái đất, khí quyển giống như một lớp kính nó cho phép mặt trời xuyên qua đốt nóng trái đất, đồng thời giữ một phần nhiệt và bức xạ phần còn lại ra vũ trụ. Nhưng hiện nay với hiện trạng ô nhiễm không khí ngày một tăng đã làm nồng độ CO_2, CH_4, SO_2… phát thải ngày càng tăng khiến cho tia sáng mặt trời được hấp thụ và phát tán tạo thành nhiệt lượng trong khí quyển, dẫn tới việc sưởi ấm toàn bộ không gian bên trong chứ không phải chỉ những chỗ có ánh sáng. Vì vậy nhiệt độ trung bình toàn cầu cũng tăng lên. Và hậu quả sẽ xảy ra đó là:

- *Các nguồn nước*: Chất lượng và số lượng của nước uống, nước tưới tiêu, nước cho các máy phát điện và sức khỏe của các loài thủy sản có thể bị ảnh hưởng nghiêm trọng bởi sự thay đổi của các trận mưa rào và bởi sự tăng khí bốc hơi. Mưa tăng có thể gây lụt lội thường xuyên hơn. Khí hậu thay đổi có thể làm đầy các lòng chảo nối với sông ngòi trên thế giới.
- *Các tài nguyên bờ biển*: Nước biển sẽ dâng lên theo sự nóng lên toàn cầu, nếu nhiệt độ của trái đất đủ cao thì có thể làm tan nhanh hơn băng tuyết ở Bắc Cực và Nam Cực và do đó mực nước biển sẽ tăng, có thể dẫn đến nạn hồng thủy.
- *Sinh vật:* Sự nóng lên của trái đất làm thay đổi điều kiện sống bình thường của các sinh vật trên trái đất. Một số loài sinh vật thích nghi với điều kiện mới sẽ thuận lợi phát triển. Trong khi đó nhiều loài bị thu hẹp về diện tích hoặc bị tiêu diệt.
- *Sức khỏe:* Nhiều loại bệnh tật mới đối với con người xuất hiện, các loại dịch bệnh lan tràn, sức khoẻ của con người bị suy giảm. Số người chết vì nóng có thể tăng do nhiệt độ cao trong những chu kì dài hơn trước. Sự thay đổi lượng mưa và nhiệt độ có thể đẩy mạnh các bệnh truyền nhiễm.
- *Lâm nghiệp*: Nhiệt độ cao hơn tạo điều kiện cho nạn cháy rừng dễ xảy ra hơn.
- *Năng lượng và vận chuyển:* Nhiệt độ ấm hơn tăng nhu cầu làm lạnh và giảm nhu cầu làm nóng. Sẽ có ít sự hư hại do vận chuyển trong mùa đông hơn, nhưng vận chuyển đường thủy có thể bị ảnh hưởng bởi số trận lụt tăng hay bởi sự giảm mực nước sông. [35]

b) Mưa Axit

Mưa axit, còn được biết tới như sự lắng đọng axit, được tạo ra bởi lượng khí thải SO_2 và NO từ các nhà máy điện, ô tô và các trung tâm công nghiệp. Mưa axit cũng có thể bắt nguồn từ núi lửa, cháy rừng hay sấm sét khi mà khí SO_2 và NO_x kết hợp với hơi nước trong khí quyển và tạo thành axit dưới 2 dạng là khô như khí gas và ướt như mưa axit, tuyết, sương mù.

Mưa axit được phát hiện ra đầu tiên năm 1948 tại Thuỵ Điển nơi có rất nhiều mỏ than. Đến năm 1960 thì các nhà khoa học mới bắt đầu quan sát và nghiên cứu về hiện tượng này. Và thuật ngữ "mưa axit" được đặt ra bởi Robert Angus Smith vào năm 1972. Các nhà khoa học thể hiện tính axit của mưa bằng thang đo độ pH. Mưa axit có nồng độ pH dưới 5.6; thông thường dao động trong khoảng từ 4.3 đến 5.0 [36].

Tác hại của mưa Axit:

- *Cuộc sống thực vật:* Axit mưa thấm vào đất và cây bằng cách hòa tan các chất độc hại trong đất, chẳng hạn như nhôm được hấp thụ bởi rễ cây. Mưa này cũng hòa tan các khoáng chất có lợi và các chất dinh dưỡng trong đất sau đó được rửa sạch, trước khi các loại cây có cơ hội sử dụng chúng để phát triển. Khi có mưa axit thường xuyên, nó ăn mòn lớp phủ bảo vệ sáp của lá. Khi lớp bảo vệ này trên lá bị mất, hậu quả của nó làm cho cây dễ bị bệnh. Do lá bị hư hỏng làm mất khả năng sản sinh đủ lượng dinh dưỡng cần để khỏe mạnh. Nó là kết quả trong việc làm cho cây dễ bị tổn thương với thời tiết lạnh, côn trùng và bệnh tật, mà có thể dẫn đến cái chết.
- *Cuộc sống dưới nước:* Mưa axit cũng ảnh hưởng xấu đến sinh vật dưới nước. Một số lượng cao của acid sulfuric trong nước biển gây trở ngại cho cá để có chất dinh dưỡng, muối và oxy. Các phân tử axit trong chất nhầy hình thành trong mang cá, làm ngăn chặn hấp thụ oxy với số lượng đầy đủ. Thêm vào đó, nồng độ axit làm giảm độ pH, gây ra sự mất cân bằng muối trong các mô của cá. Sự thay đổi này trong độ pH cũng làm suy yếu một số khả năng của cá để duy trì nồng độ canxi. Nó sẽ ảnh hưởng đến quá trình sinh sản của cá. Thiếu canxi cũng gây ra biến dạng xương và cột sống bị suy yếu.
- *Đối tượng nhân tạo:* Ngoài gây nguy hại cho các hệ sinh thái, mưa axit cũng gây thiệt hại cho cấu trúc và vật liệu nhân tạo. Ví dụ, mưa axit hòa tan đá sa thạch, đá vôi, đá cẩm thạch. Nó cũng ăn mòn sứ, dệt may, sơn, và kim loại. Cao su và da xấu đi nếu tiếp xúc với mưa axit. Di tích đá và chạm khắc dần biến mất khi tiếp xúc với dạng mưa bị ô nhiễm này.
- *Con người:* Hầu hết tất cả, mưa axit ảnh hưởng xấu đến sức khỏe con người. Nó có thể làm hại chúng ta thông qua không khí và ô nhiễm đất. Mưa axit dẫn đến sự hình thành các hợp chất độc hại bằng cách phản ứng với các hợp chất hóa học tự nhiên. Một khi các hợp chất độc hại được hình thành, chúng có thể thấm vào nguồn nước, và cũng thâm nhập vào chuỗi thực phẩm. Thực phẩm bị ô nhiễm này có thể gây tổn hại các dây thần kinh ở trẻ em, hoặc dẫn đến tổn thương não nghiêm trọng, thậm chí tử vong. Các nhà khoa học nghi ngờ

rằng nhôm, một trong những kim loại bị ảnh hưởng bởi mưa axit, có liên quan đến bệnh Alzheimer. Lượng khí thải của nitơ oxit và các vấn đề nguyên nhân sulfur dioxide như kích thích cổ họng, mũi và mắt, đau đầu, hen suyễn và ho khan. [37]

1.2.1.2. Tác động tới sức khỏe con người

Hiện nay tốc độ công nghiệp hóa, đô thị hóa diễn ra với tốc độ nhanh chóng cùng với nó là hiện tượng ô nhiễm không khí tại các đô thị và khu công nghiệp ngày càng gia tăng. Các nguồn ô nhiễm không khí không những gây ra ô nhiễm không khí trong khu vực đô thị và khu công nghiệp, mà còn khuyếch tán đi xa, gây ô nhiễm không khí vùng xung quanh.

Phần lớn các chất ô nhiễm đều gây tác hại đối với sức khoẻ con người, với hai cơ quan chính của con người là mắt và đường hô hấp. Ảnh hưởng cấp tính có thể gây ra tử vong. Ảnh hưởng mãn tính gây ra bệnh ung thư phổi.

Một số chất có ảnh hưởng trực tiếp tới sức khoe con người có thể kể tới như:

- **Khí Cacbon oxit (CO)** là một loại khí không màu, không mùi, không vị. Con người đề kháng với khí CO rất khó khăn. Nó phát sinh từ sự thiêu đốt các vật liệu tổng hợp có chứa cacbon, và chiếm tỷ lệ lớn nhất trong ô nhiễm môi trường không khí. Nồng độ CO cao trong không khí có thể ảnh hưởng đến sự vận chuyển oxygen trong máu, do CO thay thế O_2, liên kết với hemoglobin trong máu [4]
- **Khí SO_2**: Do quá trình tác dụng của quang hoá học hay một xúc tác nào đó mà khí SO_2 dễ dàng bị oxi hoá và biến thành SO_3 trong khí quyển. SO_3 tác dụng với hơi nước trong môi trường không khí ẩm ướt và biến thành axit sulfuric hay là muối sulfat. SO_2 và H_2SO_4 đều có ảnh hưởng xấu đến sức khoẻ của con người và động vật. Ở nồng độ thấp đã gây ra sự kích thích đối với bộ máy hô hấp của con người và động vật, ở mức nồng độ cao sẽ gây ra biến đổi bệnh lý đối với bộ máy hô hấp và có thể gây tử vong. [4]
- **Khí NOx (nitơ oxit)** là khí có màu hơi hồng, mùi của nó có thể phát hiện thấy khi nồng độ của nó vào khoảng 0,12 ppm. Khi trời có mưa, nước mưa sẽ rửa không khí bị ô nhiễm khí NO_2 và hình thành mưa axit. Nitơ oxit (NO) với nồng độ thường có trong không khí nó không gây ra tác hại với sức khoẻ của con người, chỉ nguy hại khi nó bị oxi hoá thành NO_2. Con người tiếp xúc lâu với không khí có nồng độ khí NO_2 khoảng 0,06 ppm đã gây trầm trọng thêm các bệnh về phổi, mắt và nếu nồng độ cao có thể gây ung thư. [4]

- **Bụi**: Bụi là tên chung cho các hạt chất rắn và hạt lỏng có đường kính nhỏ cỡ vài micrômét đến nửa milimét, tự lắng xuống theo trọng lượng của chúng nhưng vẫn có thể lơ lửng trong không khí một thời gian. Bụi được quan trắc bao gồm các loại sau:
    - Bụi lơ lửng tổng số (TSP): là các hạt bụi có đường kính động học ≤100µm
    - Bụi PM10: là các hạt bụi có đường kính động học ≤10µm
    - Bụi PM2,5: là các hạt bụi có đường kính động học ≤2,5µm
    - Bụi PM1: là các hạt bụi có đường kính động học ≤1µm

  Trong các loại bụi này thì bụi PM2.5 có khả năng đi sâu vào các phế nang phổi, gây ảnh hưởng trực tiếp đến hệ hô hấp hơn cả. [2]
- **Pb**: Có mặt trong thành phần khói xả từ động cơ của các phương tiện giao thông (trường hợp nhiên liệu có pha chì). Ngoài ra có thể phát tán từ các mỏ quặng và các nhà máy sản xuất pin, hóa chất, sơn… Thời gian lưu trong khí quyển thường dao động từ 7,5 đến 11,5 ngày [2].

1.2.2. Các nguồn gây nên ô nhiễm không khí.

Tác nhân gây ô nhiễm môi trường không khí chủ yếu bao gồm: Bụi lơ lửng tổng số (TSP), bụi PM10 (bụi ≤ 10µm), chì (Pb), ôzôn (O_3); các chất vô cơ như cacbon monoxit (CO), lưu huỳnh đioxit (SO_2), oxit nitơ (NOx), hydroclorua (HCl), hydroflorua (HF)…; các chất hữu cơ như hydrocacbon (CnHm), benzen (C_6H_6)…; các chất gây mùi khó chịu như amoniac (NH_3), hydrosunfua (H_2S)…; nhiệt, tiếng ồn…. Các tác nhân này được sinh ra và phát tán vào không khí bởi nhiều những nguyên nhân khác nhau có thể phân ra thành hai nguồn chính đó là nguồn do thiên nhiên và nguồn do các hoạt động của con người.

1.2.2.1. Các nguồn từ thiên nhiên.

- Phun núi lửa: Núi lửa phun ra những nham thạch nóng và nhiều khói bụi giàu sulfua, mêtan và những loại khí khác. Không khí chứa bụi lan tỏa đi rất xa vì nó được phun lên rất cao.
- Cháy rừng: Các đám cháy rừng và đồng cỏ bởi các quá trình tự nhiên như sấm chớp, cọ sát giữa thảm thực vật và cỏ khô. Các đám cháy này thường lan truyền rộng, phát thải nhiều bụi và khí.
- Bão bụi gây ra do gió mạnh và bão: Mưa bào mòn đất sa mạc và đất trồng và gió thổi tung lên thành bụi. Nước biển bốc hơi cùng với sóng biển tung bọt mang theo bụi muối lan truyền vào không khí.

- Các quá trình thối rữa của các động vật và thực vật chết ở tự nhiên cũng thải ra các chất khí ô nhiễm.
- Các phản ứng hóa học giữa các khí tự nhiên hình thành các khí sulfua, nitric, các loại muối... [4]

1.2.2.2. Các hoạt động bởi con người.

a) Các hoạt động công nghiệp.

Các ống khói của các nhà máy trong quá trình sản xuất do đốt nhiên liệu đã thải vào môi trường các chất khí như: SO_2, CO_2, CO... bụi và các khí độc hại khác. Hoặc các chất khí bị bốc hơi, rò rỉ thất thoát trong dây chuyền sản xuất, trên các đường dẫn, đã thải vào không khí rất nhiều chất khí độc hại.

Đặc điểm của chất thải công nghiệp là có nồng độ chất độc hại cao và tập trung, Đăc biệt là các ngành công nghiệp năng lượng, công nghiệp dầu khí, công nghiệp hoá chất, công nghiệp luyện kim, công nghiệp cơ khí, công nghiệp vật liệu xây dựng và các ngành công nghiệp nhẹ.. Gây ô nhiễm chính cho môi trường, cụ thể xem tại Bảng 2. Nhìn chung do tính đa dạng của nguồn ô nhiễm công nghiệp mà việc xác định và tìm các biện pháp xử lý ở các khu công nghiệp lớn gặp rất nhiều rất khó khăn. [4]

Bảng 2. Nhóm ngành công nghiệp và khí thải phát sinh điển hình [2].

Nhóm ngành sản xuất	Khí thải
Các ngành có lò hơi, lò sấy, máy phát điện đốt nhiên liệu nhằm cung cấp hơi, điện, nhiệt	Bụi, SO_2, CO, CO_2, NO_2, VOCs, muội khói
Nhóm ngành nhiệt điện	Bụi, CO, CO_2, H_2S, SO_2, và NOx
Nhóm ngành sản xuất xi măng	Bụi, NO_2, CO_2, F
Nhóm ngành sản xuất gang thép	Bụi, gỉ sắt chứa các oxit kim loại (FeO, MnO, Al_2O_3, SiO_2, CaO, MgO); khí thải chứa CO_2, SOx.
Nhóm ngành may mặc: từ công đoạn cắt may, giặt tẩy, sấy	Bụi, Cl, SO_2, Pingment, formandehit, HC, NaOH, NaClO

Nhóm ngành sản xuất cơ khí, luyện kim	Bụi, hơi kim loại nặng, HCl, SiO_2, CO, CO_2
Nhóm ngành sản xuất các sản phẩm từ kim loại	Bụi kim loại đặc thù, hơi hóa chất, hơi dung môi hữu cơ, SO_2, NO_2
Nhóm ngành sản xuất hóa chất	Bụi H_2S, NH_3, hơi dung môi hữu cơ, hóa chất đặc thù, bụi, SO_2, CO, NO_2
Nhóm ngành khai thác dầu thô, khí	CO, SO_2, NOx
Nhóm ngành khai thai sản xuất than và khoáng sản	Bụi, SO_2, NOx, CO, CO_2

b) Giao thông vận tải.

Nguồn ô nhiễm do giao thông vận tải sản sinh ra gần 2/3 khí CO_2 và 1/2 khí CO cùng với khí NO, nó được xem là nguồn gây ô nhiễm lớn đối với môi trường không khí. Đặc điểm nổi bật của các nguồn này là tuy nguồn gây ô nhiễm tính theo đơn vị phương tiện vận tải có quy mô nhỏ nhưng lại tập trung suốt dọc tuyến giao thông nên tác hại lớn. Đặc biệt, ô tô còn gây bụi đất đá đối với môi trường không khí và bụi rất độc hại qua ống xả là bụi hơi chì và tàn khói. Tàu hỏa, tàu thủy, chạy bằng nhiên liệu than hay xăng dầu cũng gây ô nhiễm môi trường tương tự như ôtô.

Các chất gây ô nhiễm không khí chủ yếu sinh ra do khí thải từ quá trình đốt nhiên liệu động cơ bao gồm CO, NOx, SO_2, hơi xăng dầu (CnHm, VOCs), PM10... và bụi do đất cát cuốn bay lên từ mặt đường phố trong quá trình di chuyển (TSP).

Đặc điểm nổi bật của nguồn ô nhiễm giao thông vận tải là nguồn ô nhiễm thấp, di động, khả năng khuyếch tán các chất ô nhiễm giao thông vận tải rất phụ thuộc vào địa hình và quy hoạch kiến trúc các phố phường hai bên đường. Máy bay cũng là nguồn gây ô nhiễm bụi, hơi độc hại và tiếng ồn. Bụi và hơi độc hại do máy bay thải ra nói chung là nhỏ, tính tỷ lệ trên nhiên liệu tiêu hao trên đường bay cũng ít hơn ô tô. Một điều đáng chú ý là máy bay siêu âm bay ở độ cao lớn thải ra khí Nitơ oxit (NO_2) gây hư hại tầng ozon. [4]

c) Sinh hoạt hàng ngày.

Nguồn ô nhiễm do sinh hoạt của con người gây ra chủ yếu là do bếp đun và các lò sưởi sử dụng nhiên liệu gỗ, củi, than, dầu mỏ hoặc khí đốt. Quá trình đốt nhiên liệu

không hoàn toàn đã tạo ra CO_2 và CO. Nhìn chung nguồn ô nhiễm này nhỏ nhưng có đặc điểm là tác động cục bộ trực tiếp trong mỗi gia đình nên có thể để lại hậu quả lớn về lâu dài.

Hiện nay việc dùng than đá để đun nấu tràn lan trong đô thị, đó là điều đáng quan tâm đối với các nhà tập thể có hành lang kín và các căn hộ khép kín, nồng độ CO_2 tại bếp đun thường lớn, có thể gây tại nạn đối với con người.

Cống rãnh và môi trường nước mặt như ao hồ, kênh rạch, sông ngòi bị ô nhiễm cũng bốc hơi, thoát khí độc hại gây ô nhiễm môi trường không khí. Ở các đô thị chưa thu gom và xử lý rác tốt thì sự thối rữa, phân hủy rác hữu cơ do vứt rác bừa bãi hoặc chôn không đúng kỹ thuật cũng là một nguồn gây ô nhiễm không khí.

Các khí ô nhiễm từ các nguồn thải sinh hoạt trên chủ yếu là khí CH_4, H_2S, NH_4, mùi hôi thối làm ô uế không khí các khu dân cư trong đô thị. [4]

d) Nông nghiệp.

Hoạt động chăn nuôi hiện nay đang tồn tại ở hai loại hình đó là trang trại và hộ gia đình. Loại hình chăn nuôi theo mô hình hộ gia đình đang là nguồn gây ô nhiễm khó kiểm soát đối với môi trường không khí tại các khu vực nông thôn. Theo thống kê mỗi năm ngành chăn nuôi gia súc, gia cầm thải ra khoảng 75-85 triệu tấn chất thải. Các chất này sẽ sinh ra các khí như CO_2, H_2S, NOx, CH_4, NH_3… [2]

Về hoạt động trồng trọt không ngừng gia tăng về sản lượng, theo đó là sự gia tăng liều lượng và chủng loại thuốc bảo vệ thực vật, phân bón hóa học. Công tác thu gom, lưu giữ và xử lý các loại hóa chất, vỏ bao bì hóa chất bảo vệ thực vật chưa được quan tâm đúng mức. Tại nhiều nơi, rác thải bỏ ngay tại đồng ruộng, từ đó phát sinh mùi, khí thải gây ảnh hưởng đến môi trường không khí.

Thêm nữa hiện nay, tại các vùng nông thôn, rơm rạ không còn là chất đốt sinh hoạt chủ yếu do có các nhiên liệu khác thay thế như điện, khí gas… Ngoài ra, việc gia tăng số mùa vụ canh tác hàng năm cũng làm gia tăng lượng rơm rạ thải ra môi trường. Biện pháp chính được người dân sử dụng đối với lượng rơm rạ dư thừa nói trên là đốt ngay trên đồng ruộng. Chính vì vậy, sau mỗi vụ thu hoạch, hoạt động đốt rơm rạ đã gây hiện tượng khói mù cho các vùng lân cận. Việc đốt rơm rạ ngoài trời là quá trình đốt không kiểm soát, trong đó sản phẩm chủ yếu là các chất khí: bụi, CO_2, CO, NOx. Khi rơm rạ cháy không hết có thể tạo ra hợp chất Anđêhit và bụi mịn là những chất gây ảnh hưởng xấu tới sức khỏe con người.

e) Làng nghề.

Ô nhiễm môi trường không khí tại các làng nghề có nguồn gốc chủ yếu từ việc sử dụng than làm nhiên liệu (phổ biến là than chất lượng thấp), sử dụng nguyên vật liệu và hóa chất trong dây chuyền công nghệ sản xuất, khí thải chứa các thành phần đặc trưng là bụi, CO_2, CO, SO_2, NOx và chất hữu cơ bay hơi.

f) Chôn lấp và xử lý chất thải rắn.

Bãi rác lộ thiên là nơi tập hợp các loại chất thải rắn, chủ yếu là chất thải rắn sinh hoạt có thành phần hữu cơ cao. Dưới tác động của nhiệt độ, độ ẩm và các vi sinh vật, chất thải rắn hữu cơ bị phân hủy và sản sinh ra các chất khí (CH_4 – 63.8%, CO_2 – 33.6%, và một số khí khác). Ước tính, lượng khí CH_4 và CO_2 phát sinh từ các bãi rác lộ thiên và các khu chôn lấp chiếm 3-19% tổng lượng phát sinh. Lượng khí phát thải tăng khi nhiệt độ tăng [2].

Đối với các bãi chôn lấp, ước tính 30% các chất khí phát sinh trong quá trình phân hủy rác có thể thoát lên trên mặt đất mà không cần một sự tác động nào. Quá trình vận chuyển và lưu giữ chất thải rắn cũng phát sinh mùi từ quá trình phân hủy các chất hữu cơ gây ô nhiễm môi trường không khí. Các khí phát sinh từ quá trình phân hủy chất hữu cơ trong chất thải rắn bao gồm: Amoni có mùi khai, Hy-drosunfur mùi trứng thối, Sunfur hữu cơ mùi bắp cải thối rữa, Mecaptan hôi nồng, Amin mùi cá ươn, Diamin mùi thịt thối [2].

1.3. Thực trạng ô nhiễm không khí ở Việt Nam.

1.3.1. Quy chuẩn đánh giá mức độ ô nhiễm không khí ở Việt Nam.

Bảng 3. Giá trị giới hạn các thông số cơ bản trong môi trường không khí xung quanh theo QCVN 05:2013/BTNMT.

Đơn vị tính µg/m^3

Thông số	TB 1 giờ	TB 8 giờ	TB 24 giờ	TB năm
SO_2	350		125	50
CO	30.000	10.000		
NO_2	200		100	40
O_3	200	120		
TSP	300		200	100
PM10			150	20

PM2.5			20	25
Pb			1,5	0.5

Để đánh giá chất lượng và mức độ ảnh hưởng đến sức khỏe con người do ô nhiễm môi trường không khí, Bộ Tài Nguyên Môi Trường đã đưa ra văn bản qui định Quy chuẩn kỹ thuật quốc gia về chất lượng không khí xung quanh (Bảng 3). Đồng thời sử dụng chỉ số chất lượng không khí (AQI) để thể hiện và đưa ra các cảnh báo cho người dân.

AQI là chỉ số tổng hợp đại diện cho nồng độ của một nhóm các chất ô nhiễm cơ bản trong không khí xung quanh. Giá trị AQI được tính dựa trên kết quả quan trắc các thông số SO_2, CO, NOx, O_3, PM10. Giá trị AQI của từng thông số được hiểu là tỷ lệ giữa giá trị quan trắc được của thông số đó so với giá trị quy chuẩn cho phép tính theo phần trăm. Giá trị AQI tổng hợp là giá trị cao nhất trong các giá trị AQI của từng thông số và được đánh giá theo 5 thang như Bảng 4.

Bảng 4. Các mức cảnh báo AQI ở Việt Nam và ảnh hưởng tới sức khỏe con người. [2]

Khoảng giá trị AQI	Chất lượng không khí	Ảnh hưởng sức khỏe
0 – 50	Tốt	Không ảnh hưởng đến sức khỏe
51 – 100	Trung bình	Nhóm nhạy cảm nên hạn chế thời gian ở bên ngoài
101 – 200	Kém	Nhóm nhạy cảm hạn chế thời gian ở bên ngoài
201 – 300	Xấu	Nhóm nhạy cảm tránh ra ngoài. Những người khác hạn chế ở bên ngoài
Trên 300	Nguy hại	Mọi người nên ở trong nhà

1.3.2. Hệ thống các trạm quan trắc chất lượng không khí.

Hệ thống trạm quan trắc môi trường không khí tự động do Bộ TNMT quản lý, gồm 2 mạng lưới [2]:

- Mạng lưới quan trắc khí tượng thủy văn và môi trường quốc gia: Gồm 10 trạm quan trắc chất lượng không khí tự động và các điểm quan trắc khí tượng do các đài khí tượng thủy văn thực hiện tại các tỉnh/thành phố là Hà Nội, Hải Phòng, Ninh Bình, Vinh, Đà Nẵng, Hồ Chí Minh, Pleiku, Cần Thơ, Sơn La.
- Mạng lưới quan trắc môi trường quốc gia do Tổng cục Môi trường quản lý gồm: 07 trạm
    - Trạm đặt tại địa chỉ 556 Nguyễn Văn Cừ (Hà Nội) vận hành từ tháng 6/2009.
    - Trạm Lăng Chủ tịch Hồ Chí Minh (Hà Nội) vận hành từ tháng 10/2012.
    - Trạm Đà Nẵng vận hành từ tháng 6/2011.
    - Trạm Khánh Hòa vận hành từ tháng 5/2012.
    - Trạm Huế vận hành từ tháng 6/2013.
    - Trạm Phú Thọ vận hành từ tháng 6/2013.
    - Trạm Quảng Ninh bắt đầu vận hành từ tháng 12/2013.

  Hệ thống trạm quan trắc không khí tự động, cố định do địa phương quản lý:

    - Trạm Vĩnh Phúc đi vào vận hành từ 2013.
    - Trạm Đồng Nai vận hành từ năm 2012.

Hệ thống các trạm quan trắc tự động đo nồng độ ô nhiễm không khí của các chất phổ biến như carbon monoxide (CO), oxit nitric (NO), nitơ đioxit (NO_2), sulfur dioxide (SO_2), ozone (O_3) và PM10. Ngoài ra các trạm này còn đo thêm một số các thông tin khí tượng để hỗ trợ cho quá trình đánh giá dữ liệu quan trắc.

Nhìn chung số lượng trạm quan trăc sở Việt Nam có rất ít, nhiều khi dữ liệu của một trạm không đủ để đánh giá chất lượng không khí của cả một khu vực rộng lớn xung quanh đó. Các trạm quan trắc được đo tự động và lưu dữ liệu cục bộ ngay trên bộ nhớ của trạm. Và được định kỳ hàng ngày/tuần được lấy ra và cập nhật vào bộ lưu trữ chung của nhà quản lý. Một số trạm quan trắc hiện đại hơn đã cho phép truyền dữ liệu trực tiếp về hê thống quản lý thông qua các kết nối internet hay các kết nôi không dây nhưng tỉ lệ này là không nhiều.

1.3.3. Ô nhiễm không khí tại nông thôn và các thành phố lớn.

Hiện trạng ô nhiễm không khí tại khu vực nông thôn.

Tình trạng ô nhiễm môi trường tại khu vực nông thôn đang ngấp nghé ở mức báo động bởi ô nhiễm môi trường hiện đang là nỗi bức xúc của nhiều người dân.

Nguyên nhân chính là do việc xử lý chất thải, lạm dụng thuốc bảo vệ thực vật, thuốc trừ sâu… làm cho nguồn nước và không khí ô nhiễm trầm trọng. Người dân tại các khu vực ô nhiễm thường xuyên phải đối mặt với nhiều dịch bệnh nguy hiểm.

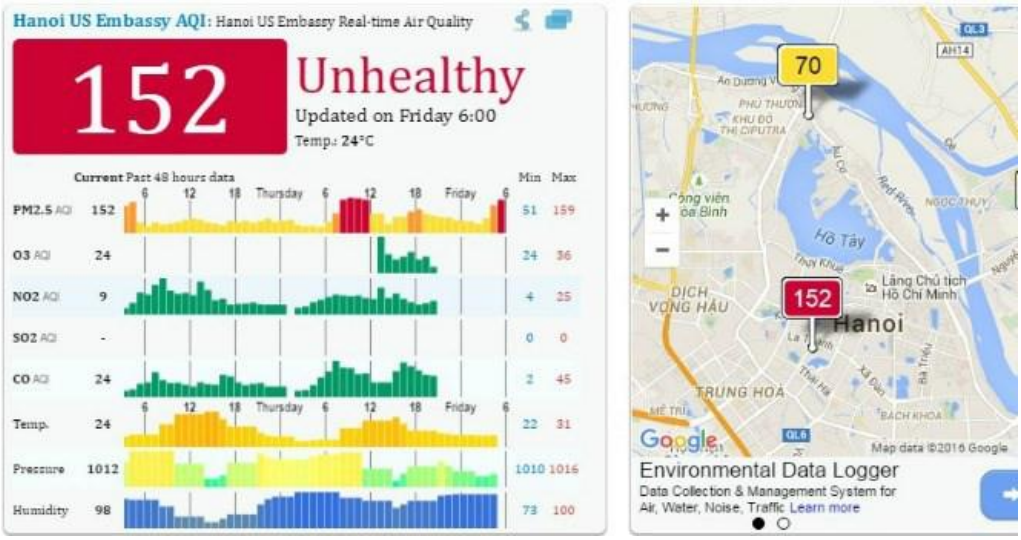
Ngoài các nguyên nhân trên thì chất thải rắn cũng là một nguyên nhân thiết yếu gây nên ô nhiễm không khí. Khác với khu vực thành phố rác thải được thu gom và xử lý tại một khu vực riêng thì ở nông thôn rác thải sinh hoạt của người dân cũng như ở các làng nghề được để tự do không theo quy định như để rác bên lề đường, đổ xuống các ao, hồ, sông ngòi… rác thải không được thu gom và xử lý một cách an toàn. Có những khu vực mà rác thải cách quán ăn, nhà dân, chợ… không quá 20 mét và mùi xú uế của rác thải khiến cho môi trường sống của chính người nông dân ít nhiều bị ảnh hưởng nghiêm trọng . Tại một số khu vực khác thì người dân tự xử lý bằng cách đào hố trôn hay đốt. Với những biển pháp xử lý rác thải còn nghèo nàn cộng với ý thức người dân chưa được cao đã dẫn tới các hiện trạng như ô nhiễm không khí, ô nhiễm nguồn nước mà không biết rằng đó là những thứ mà họ đang trực tiếp sử dụng hàng ngày.

Việt Nam có thể nói là một trong các nước xuất khẩu gạo lớn nhất thế giới tương đồng với điều đó là những ảnh hưởng từ việc trồng trọt, sản xuất nông nghiệp tới môi trường là không hề nhỏ. Đặc biệt tại các khu vực sản xuất nông nghiệp người dân sống chủ yếu bằng nghề làm ruộng. Vào các ngày mùa thì thường xuyên xảy ra các hiện tượng đốt rơm rạ để lấy tro bón ruộng đã gây ra hiện tượng khói mù cộng với gió làm khuyếch tán, tạo ra một khu vực lan tỏa rộng lớn. Khí cacbonic được sinh do đốt cháy nhiên liệu hóa thạch (xăng dầu) và chất thải nông nghiệp (rơm rạ, trấu..) chiếm lượng lớn khí thải nhà kính trên toàn cầu. Theo một nghiên cứu của Gadde [18] cho thấy nếu đốt 1kg rơm rạ sẽ phát thải 1.46 kg carbon dioxide, 34,7 gram carbon mono oxide và 56 gram bụi. Nhìn vào thực tế Việt Nam với hoạt động sản xuất nông nghiệp hàng năm có thể phát thải ra hàng chục triệu tấn carbon dioxide [19]. Ngoài ra với việc đốt rơm rạ như vậy còn ảnh hưởng tới chất lượng đất, những vi sinh vật sống trong môi trường đất [17] ảnh hưởng trực tiếp tới năng xuất sản xuát nông nghiệp. Do đó, môi trường nông thôn đang chịu những áp lực ngay chính từ hoạt động sản xuất và sinh hoạt, đồng thời còn chịu sự tác động từ các cụm công nghiệp, các làng nghề… và các khu đô thị lân cận, đòi hỏi các nhà quản lý phải có những giải pháp đồng bộ nhằm ngăn ngừa và giảm thiểu ô nhiễm.

Ô nhiễm không khí tại các thành phố lớn

Theo số liệu quan trắc mức độ ô nhiễm không khí của Đại sứ quán Mỹ[2] cho thấy, chỉ số AQI - chỉ số dùng để đánh giá chất lượng không khí và khả năng tác động sức khỏe tại Hà Nội lúc 6h sáng ngày 22/04/2016 là "Ulhealthy" với mức đo là 152. Kết quả quan trắc này được đánh giá là ô nhiễm bậc 4/6 theo thang đo ô nhiễm không khí 6 bậc của Mỹ. Đây là mức phản ánh chất lượng không khí kém, không tốt cho sức khỏe với nhóm nhạy cảm là trẻ em và người già, ảnh hưởng trực tiếp tới hệ hô hấp và tim mạch. Trẻ em, người già được khuyến cáo nên hạn chế vận động, tránh các hoạt động gắng sức ngoài trời để bảo đảm sức khỏe.

Nhìn chung những cảnh báo về chất lượng không khí như trên không được phổ cập tới người dân một cách thường trực bởi nhiều nguyên nhân khác nhau cũng như không có các kênh thông tin chính thống để phổ biến. Vì vậy đã số mọi người không hiểu rõ về tầm quan trọng của việc bảo vệ chính sức khỏe của họ cũng như bảo vệ môi trường xung quanh mình.



Hình 2. Kết quả quan trắc không khí của Đại sứ quán Mỹ tại Hà Nội ngày 22/04/2016.

Hiện trạng ô nhiễm không khí ở Việt Nam hiện nay chủ yếu xảy ra ở các thành phố lớn như Hà Nội và Hồ Chí Minh. Ở một vài địa phương khác nồng độ ô nhiễm vẫn nằm trong tiêu chuẩn cho phép. Ở các thành phố lớn như Hà Nội và Hồ Chí Minh

[2] http://aqicn.org/city/vietnam/hanoi/us-embassy/

với dân số tập trung đông và tăng dần hàng năm cộng với quá trình đô thị hóa nhanh chóng đã khiến các thành phố này trở nên đông đúc. Theo thống kê thì dân số ở Hà Nội là gần 7 triệu người và Hồ Chí Minh là 7.5 triệu người[3]. Đây mới chỉ là những con số dựa trên số liệu được đăng kí của người dân còn trên thực tế số lượng này chắc chắn cao hơn rất nhiều. Điều này đã tạo nên những áp lực lớn về cơ sở hạ tầng, đô thị cũng như các vấn đề giao thông như ùn tắc, quá tải. Từ đó dẫn tới những ảnh hưởng về chất lượng không khí mà đứng đầu có thể nói tới đó là ô nhiễm bởi giao thông (Bảng 5, Bảng 6), sau đó là hoạt động công nghiệp.

Với số lượng lớn sử dụng các phương tiện cá nhân mà chủ yếu là xe máy đã gây ra hậu quả là sự gia tăng liên tục một cách ồ ạt đến mức khó có thể quản lý được. Theo Báo cáo của trung tâm Quan trắc môi trường quốc gia [2] thì tác nhân giao thông là nguyên nhân chính gây ra ô nhiễm không khí ở Việt Nam.

Bảng 5. Bảng thống kê tăng trưởng phương tiện giao thông cơ giới trên địa bàn TPHCM 2000 – 2010 [38]

Năm	Tổng số phương tiện quản lý (xe)		
	Ô tô	Xe máy	Tổng số
2000	131.182	1.569.355	1.700.537
2001	144.407	1.968.872	2.113.279
2002	158.172	2.284.870	2.443.042
2003	221.665	2.305.415	2.527.080
2004	252.861	2.428.989	2.681.850
2005	267.815	2.557.621	2.825.436
2006	296.143	2.917.502	3.213.645
2007	326.679	3.338.913	3.665.592
2008	366.066	3.659.473	4.025.529
2009	408.688	4.071.567	4.480.255
2010	438.030	4.401.317	4.839.347

[3] Theo số liệu thống kê của Tổng cục Thống Kê năm 2011

2001-2005 (%/năm)	***15,3***	***10,3***	***10,7***
2006-2010 (%/năm)	***10,3***	***11,5***	***11,4***
2001-2010 (%/năm)	***12,8***	***10,9***	***11,0***

Số lượng các phương tiện cá nhân ở Việt Nam đang gia tăng nhanh chóng ở hai thành phố lớn là Hà Nội và TP. Hồ Chí Minh. Riêng ở thành phố Hồ Chí Minh từ năm 2000-2010 tổng số các phương tiện cá nhân được đăng kí mới tăng đều đặn hàng năm với tỉ lệ tăng là 11%/năm bao gồm cả ô tô và xe máy, tương ứng với hơn hai trăm nghìn phương tiện được sử dụng mới hàng năm. Tính đến hết năm 2011, tổng số phương tiện giao thông trên địa bàn thành phố là gần 5.524.000 xe, trong đó xe ô tô là hơn 494.000 xe. Tính đến hết tháng 11/2012 số phương tiện tăng lên thành hơn 5.899.000 xe, với gần 515.000 ô tô. Và đến hết tháng 3/2013 số phương tiện trên địa bàn TP.HCM đã chính thức vượt qua con số 6 triệu chiếc [38]. Đối với Hà Nội, theo số liệu thống kê thì số lượng xe máy tăng tới 400% trong giai đoạn 10 năm từ 1996 đến 2006, tổng số xe cơ giới gia tăng hàng năm ở mức 10% với ô tô và 15% với xe máy. THống kê chi tiết được mô tả ở Bảng 6.

Bảng 6. Số lượng phương tiện được thống kê ở Hà Nội từ năm 1990-2006[16]

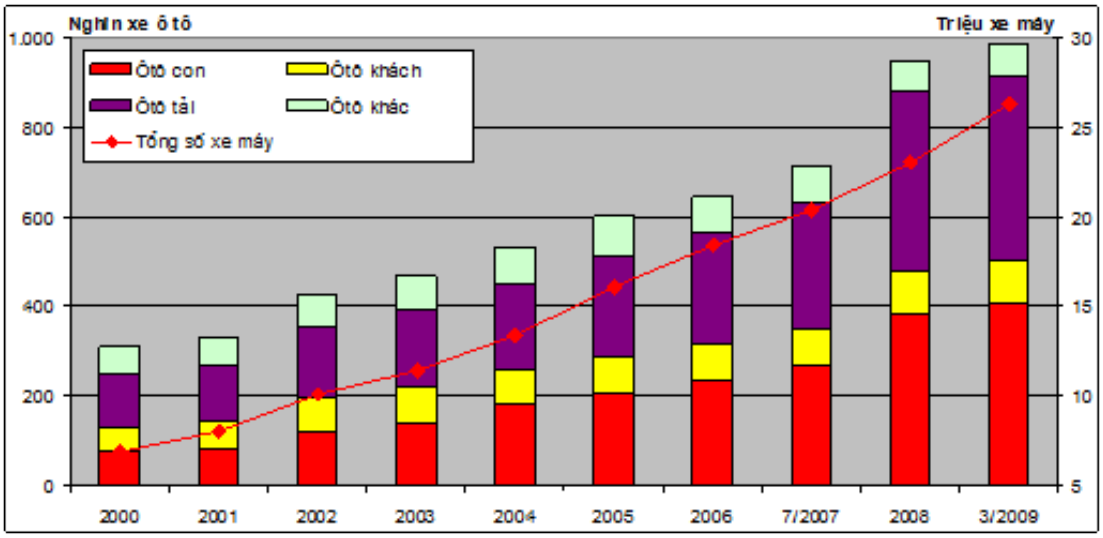
Năm	Ô tô	Xe máy	Tổng
1990	34.222	195.447	229.669
1995	60.231	498.468	558.699
2000	96.679	785.969	759.029
2003	126.478	1.179.166	1.323.644
2006	157.000	1.700.000	1.857.000

Đồng tình với nhận định giao thông là nguồn gây ô nhiễm chính tại các thành phố lớn, Dang.PN [5] cho rằng các hoạt động giao thông vận tải chính là nguồn phát thải lớn nhất ứng với xu hướng tăng nhanh chóng về số lượng phương tiện cá nhân, được mô tả như Hình 3. Ông cho rằng dựa trên các nguồn thải gây ra ô nhiễm không khí trên phạm vi toàn quốc (bao gồm cả khu vực đô thị và khu vực khác). Ước tính cho thấy (Bảng 7), hoạt động giao thông đóng góp tới gần 85% lượng khí CO, 95% lượng VOCs. Trong khi đó, các hoạt động công nghiệp là nguồn đóng góp khoảng 70% khí

SO_2. Đối với NO_2, hoạt động giao thông và hoạt động sản xuất công nghiệp có tỷ lệ đóng góp xấp xỉ nhau

Bảng 7. Ước tính thải lượng các chất gây ô nhiễm từ các nguồn thải chính của Việt Nam năm 2005 (Đơn vị: tấn/năm)[34]

TT	Ngành sản xuất	CO	NO_2	SO_2	VOCs
1	Nhiệt điện	4.562	57.263	123.665	1.389
2	Sản xuất công nghiệp, dịch vụ, sinh hoạt	54,004	151,031	272,497	854
3	Giao thông vận tải	301.779	92.728	18.928	47.462
	Tổng	**360.345**	**301.022**	**415.090**	**49.705**



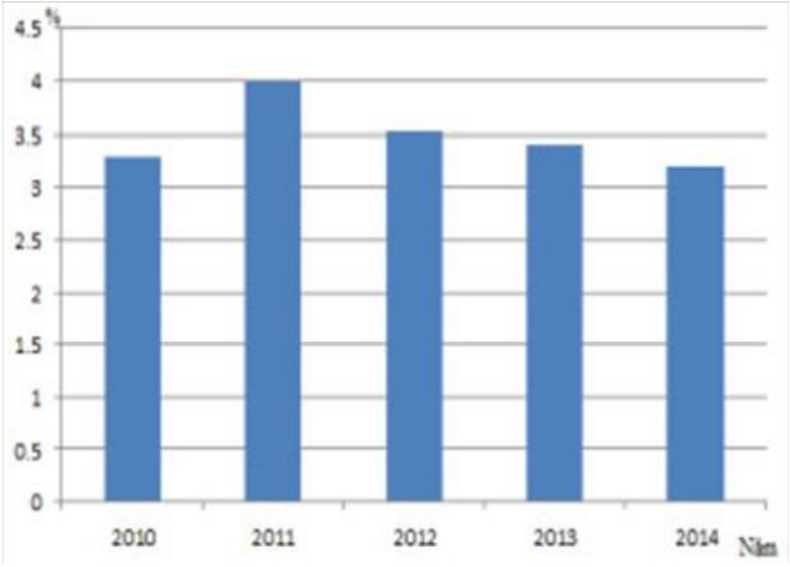
Hình 3. Xu hướng các phương tiện cá nhân gia tăng nhanh chóng từ 1996-2007 ở Việt Nam [34]

Ảnh hưởng của ô nhiễm không khí tới sức khỏe

Ô nhiễm không khí có ảnh hưởng rất lớn đến sức khỏe con người, đặc biệt đối với đường hô hấp. Khi môi trường không khí bị ô nhiễm, sức khỏe con người bị suy giảm, quá trình lão hóa trong cơ thể bị thúc đẩy, chức năng phổi bị suy giảm, gây bệnh hen suyễn, ho, viêm mũi, viêm họng, viêm phế quản, suy nhược thần kinh, tim mạch và làm giảm tuổi thọ con người. Nguy hiểm nhất là có thể gây ra bệnh ung thư phổi. Các nhóm cộng đồng nhạy cảm nhất là những người cao tuổi, phụ nữ mang thai, trẻ em dưới 15 tuổi, người đang mắc bệnh phổi và tim mạch, người thường xuyên làm việc ngoài trời...

Mức độ ảnh hưởng của ô nhiễm không khí đối với từng người tùy thuộc và tình trạng sức khỏe, nồng độ, loại chất ô nhiễm và thời gian tiếp xúc với môi trường ô nhiễm.

Theo số liệu thống kê của Sở Y tế Quảng Trị [40], tỷ lệ người dân mắc bệnh hô hấp liên quan đến ô nhiễm không khí trong tỉnh không biến động nhiều trong giai đoạn 2010 – 2014 như Hình 4. Mức giao động tại khu vực này nhìn chung là không cao và xu hướng giảm cũng chỉ ở mức độ nhẹ. do Quảng Trị có lượng dân cư thấp cùng với tỉ lệ các nhà máy, khu công nghiệp còn nhỏ lẻ. Trái lại ở Hà Nội tỷ lệ mắc các bệnh về đường hô hấp của dân cư sống gần các khu công nghiệp (KCN) cao hơn nhiều so với vùng nông thôn (Bảng 8). Tỷ lệ mắc bệnh viêm phế quản mãn tính ở vùng đô thị, công nghiệp (khu Thượng Đình (Hà Nội) chiếm 14,6%) cao gấp 2,32 lần so với vùng nông thôn (khu Kim Bảng (Hà Nam) chiếm 6,3%). Tại Hải Phòng, nghiên cứu cho thấy tất cả các triệu chứng và bệnh tật liên quan đến đường hô hấp ở nơi bị ô nhiễm đều cao hơn nơi không bị ô nhiễm từ 1,91 đến 7,6 lần [39].



Hình 4. Tỷ lệ mắc bệnh vè đường hô hấp liên quan đến ô nhiễm không khí từ năm 2010-2014 tại Quảng Trị (%) [40]

Năm 2006, nhận thấy rằng các bệnh lý liên quan đến ô nhiễm không khí ngày càng gia tăng, nhất là ở trẻ em là thực trạng rất đáng lo ngại. Số lượng trẻ đến khám, điều trị các bệnh đường hô hấp tại Bệnh viện Nhi đồng 1 (TP.HCM) đã cho thấy điều đó: Nhiễm khuẩn ở đường hô hấp từ gần 2.800 trường hợp năm 1996 tăng lên gần 3.800 trường hợp vào năm 2005; bệnh suyễn từ hơn 3.000 trường hợp năm 1996 tăng

lên trên 11.000 trường hợp vào năm 2005; bệnh viêm tai giữa: từ chỉ 441 trường hợp năm 1996 tăng lên gần 2.000 trường hợp năm 2005... . [39]

Tương tự, tại Bệnh viện Nhi đồng 2 (TP.HCM), lượng bệnh nhi mắc các bệnh lý đường hô hấp (như: viêm họng, viêm phế quản, viêm phổi, hen phế quản...) đến khám cũng ngày càng gia tăng - chiếm 40% - 50% số bệnh nhi nhập viện điều trị nội trú tại đây. [22]

Bảng 8. Bảng thống kê và dự báo số trường hợp bị ảnh hưởng đến sức khoẻ do ô nhiễm không khí ở Hà Nội tới năm 2020. [39]

Các tác động	Số trường hợp		
	2005	2010	2020
Chết	572	1.260	2.824
Viêm phổi mãn tính ở người lớn	987	2.174	4.872
Viêm phổi cấp tính ở trẻ em	8.890	19.580	43.889
Nhập viện vì đường hô hấp	233	513	1.150
Nhập viện vì tim mạch	204	450	1.008
Cấp cứu	9.617	21.181	47.479
Khó thở	18.478	260.942	584.916
Hạn chế các hoạt động trong ngày	1.563.910	3.444.434	7.720.888
Ngày có triệu chứng đường hô hấp	7.476.373	16.466.340	36.910.203

1.4. Kết luận

Chương này tổng hợp và tìm hiểu về tất cả các khái niệm, kiến thức cũng như hiểu biết chung về ô nhiễm không khí, các nguyên nhân cũng như tác hại của ô nhiễm không khí gây ra với con người cũng như môi trường sống của các loài động, thực vật trên trái đất. Đây là những thông tin tổng quát và cốt lõi nhất hỗ trợ cho quá trình xử lý dữ liệu quan trắc môi trường. Ngoài ra với việc đánh giá những hiện trạng về ô nhiễm không khí ở Việt Nam hiện nay đã đưa ra một cái nhìn khách quan và tổng quát nhất làm rõ hơn những mối liên hệ qua lại, ràng buộc giữa những tác nhân gây ra ô nhiễm không khí. Với những thông tin thực tế thu được từ hiện trạng ô nhiễm không khí ở Việt Nam ta thấy rằng dữ liệu về ô nhiễm không khí có một tầm quan trọng vô cùng lớn. Nếu muốn thực hiện bất cứ một quyết định nào về giao thông, y tế, quy hoạch đô thị… thì cũng đều cần có một bộ giữ liệu tin cậy hỗ trợ quá trình ra quyết định của nhà quản lý một cách an toàn và chính xác.

CHƯƠNG 2. NGHIÊN CỨU VÀ ĐỀ XUẤT QUY TRÌNH CHUẨN HÓA DỮ LIỆU QUAN TRẮC MÔI TRƯỜNG Ở VIỆT NAM.

2.1 Tổng quan về quy trình làm sạch dữ liệu.

Thực tế thường có 4 lĩnh vực liên quan tới phân tích, trích xuất thông tin từ dữ liệu bao gồm thống kê (statistics), máy học (Machine Learning), cơ sở dữ liệu (Database) và biểu diễn tri thức (Visualization). Trong 4 lĩnh vực này thì thống kê đóng vài trò rất quan trọng trong quá trình khai phá dữ liệu nhất là trong kiểm định kết quả của mô hình và trong đánh giá tri thức phát hiện được. Trong chương này tôi sẽ trình bày những khái niệm cơ bản nhất và hay được sử dụng trong thống kê được dùng tỏng công tác xử lý dữ liệu quan trắc môi trường. Từ đó đưa ra phương pháp cũng như quy trình cụ thể để chuẩn hóa dữa liệu quan trắc môi trường ở Việt Nam.

2.1.1 Đánh giá dữ liệu dựa trên thống kê.

Trước khi tiến hành, thực hiện các phương pháp xử lý dữ liệu chúng ta cần phải có những cái nhìn tổng quát nhất về dữ liệu, từ đó có thể phát hiện ra những đặc tính của dữ liệu cũng như phát hiện ra những dữ liệu nhiễu hay ngoại lai. Quan trọng hơn cả là đưa ra được những phương pháp xử lý phù hợp với dạng dữ liệu mà chúng ta quan tâm. Một trong những phương pháp cơ bản nhất đó là dựa trên thống kê để mô tả dữ liệu như các tham số đo lường xu hướng tập trung của dữ liệu (Mean, Median, Mode) và đo lường sự biến thiên của dữ liệu (Range, Variance và Standard Deviation).

2.1.1.1 Độ tập trung dữ liệu

Để dễ hình dung, ta bắt đầu với ví dụ đơn giản sau:

Giả sử rằng bạn chạy 100 m trong sáu lần, mỗi lần chạy bạn dùng đồng hồ đo lại thời gian chạy (tính bằng giây) và kết quả 6 lần chạy của bạn gồm sáu giá trị (còn gọi là quan sát) như sau:

X = {25.1, 21.2, 17.9, 23.0, 24.6, 19.5}

Dữ liệu này cho ta biết những thông tin gì? Sau đây là một số thống kê đơn giản của dữ liệu về thời gian chạy 100m:

- Thời gian chạy trung bình (mean) là 21.9 giây
- Giá trị giữa (còn gọi là trung vị - median) là 22.1 giây
- Thời gian chạy lớn nhất (maximum) là 25.1 giây và thời gian chạy nhỏ nhất (minimum) là 17.9 giây.

- Phương sai (variance) là 8.2 giây bình phương và độ lệch chuẩn (standard Deviation) là 2.9 giây

Vậy để đo lường xu hướng tập trung của dữ liệu người ta thường dùng 3 tham số đó là số trung bình (trung bình số học - Arithmetic mean hay average), số trung vị (median) và số mode

Mean (số trung bình): Trung bình số học được tính đơn giản bằng tổng của tất cả các giá trị của dữ liệu trong mẫu chia cho kích thước mẫu n.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1)$$

Với dữ liệu về chạy 100m trên ta có

$$\bar{x} = \frac{25.1 + 21.2 + 17.9 + 23.0 + 24.6 + 19.5}{6} = 21.9\ (giây)$$

Median (trung vị): Trong lý thuyết xác suất và thống kê, số trung vị (Median) là giá trị giữa trong một phân bố được chia thành 2 nhóm, mà trong đó số lượng các số trong mỗi nhóm bằng nhau. Nói cách khác, nếu m là trung vị của một phân bố nào đó thì 1/2 cá thể trong phân bố đó có giá trị nhỏ hơn hay bằng m và một nửa còn lại có giá trị bằng hoặc lớn hơn m.

Công thức chung để tính median là:

$$median = \frac{x\left(\frac{n}{2}\right) + x(\frac{n}{2} + 1)}{2} \quad (2)$$

Để tính được median trước hết dãy số phải được sắp xếp theo thứ tự tăng dần (17.9, 19.5, 21.2, 23.0, 24.6, 25.1), sau đó là xác định vị trí 2 số nằm giữa của dãy số để tính trung bình. Nếu số giá trị là một số chẵn thì median là trung bình của 2 giá trị ở giữa. Với số liệu X quan sát trên ta có n bằng 6. Áp dụng vào công thức (2), vậy median là

$$median = \frac{x_3 + x_4}{2} = \frac{21.2 + 23.0}{2} = 22.1$$

Mode: Mode là độ đo thể hiện dữ liệu xuất hiện với tần xuất cao nhất trong tập dữ liệu, với quan sát X trên ta có thể hiểu nôm na là số có tần suất xuất hiện nhiều nhất trong mẫu. Nếu trong mẫu không có số nào xuất hiện lặp lại thì không có mode. Mode rất hữu ích đối với dữ liệu có kiểu dữ liệu phân loại (nominal). Đối với các dữ liệu có

kiểu phân loại ta không thể dùng Mean hay Median vì nó không có ý nghĩa gì mà phải dùng Mode. Ví dụ nếu dữ liệu mô tả giới tính dạng nominal với 1 là nam, 0 là nữ thì Mean hay Median là 0.5 không có ý nghĩa gì. Trong khi đó Mode cho biết tần suất nam hay nữ xuất hiện nhiều nhất trong tập quan sát.

Trong 3 tham số Mean, Mode và Median thì Median có khả năng đo lường xu hướng tập trung của dữ liệu mạnh nhất vì nó không bị ảnh hưởng nhiều bởi dữ liệu ngoại lai.

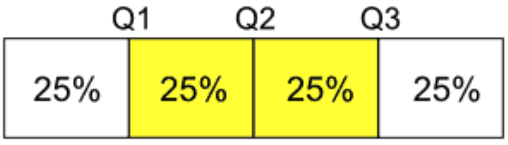
2.1.1.2 Độ phân tán dữ liệu

Quartiles (tứ phân vị)

Tứ phân vị là đại lượng mô tả sự phân bố và sự phân tán của tập dữ liệu. Tứ phân vị có 3 giá trị, đó là tứ phân vị thứ nhất (Q1), thứ nhì (Q2), và thứ ba (Q3). Ba giá trị này chia một tập hợp dữ liệu (đã sắp xếp dữ liệu theo trật từ từ bé đến lớn) thành 4 phần có số lượng quan sát đều nhau.

Tứ phân vị được xác định như sau:

- Sắp xếp các số theo thứ tự tăng dần.
- Cắt dãy số thành 4 phần bằng nhau.
- Tứ phân vị là các giá trị tại vị trí cắt.



Công thức xác định vị trí tứ phân vị:

$$Q_1 = \frac{25 * (n+1)}{100} \quad (3)$$

$$Q_3 = \frac{75 * (n+1)}{100} \quad (4)$$

$$Q_2 = Median \quad (5)$$

Ví dụ với dãy số sau: 5, 8, 4, 4, 6, 3, 8

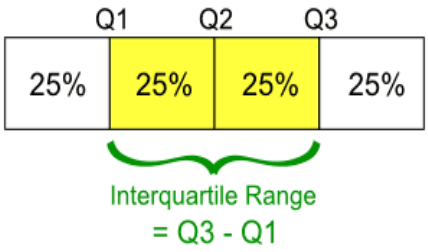
- Sắp xếp: 3, 4, 4, 5, 6, 8, 8
- Chia thành 4 phần:

Kết quả là

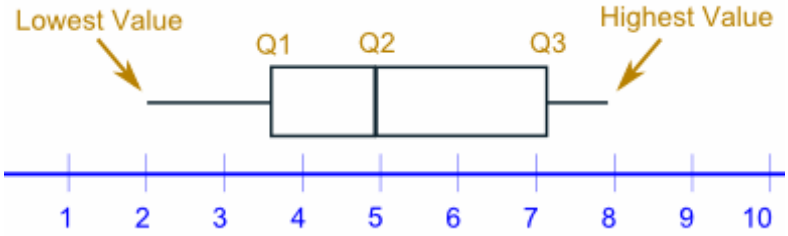
- Quartile 1 (Q1) = 4
- Quartile 2 (Q2/Tứ phân vị thứ 2 chính là số trung vị Median) = 5
- Quartile 3 (Q3) = 8

Khoảng tứ phân vị (Interquartile Range - IQR)

Interquartile Range được xác định như sau: Lấy giá trị tứ phân vị Q3 – Q1. IQR là thước đo độ biến thiên,cho biết độ biến thiên của 50% số đơn vị ở giữa thông qua sự chênh lệch giữa tứ phân vị thứ ba và tứ phân vị thứ nhất.



Quartiles được thể hiện một cách chi tiết hơn thông qua biểu đồ box plot, box plot giúp ta biểu diễn các đại lượng quan trọng của dãy số như Min, Max, Quartile, Interquartile Range một cách trực quan, dễ hiểu. Một Box plot được mô tả như sau



Hình 5. Mẫu mô tả các giá trị của một boxplot.

2.1.1.3 Độ biến thiên dữ liệu

Để biết xu hướng tập trung của dữ liệu ta dùng các tham số như Mean, Median, Mode. Tuy nhiên, một câu hỏi quan trọng nữa cần phải trả lời khi xem xét một chất lượng của mẫu là "làm sao đo lường sự biến thiên (hay sự phân tán) của dữ liệu trong mẫu?". Vì có thể 2 mẫu có cùng trung bình nhưng sự biến thiên của dữ liệu là khác nhau.

Để đo lường sự biến thiên (thường so với giá trị trung bình) của dữ liệu người ta thường dùng các tham số Range (phạm vi), Standard Deviation (độ lệch chuẩn), Variance (phương sai)

Khoảng biến thiên (Range): Được tính bằng cách lấy giá trị lớn nhất trừ giá trị nhỏ nhất

$$Range = Max - Min \quad (6)$$

Trong sample gồm 6 quan sát về thời gian chạy 100 m trong ví dụ trên ta có Range = 25.1- 17.9 = 7.2 giây

Phương sai (Variance) và độ lệch chuẩn (Standard Deviation)

Để tránh tổng các độ lệch bằng 0 và loại bỏ ảnh hưởng của kích thước mẫu người ta tính tổng bình phương các độ lệch và chia cho kích thước mẫu trừ 1 (hiệu chỉnh). Ta có kết quả là "trung bình tổng bình phương các độ lệch" và gọi là phương sai mẫu (Sample Variance - s^2)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (7)$$

Phương sai là tham số rất tốt để đo lường sự biến thiên (hay phân tán) của dữ liệu trong mẫu vì nó đã quan tâm đến độ lệch của mỗi quan sát so với số trung bình, loại bỏ ảnh hưởng của kích thước. Tuy nhiên, điểm yếu của phương sai là không cùng đơn vị tính với Mean. Đơn vị tính của phương sai là bình phương của đơn vị tính của trung bình. Chẳng hạn, đơn vị tính của thời gian chạy trung bình là giây, trong khí đó đơn vị tính của phương sai là giây bình phương. Để giải quyết vấn đề này, người ta lấy căn bậc 2 của phương sai và kết quả này gọi là độ lệch chuẩn (Standard Deviation - *s*).

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (8)$$

2.1.2 Khử nhiễu và điền dữ liệu thiếu.

Kỹ thuật này là một trong những bước tiền xử lý dữ liệu với mục đích, loại bỏ nhiễu, xử lý giá trị ngoại lai và điền giá trị thiếu cho bộ dữ liệu. Chúng ta không thể tin tưởng vào kết quả thu được từ bất kì thuật toán khai phá dữ liệu nào nếu ta biết chắc rằng dữ liệu chưa được làm sạch và có chất lượng tốt. Kỹ thuật này tìm cách tính toán, loại bỏ và làm mịn các giá trị nhiễu trong quá trình xác định đặc trưng và đưa ra hướng xử lý với những dữ liệu nhiễu, thiếu nhằm đảm bảo bộ dữ liệu cuối cùng được xây dựng có độ tin cậy và hoàn thiện cao nhất.

2.1.2.1 Dữ liệu thiếu

Trong trường hợp dữ liệu có rất nhiều bản ghi có các thuộc tính không có dữ liệu, liệu có cách nào để xử lý lấp đầy những vị trí thiếu như vậy không? Ta có một số phương pháp như sau:

- Bỏ qua các bộ: Điều này thường được thực hiện khi thông tin nhãn dữ liệu bị mất. Phương pháp này không phải lúc nào cũng hiệu quả trừ khi các bộ có chứa một số thuộc tính không thực sự quan trọng.
- Điền vào các giá trị thiếu bằng tay: Phương pháp này thường tốn thời gian và có thể không khả thi cho một tập dữ liệu nguồn lớn với nhiều giá trị bị thiếu.
- Sử dụng các giá trị quy ước để điền vào cho giá trị thiệu: Thay thế các giá trị thuộc tính thiếu bởi cùng một hằng số quy ước, chẳng hạn như một nhãn ghi giá trị "Không biết" hoặc "∞". Tuy vậy điều này cũng có thể khiến cho chương trình phân tích dữ liệu hiểu nhầm trong một số trường hợp và đưa ra các kết luận không hợp lý.
- Sử dụng các thuộc tính có nghĩa là để điền vào cho giá trị thiếu: Ví dụ, ta biết thu nhập bình quân đầu người của một khu vực là 800.000đ, giá trị này có thể được dùng thể thay thế cho giá trị thu nhập bị thiếu của khách hàng trong khu vực đó.
- Sử dụng các giá trị của các bộ cùng thể loại để thay thế cho giá trị thiếu: Ví dụ, nếu khách hàng A thuộc cùng nhóm phân loại theo rủi ro tín dụng với một khách hàng B khác trong khi đó khách hàng này có thông tin thu nhập bình quân. Ta có thể sử dụng giá trị đó để điền vào cho giá trị thu nhập bình quân của khách hàng A .
- Sử dụng giá trị có tỉ lệ xuất hiện cao để điền vào cho các giá trị thiếu.: Điều này có thể xác định bằng phương pháp hồi quy, các công cụ suy luận dựa trên lý thuyết Bayersian hay cây quyết định.

2.1.2.2 Dữ liệu nhiễu

Nhiễu dữ liệu là một lỗi ngẫu nhiên hay do biến động của các biến trong quá trình thực hiện, hoặc sự ghi chép nhầm lẫn không được kiểm soát…vậy làm cách nào để có thể làm mịn để loại bỏ dữ liệu nhiễu.

Ví dụ ta có mảng lưu giá cả các mặt hàng được bán là: {4, 8, 15, 21, 24, 25, 28, 34}

- Phân thành các bin

  - Bin1: 4, 8,15
  - Bin2: 21, 21, 24
  - Bin3: 25, 28, 34
- Làm mịn sử dụng phương pháp trung vị
  - Bin1: 9, 9, 9
  - Bin2: 22, 22, 22
  - Bin3: 29, 29, 29
- Làm mịn biên
  - Bin1: 4, 4, 15
  - Bin2: 21, 21, 24
  - Bin3: 25, 25, 34

Ví dụ trên áp dụng phương pháp làm mịn có tên là Binning

Binning: Làm mịn một giá trị dữ liệu được xác định thông qua các giá trị xung quanh nó. Ví dụ, các giá trị giá cả được sắp xếp trước sau đó phân thành các dải khác nhau có cùng kích thước 3 (tức mỗi "Bin" chứa 3 giá trị).

- Khi làm mịn trung vị trong mỗi bin, các giá trị sẽ được thay thế bằng giá trị trung bình các giá trị có trong bin
- Làm mịn biên: các giá trị nhỏ nhất và lớn nhất được xác định và dùng làm danh giới của bin.

Các giá trị còn lại của bin sẽ được thay thế bằng một trong hai giá trị trên tùy thuộc vào độ lệch giữa giá trị ban đầu với các giá trị biên đó. Ví dụ, bin 1 có các giá trị 4, 8, 15 với giá trị trung bình là 9. Do vậy nếu làm mịn trung vị các giá trị ban đầu sẽ được thay thế bằng 9. Còn nếu làm mịn biên giá trị 8 ở gần giá trị 4 hơn nên nó được thay thế bằng 4.

Hồi quy: Phương pháp thường dùng là hồi quy tuyến tính, để tìm ra được một mối quan hệ tốt nhất giữa hai thuộc tính (hoặc các biến), từ đó một thuộc tính có thể dùng để dự đoán thuộc tính khác. Hồi quy tuyến tính đa điểm là một sự mở rộng của phương pháp trên, trong đó có nhiều hơn hai thuộc tính được xem xét, và các dữ liệu tính ra thuộc về một miền đa chiều. Nội dung cụ thể được trình bày trong 2.1.3.

Nhóm/cụm: Các giá trị tương tự nhau được tổ chức thành các nhóm hay "cụm" trực quan. Các giá trị rơi ra bên ngoài các nhóm này là những giá trị nhiễu sẽ được xem xét để làm mịn.

2.1.3 Phân tích tương quan và hồi quy phục vụ khử nhiễu và điền dữ liệu thiếu.

2.1.3.1 Phân tích tương quan

Trong lý thuyết xác suất và thống kê, hệ số tương quan (Coefficient Correlation) cho biết độ mạnh của mối quan hệ tuyến tính giữa hai biến số ngẫu nhiên. Từ tương quan (Correlation) được thành lập từ Co- (có nghĩa "together") và Relation (quan hệ).

Một trong những mục tiêu của phân tích và xử lý dữ liệu môi trường là tìm hiểu những mối tương quan giữa các yếu tố khí tượng cũng như các chỉ tiêu quan trắc với nhau, và qua đó có thể tiên lượng một yếu tố phụ thuộc từ các yếu tố độc lập. "Mối tương quan" ở đây bao gồm các đặc điểm như mức độ tương quan và xây dựng một mô hình tiên đoán. Mô hình ở đây chính là hàm số nối kết hai biến với nhau, và hàm số này phải có độ tin cậy nhất định và có ý nghĩa để giải thích được dữ liệu.

Gọi x_i và y_i là hai biến quan sát giá trị x và y của đối tượng *i*. Giả sử chúng ta có *n* đối tượng thì $i = 1, 2, 3, ..., n$. Gọi $\bar{x}$ và $\bar{y}$ là hai giá trị trung bình của biến quan sát được x và y; s_x^2 và s_y^2 lần lượt là phương sai của hai biến, được định nghĩa như sau:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (9)$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (10)$$

Do đó, nếu x và y độc lập, chúng ta có thể viết:

$$s_{x+y}^2 = s_x^2 + s_y^2 \quad (11)$$

Nhưng nếu x và y có liên hệ với nhau, công thức trên không đáp ứng được vấn đề mô tả. Chúng ta cần tìm một chỉ số khác mô tả mối liên hệ giữa hai biến, bằng cách nhân độ lệch của biến x từ số trung bình, $(x_i - \bar{x})$, cho độ lệch của biến y, $(y_i - \bar{y})$, thay vì bình phương độ lệch từng biến riêng lẻ như công thức (11) . Nói cách khác, tích số hai độ lệch chính là hiệp biến. Đối với mỗi cá nhân, hiệp biến kí hiệu là "Cov", viết tắt của Covariance.

$$cov(x_i, y_i) = (x_i - \bar{x})(y_i - \bar{y}) \quad (12)$$

Nhưng ở đây chúng ta có n đối tượng, cho nên cần phải cộng tất cả lại và chia cho số đối tượng:

$$cov(x_i, y_i) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad (13)$$

Một cách để "chuẩn hóa" hiệp biến và phương sai là lấy tỉ số của hai chỉ số này như công thức (14). Và đây chính là định nghĩa của hệ số tương quan. Hệ số tương quan thường được kí hiệu bằng r:

$$r = \frac{cov(x,y)}{\sqrt{var(x).var(y)}} = \frac{cov(x,y)}{s_x * s_y} \quad (14)$$

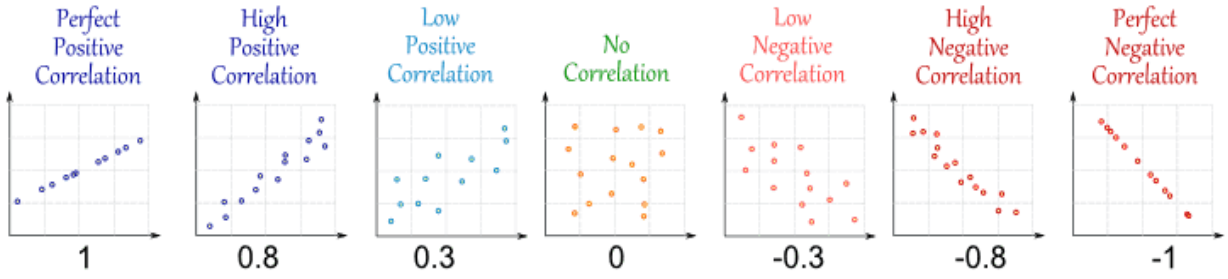
(Chú ý rằng căn số bậc hai của phương sai là độ lệch chuẩn cho nên công thức trên được mô tả bằng độ lệch chuẩn, thay vì phương sai). Với vài thao tác đại số, có thể viết lại công thức sẽ được chuyển đổi như sau:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{(n-1)s_x s_y} \quad (15)$$

Công thức trên còn được biết đến như là hệ số Pearson (Pearson's correlation coefficient) để ghi nhận cống hiến của nhà thống kê học nổi tiếng Karl Pearson, người đầu tiên phát triển lí thuyết về tương quan.

Hệ số tương quan giữa 2 biến có thể mang giá trị dương hoặc âm. Hệ số tương quan dương cho biết rằng giá trị 2 biến tăng cùng nhau còn hệ số tương quan âm thì nếu một biến tăng thì biến kia sẽ giảm.

Độ mạnh và hướng tương quan của 2 biến được mô tả như sau:



Hình 6. Minh họa ý nghĩa giá trị của hệ số tương quan.

Hệ số tương quan có thể nhận giá trị từ -1 đến 1 và có những ý nghĩa khác nhau:

Bảng 9. Bảng ý nghĩa ứng với các khoảng giá trị hệ số tương quan.

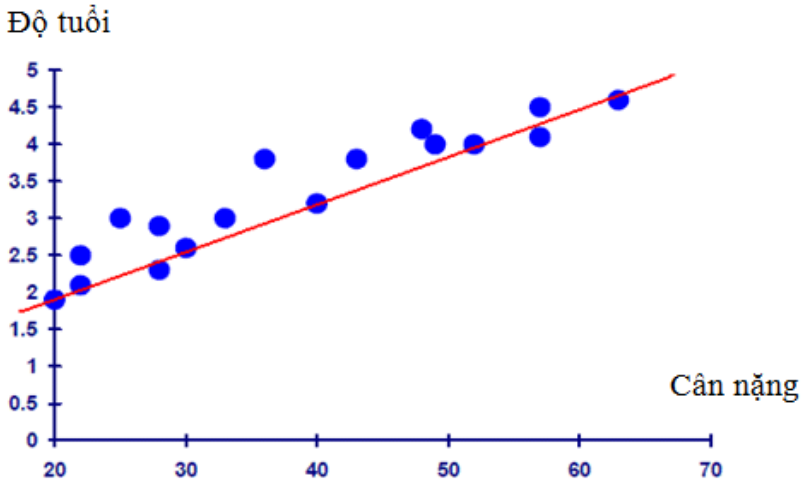
Hệ số tương quan	Ý nghĩa
±0.01 đến ±0.1	Mối tương quan quá thấp, không đáng kể
±0.2 đến ±0.3	Mối tương quan thấp
±0.4 đến ±0.5	Mối tương quan trung bình
±0.6 đến ±0.7	Mối tương quan cao
±0.8 trở lên	Mối tương quan rất cao

2.1.1.1 Hồi quy tuyến tính

a) Hồi quy đơn biến.

Phân tích hồi qui tuyến tích đơn giản hay đơn biến (Simple Linear Regression Analysis) là tìm sự liên hệ giữa 2 biến số liên tục là biến độc lập (biến dự đoán) trên trục hoành x với biến phụ thuộc (biến kết cục) trên trục tung y. Sau đó vẽ một đường thẳng hồi qui và từ phương trình đường thẳng này ta có thể dự đoán được biến y.

Ví dụ: Biểu đồ của cân nặng (trục y) so với độ tuổi (trục x) cho ra mối quan hệ như Hình 7:



Hình 7. Biểu đồ minh họa đường hồi quy tuyến tính

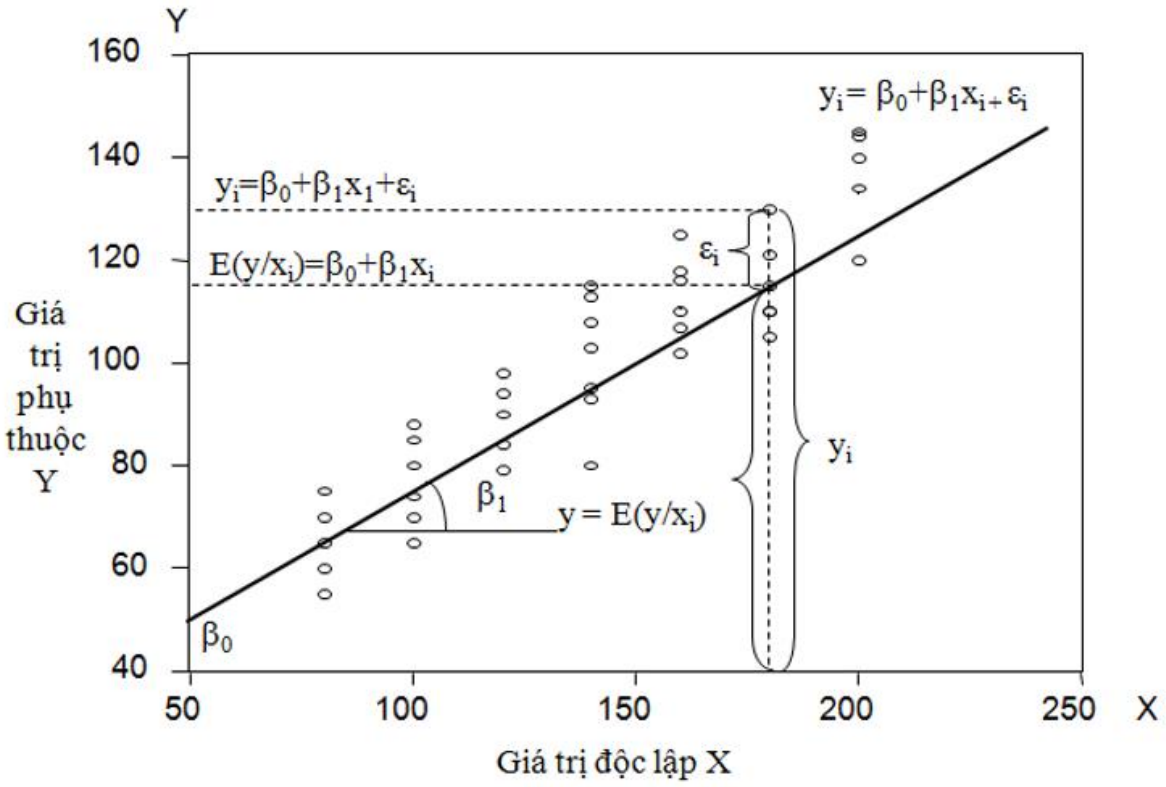
Từ đồ thị này, chúng ta có thể thấy rằng mức cân nặng dường như thay đổi một cách hệ thống với tuổi, tuổi tăng thì cân nặng cũng tăng theo. Hơn thế nữa, các điểm dữ liệu dường như nằm rãi rác xung quanh đường thẳng nối liền hai điểm (20; 2) và

(65; 4,5). Như chúng ta đã biết rằng qua hai điểm bất kì có thể dựng được duy nhất một đường thẳng. Cùng một nguyên tắc được áp dụng ở đây nhưng các kĩ thuật ước tính thì hơi phức tạp hơn.

Gọi các cặp giá trị quan sát của x và y là (x_1, y_1), (x_2, y_2),. . . , (x_n, y_n). Bản chất của một phân tích hồi quy có liên quan với các quan hệ giữa biến phụ thuộc (y) và biến độc lập (x). Quan hệ đơn giản nhất là mô hình đường thẳng:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (16)$$

Trong mô hình này, β_0 và β_1 là tham số chưa biết và phải được ước tính từ dữ liệu quan sát, ε_i là sai số ngẫu nhiên tuân theo luật phân phối chuẩn với giá trị trung bình là 0. Trong phương trình trên β_0 là chặn (intercept) và β_1 là độ dốc (slope hay gradient). Trong thực tế β_0 và β_1 được gọi là hệ số hồi quy (regression coeficent).



Hình 8. Biểu đồ mô tả tổng quan về phép hồi quy tuyến tính.

Để ước lượng β_0 và β_1 từ một loạt các điểm dữ liệu (x_1, y_1), (x_2, y_2),. . . , (x_n, y_n) chúng ta sử dụng phương pháp bình phương nhỏ nhất.

Ý tưởng chính của phương pháp bình phương nhỏ nhất là khi nhận thấy các điểm dữ liệu trên đồ thị phân tán gần như thẳng hàng thì dò tìm đường thẳng mà tất cả các điểm dữ liệu "gần" với đường thẳng đó nhất. Đường thẳng này gọi là đường thẳng hồi quy. Về mặt toán học, việc tìm đường thẳng hồi quy thu về việc tìm tung độ gốc b_0 còn gọi là hệ số chặn) và độ dốc b_1 của nó. Trong thực hành, phương pháp này ước lượng β_0 và β_1 bằng hai hệ số b_0 và b_1 của đường thẳng $y = b_0 + b_1x$ sao cho hai hệ số này làm cho tổng các bình phương độ lệch giữa tung độ y_i của các điểm dữ liệu với tung độ $\widehat{y_i} = b_0 + b_1x_i$ của các điểm cùng hoành độ trên đường thẳng có giá trị nhỏ nhất. Nói cách khác, chúng ta phải tìm cặp số (b_0, b_1) sao cho $Q = \sum_{i=1}^{n}[y_i - (b_0 + b_1x_i)]^2$ có giá trị nhỏ nhất.

Theo toán học, Q nhỏ nhất khi các đạo hàm riêng của Q theo b_0 và theo b_1 đồng thời bằng 0, tức lá chúng ta có hệ phương trình bậc nhất hai ẩn sau:

$$\begin{cases} \sum_i y_i = nb_0 + b_1 \sum_i x_i \\ \sum_i x_i y_i = b_0 \sum_i x_i + b_1 \sum_i {x_i}^2 \end{cases} \quad (17)$$

Giải hệ phương trình (17) trên ta được kết quả sau:

$$b_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{cov(x,y)}{var(x)} \quad (18)$$

và

$$b_0 = \bar{y} - b_1\bar{x} \quad (19)$$

Phương trình (19) còn cho thấy đường thẳng hồi quy $y = b_0 + b_1x$ tìm được, đi qua điểm $(\bar{x}, \bar{y})$.

Giả sử ta có phương trình hồi quy giữa thu nhập (x) và chi tiêu (y) là y = 1923 + 0,3815x. Phương trình này hàm ý rằng nếu thu nhập của hộ gia đình tăng lên 1.000đ thì trung bình chi tiêu tăng thêm là 381,5 đ. Còn 1923đ là phần chi tiêu do các nguồn khác. Từ đó có thể dự đoán giá trị của y thông qua các giá trị của x.

b) Hồi quy đa biến

Như đã thảo luận trong mô hình hồi quy đơn giản với một biến độc lập, mô hình này khá thường xuyên được sử dụng trong các bài toán phân tích dữ liệu, nhưng với một số bài toán khác nhau chúng ta không chỉ sử dụng 1 biến độc lập mà có thể sử dụng nhiều hơn là 2,3,…k biến độc lập tùy từng yêu cầu cũng như mục đích phân tích dữ liệu. Trong phần này ta sẽ mở rộng ý tưởng để bao gồm nhiều hơn một biến độc lập trong phương trình hồi quy. Kĩ thuật này được gọi là hồi quy tuyến tính đa biến .

Một cách tổng quát, phương trình hồi qui tuyến tính đa biến có dạng:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i \quad (20)$$

Cũng giống như đối với phương trình hồi quy đơn giản là sử dụng phương pháp bình phương nhỏ nhất để ước lượng tham số thì đối với hồi quy đa biến các tham số $\beta_0, \beta_1, \beta_2, \ldots \beta_k$ có thể được ước lượng dễ dàng nhờ các phần mềm hỗ trợ hoặc giải theo phương pháp ma trận. Hệ phương trình sinh ra sẽ càng phức tạp nếu số lượng biến độc lập trong phương trình hồi quy lớn.

Để dễ hiểu hơn ta so sánh với mô hinh hồi quy đơn giản:

- β_0 vẫn là hệ số chặn
- $\beta_1, \beta_2, \ldots \beta_k$ là các hệ số hồi quy xác định độ dốc đường hồi quy
- ε_i vẫn là sai số ngẫu nhiên tuân theo luật phân phối chuẩn với giá trị trung bình là 0
- Vẫn sử dụng phương pháp bình phương nhỏ nhất, nhưng do có k biến độc lập nên sẽ có tương ứng k+1 phương trình xác định hệ hối hồi quy.

c) Hệ số xác định R^2

Hệ số xá định (Multiple coefficient of determination) R^2 được định nghĩa như là tỉ lệ (hay phần trăm) biến động của biến phụ thuộc (y) được giải thích bởi các biến độc lập (x_i). Giá trị R^2 càng cao là một dấu hiệu cho thấy mối liên hệ giữa biến độc lập và biến phụ thuộc càng chặt chẽ.

Giả sử ta có phương trình hồi quy giữa thu nhập (x) và chi tiêu (y). Ta có phương trình hồi quy y = 1923 + 0,3815x và hệ số xác định R^2 có giá trị là 0.88. Điều này có nghĩa là mô hình hồi quy sẽ giải thích khoảng 88% các khác biệt về chi tiêu giữa các cá nhân.

- Hệ số xác định được tính như sau:

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} \qquad 0 \le R^2 \le 1 \quad (21)$$

- TSS (Total sum of squares): Tổng bình phương toàn phần

$$TSS = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n(\bar{y})^2 \quad (22)$$

- ESS (Explained sum of squares): Tổng bình phương hồi quy

$$ESS = \sum (\hat{y}_i - \bar{y})^2 \quad (23)$$

- RSS(Residual sum of square): Tổng bình phương phần dư

$$RSS = \sum (y_i - \hat{y}_i)^2 = \sum \varepsilon_i^2 \quad (24)$$

- R^2 có giá trị từ 0 đến 1 ($0 \leq R^2 \leq 1$)
  - R^2 càng gần 1: Mô hình phù hợp hoàn toàn với mẫu nghiên cứu.
  - R^2 càng gần 0: Mô hình hoàn toàn không phù hợp với mẫu nghiên cứu.

2.2 Chuẩn hóa dữ liệu quan trắc môi trường.

Đối với dữ liệu, việc hiểu được đặc tính dữ liệu giúp cho quá trình phân tích, đánh giá và gia quyết định hiệu quả hơn. Trong thực tế với mọi loại dữ liệu nói chung cũng như những dữ liệu quan trắc môi trường nói riêng đều không thể tránh khỏi bị nhiễu, không đầy đủ, không nhất quán là những vấn đề thường xuyên xảy ra.

Từ những nguyên nhân trên cũng như thực tế gặp phải trong công việc xử lý dữ liệu quan trắc môi trường tôi thấy rằng cần có một bộ dữ liệu được "chuẩn hóa". "Chuẩn hóa" ở đây bao gồm:

- Chuẩn về cấu trúc dữ liệu: Dữ liệu được qui ước tập hợp về đúng định dạng về thời gian đo, đơn vị đo, về tên trường, thứ tự sắp xếp… theo qui ước cụ thể đảm bảo vấn đề đồng nhất dữ liệu.
- Chuẩn về chất lượng dữ liệu: Dữ liệu được làm sạch (loại bỏ nhiễu và dữ liệu thiếu được bổ sung)

Kết quả cuối cùng đó là một bộ dữ liệu theo chuẩn qui ước, hỗ trợ cho các công tác nghiên cứu, đánh giá, gia quyết định … một cách hiệu quả.

2.3 Phương pháp đề xuất.

Theo các nghiên cứu hiện tại ở Việt Nam chứa có một qui trình chính thống nào được sử dụng với những dữ liệu quan trắc môi trường. Dựa trên những đặc điểm dữ liệu, hiện trạng dữ liệu, công tác xử lý dữ liệu quan trắc môi trường hiện tại ở Việt Nam cũng như thông qua việc tổng hợp các nghiên cứu, tập hợp tài liệu tôi đề xuất quy trình chuẩn hóa dữ liệu quan trắc môi trường theo 5 bước. Từ cơ sở đó xây dựng công cụ tự động hóa hỗ trợ công tác xử lý dữ liệu quan trắc môi trường ở Việt Nam.

Phương pháp chuẩn hóa dữ liệu quan trắc môi trường được đề xuất bao gồm 5 bước như sau (Hình 9):

1. Thu thập dữ liệu.
2. Đánh giá dữ liệu tổng quan (dựa trên thống kê).
3. Xử lý dữ liệu nhiễu.
4. Xử lý dữ liệu thiếu.
5. Đánh giá dữ liệu sau mỗi bước.

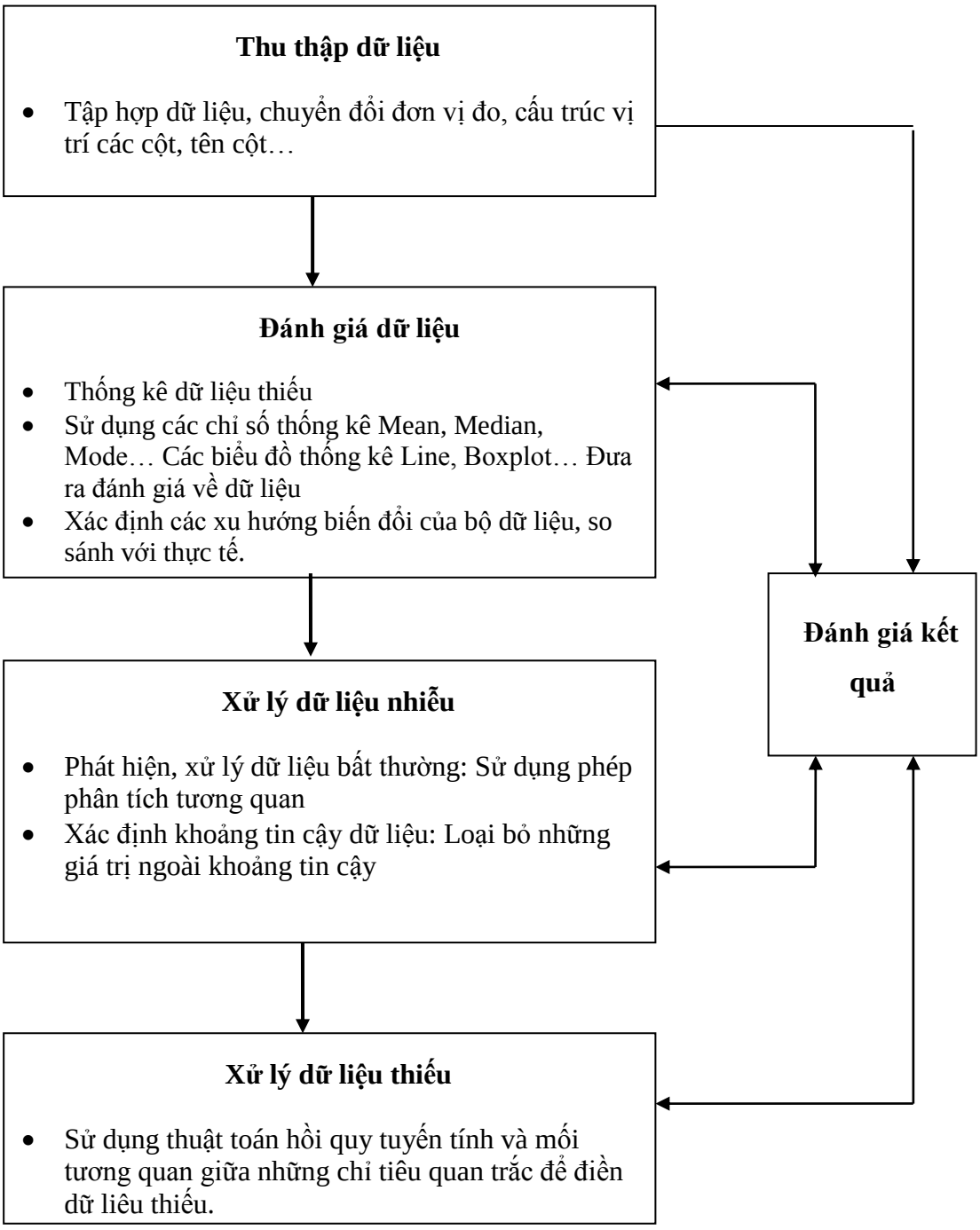
Các bước con trong quy trình được mô tả chi tiết như Bảng 10. Các qui trình con có thể chạy đôc lập và có thể xoay vòng thông qua công tác đánh giá dữ liệu sau mỗi bước xử lý. Với những kết quả đanh giá cụ thể nhà phân tích sẽ đưa ra những hướng xử lý khác nhau để sinh ra bộ dữ liệu cuối cùng đạt kết quả tốt nhất.

Phương pháp đề xuất trên sẽ được mô tả, trình bày chi tiết với dữ liệu quan trắc thực tế tại Chương 3.

Bảng 10. Bảng thông tin chi tiết từng quy trình con trong quy trình chuẩn hóa dữ liệu quan trắc môi trường được đề xuất.

TT	Tên quy trình	Nội dung	Người thực hiện	Điều kiện áp dụng	Công cụ xử lý
1	Thu thập dữ liệu	Thu thập dữ liệu quan trắc ô nhiễm không khí và thông số khí tượng. Sau đó tập hợp, xây dựng bộ dữ liệu chung theo qui ước đã định. Mục đích là xây dựng một bộ dữ liệu chuẩn về cấu trúc sao cho đơn giản hóa quá trình quản lý và phân tích dữ liệu.	- Nhân viên quản lý trạm quan trắc - Nhà phân tích dữ liệu	Áp dụng hàng ngày/tuần định kỳ khi có dữ liệu mới	- Excel - EnvPro
2	Đánh giá dữ liệu tổng quan	Sử dụng các phương pháp thống kê dữ liệu đưa ra những đánh giá, đặc trưng nhất của dữ liệu. Với những phân tích sơ bộ ta có thể thấy và so sánh được nhưng xu hướng (Trend) của dữ liệu và so sánh với thực tế từ đó kết luận độ tin cậy của dữ liệu	Nhà phân tích dữ liệu	Áp dụng hàng ngày/tuần/tháng/năm định kỳ.	EnvPro
3	Xử lý dữ liệu nhiễu	- Loại bỏ dữ liệu theo khoảng dữ liệu tin cậy. - Xác định và đánh giá dữ liệu	Nhà phân tích dữ liệu	Áp dụng hàng tháng	EnvPro

		bất thường: Sử dụng phương pháp phân tích tương quan phát hiện những ngày có dữ liệu quan trắc bất thường. Từ đó đưa ra những phương pháp để xử lý.			
4	Xử lý dữ liệu thiếu	Sử dụng thuật toán hồi quy tuyến tính và phân tích tương quan giữa những chỉ tiêu quan trắc... để điền giá trị quan trắc cho những bản ghi thiếu dữ liệu.	Nhà phân tích dữ liệu	Áp dụng hàng tháng	EnvPro
5	Đánh giá dữ liệu sau mỗi bước	Đánh giá kết quả sau mỗi qui trình con. Tùy vào kết quả đạt được, nhà phân tích tiền hành chạy lại quy trình đang thực hiện hoặc thực hiến quy trình kế tiếp.	Nhà phân tích dữ liệu	Sau mỗi qui trình từ 2-4	Kinh nghiệm phân tích dữ liệu quan trắc môi trường



Hình 9. Sơ đồ tổng quan quy trình đề xuất chuẩn hóa dữ liệu quan trắc môi trường tại Việt Nam

CHƯƠNG 3. ĐÁNH GIÁ QUY TRÌNH CHUẨN HÓA DỮ LIỆU QUAN TRẮC PM10 TẠI TRẠM NGUYỄN VĂN CỪ, HÀ NỘI.

3.1 Tổng quan khu vực nghiên cứu.

3.1.1 Vị trí địa lý

Hà Nội là thủ đô của Việt Nam có tọa độ từ 20°53' đến 21°23' vĩ độ Bắc và 105°44' đến 106°02' kinh độ Đông, tiếp giáp với các tỉnh Thái Nguyên - Vĩnh Phúc ở phía Bắc; Hà Nam - Hòa Bình ở phía Nam; Bắc Giang- Bắc Ninh- Hưng Yên ở phía Đông và Hòa Bình- Phú Thọ ở phía Tây. Thủ đô Hà Nội có diện tích tự nhiên 334.470,02 ha, dân số hơn 7 triệu người, gồm 30 đơn vị hành chính cấp quận, huyện, thị xã, 577 xã, phường, thị trấn.

Hà Nội hiện nay vừa có núi, có đồi và địa hình thấp dần từ Bắc xuống Nam, từ Tây sang Đông, trong đó đồng bằng chiếm tới ¾ diện tích tự nhiên của thành phố. Độ cao trung bình của Hà Nội từ 5 đến 20 mét so với mặt nước biển, các đồi núi cao đều tập trung ở phía Bắc và Tây. [43]

3.1.2 Khí hậu, khí tượng

Nằm trong vùng nhiệt đới gió mùa, khí hậu Hà Nội có đặc trưng nổi bật là gió mùa ẩm, nóng và mưa nhiều về mùa hè, lạnh và ít mưa về mùa đông; được chia thành bốn mùa rõ rệt trong năm: Xuân, Hạ, Thu, Đông. Mùa xuân bắt đầu vào tháng 2 (hay tháng giêng âm lịch) kéo dài đến tháng 4. Mùa hạ bắt đầu từ tháng 5 đến tháng 8, nóng bức nhưng lại mưa nhiều. Mùa thu bắt đầu từ tháng 8 đến tháng 10, trời dịu mát. Mùa đông bắt đầu từ tháng 11 đến tháng 1 năm sau, thời tiết giá lạnh, khô hanh. Ranh giới phân chia bốn mùa chỉ có tính chất tương đối, vì Hà Nội có năm rét sớm, có năm rét muộn, có năm nóng kéo dài, nhiệt độ lên tới 40°C, có năm nhiệt độ xuống thấp dưới 8°C. Nhiệt độ trung bình năm 24,9°C, độ ẩm trung bình 80 - 82%. Lượng mưa trung bình trên 1700mm/năm (khoảng 114 ngày mưa/năm). [43]

3.1.3 Phạm vi dữ liệu nghiên cứu.

Như đã đề cập trong 1.3.2, hiện tại ở Hà Nội có 3 trạm quan trắc không khí, khí tượng bao gồm 1 trạm thuộc mạng lưới quan trắc khí tượng thủy văn và môi trường quốc gia (đặt tại Pháo Đài Láng, Đống Đa), 2 trạm thuộc Mạng lưới quan trắc môi trường quốc gia, Tổng cục Môi trường (đặt tại 556 Nguyễn Văn Cừ và Lăng Chủ Tịch Hồ Chí Minh) cụ thể:

Bảng 11. Bảng thông tin các trạm quan trắc hiện có trên địa bàn Hà Nội.

Mạng lưới	Vị trí hoạt động	Năm hoạt động	Chỉ tiêu quan trắc
Mạng lưới quan trắc khí tượng thủy văn và môi trường quốc gia	Pháo Đài Láng, Đống Đa	2002	SO2, NO, NO2, NH3, CO, O3, NMHC, CH4, TSP, PM10, OBC, WD, WS, Temp, Hum, SR, UV, ATP, Rain
Mạng lưới quan trắc môi trường quốc gia, Tổng cục Môi trường	556 Nguyễn Văn Cừ	2009	Wind.Spd, Wind.Dir, Temp, RH, Barometer, Radiation, InnerTemp, NO, NO2, NOx, SO2, CO, O3, PM.10, PM.2.5, PM.1, CH4, NMHC, THC, BENZEN, TOLUEN, ETHYL.BENZEN, MP.XYLEN, O.XYLEN
	Lăng Chủ Tịch Hồ Chí Minh	2012	

Với 3 trạm quan trắc khí tượng và không khí tại Hà Nội hiện tại thì đối với:

- Trạm quan trắc đặt tại Láng: Tính đến nay cũng đã hoạt động được 16 năm thêm vào đó là hoạt động bảo dưỡng, bảo trì còn kém cũng như chi phí duy trì lớn dẫn tới nhiều module quan trắc đã bị hỏng. Vì vậy dữ liệu quan trắc có nhiều sai sót, đặc biệt là dữ liệu thiếu rất nhiều.
- Trạm quan trắc Nguyễn Văn Cừ: Thời gian đưa vào hoạt động cũng khá gần đây, các hoạt động bảo trì bảo dưỡng cũng được thực hiện thường xuyên.

Trạm quan trắc này khá quan trọng vì là trạm đầu tiên được lắp đặt tại Hà Nội bởi Trung tâm quan trắc môi trường quốc gia và được lắp đặt, vận hành ngay tại khuôn viên của Trung tâm quan trắc môi trường quốc gia. Chính vì vậy khả năng vận hành và duy trì có thể nói là được đảm bảo.

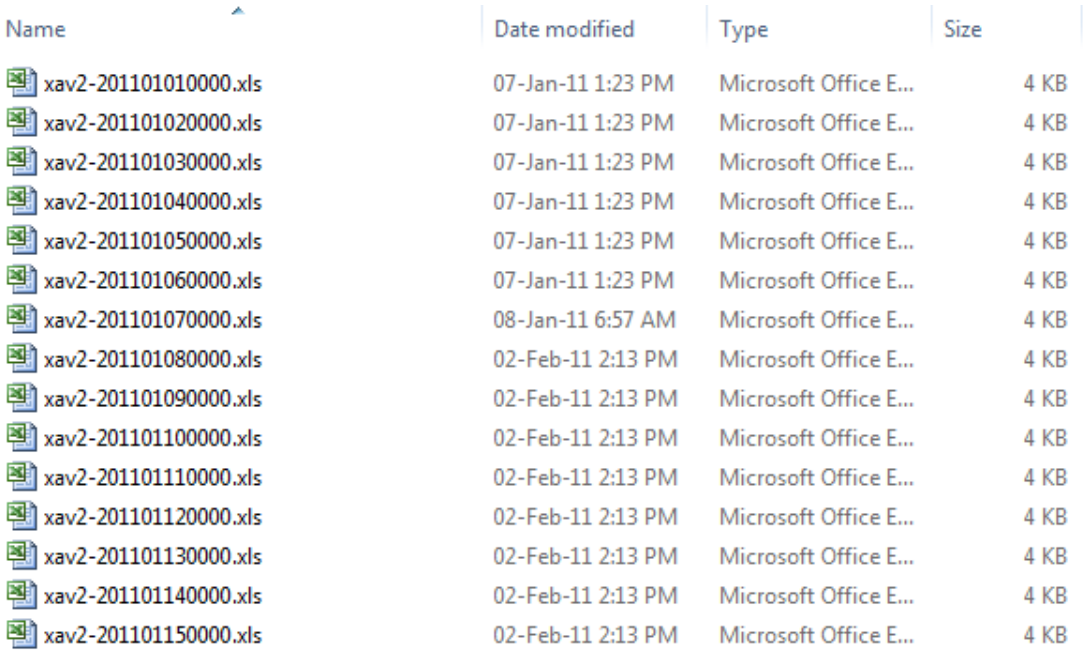
- Trạm quan trắc Lăng Chủ tịch Hồ Chí Minh: Trạm quan trắc đặt tại khu vực Lăng khá mới, được lắp đặt năm 2012. Trong địa phận nội đô Hà Nội thì đây là khu vực quan trọng có thể nói là đầu não về chính trị của Việt Nam với nhiều khu vực quan trọng, nhạy cảm xung quanh. Vì vậy khả năng tiếp cận về thông tin quan trắc cũng như dữ liệu quan trắc khá khó khăn.

Trong giới hạn phạm vi luận văn cũng như mục đích nghiên cứu và khả năng tiếp cận dữ liệu hiện tại của 3 trạm quan trắc. Tôi sẽ tập trung, áp dụng phương pháp đã đề xuất đối với chỉ tiêu quan trắc bụi PM10 với bộ dữ liệu từ 01/01/2011 đến 31/01/2011 và 01/01/2012 đến 31/01/2012 tại trạm Nguyễn Văn Cừ - Hà Nội.

3.2 Phương pháp chuẩn hóa dữ liệu quan trắc môi trường.

3.2.1 Thu thập dữ liệu

Bộ dữ liệu quan trắc tháng 01/2011 và 01/2012 quan trắc tại trạm Nguyễn Văn Cừ được cung cấp bởi Trung tâm quan trắc môi trường quốc gia. Dữ liệu quan trắc tại trạm Nguyễn Văn Cừ được đo theo giờ. Mỗi giờ mỗi chỉ tiêu quan trắc sẽ có một giá trị quan trắc và mỗi ngày quan trắc là một file dữ liêu riêng biệt. Tên file được đặt tự động theo ngày quan trắc.



Name	Date modified	Type	Size
xav2-201101010000.xls	07-Jan-11 1:23 PM	Microsoft Office E...	4 KB
xav2-201101020000.xls	07-Jan-11 1:23 PM	Microsoft Office E...	4 KB
xav2-201101030000.xls	07-Jan-11 1:23 PM	Microsoft Office E...	4 KB
xav2-201101040000.xls	07-Jan-11 1:23 PM	Microsoft Office E...	4 KB
xav2-201101050000.xls	07-Jan-11 1:23 PM	Microsoft Office E...	4 KB
xav2-201101060000.xls	07-Jan-11 1:23 PM	Microsoft Office E...	4 KB
xav2-201101070000.xls	08-Jan-11 6:57 AM	Microsoft Office E...	4 KB
xav2-201101080000.xls	02-Feb-11 2:13 PM	Microsoft Office E...	4 KB
xav2-201101090000.xls	02-Feb-11 2:13 PM	Microsoft Office E...	4 KB
xav2-201101100000.xls	02-Feb-11 2:13 PM	Microsoft Office E...	4 KB
xav2-201101110000.xls	02-Feb-11 2:13 PM	Microsoft Office E...	4 KB
xav2-201101120000.xls	02-Feb-11 2:13 PM	Microsoft Office E...	4 KB
xav2-201101130000.xls	02-Feb-11 2:13 PM	Microsoft Office E...	4 KB
xav2-201101140000.xls	02-Feb-11 2:13 PM	Microsoft Office E...	4 KB
xav2-201101150000.xls	02-Feb-11 2:13 PM	Microsoft Office E...	4 KB

Hình 10. Các file lưu trữ dữ liệu quan trắc theo ngày.

Có một vấn đề đối với dữ liệu quan trắc là đơn vị đo. Đơn vị đo nhiều khi khác nhau trong cùng một trạm, hoặc khác nhau giữa các trạm. Ví dụ với dữ liệu quan trắc CH_4 trong 6 tháng đầu năm đo với đơn vị là ug/m^3, 6 tháng cuối năm lại đo theo đơn vị là ppm hoặc ppb. Vì vậy nếu tập hợp dữ liệu mà chưa chuyển đổi đơn vị đo sẽ tạo ra những sai sót về sữ liệu thực tế. Một thủ thuật excel đơn giản để giải quyết vấn đề này là áp dụng các công thức chuyển đổi tương ứng vào nhưng chỉ tiêu quan trắc không đồng nhất về đơn vị đo trước khi tập hợp thành một bộ dữ liệu hoàn chỉnh

Tất cả những dữ liệu quan trắc có được sau khi được tập hợp và chuyển đổi đơn vị đo tương ứng tại trạm Nguyễn Văn Cừ sau khi được tập hợp được thể hiện có dạng như sau:

A	B	C	D	E	F	G	H	I	J	K
Time	MeasPt	WindSpd	WindDir	Temp	RH	Baromete	Radiation	InnerTem	NO	NO2
01-01-12 0:00		0.82168	14.2896	15.0002	78.006	1021.92	1.0822	23.1935	19.6029	25.1934
01-01-12 1:00		0.72819	23.439	14.9586	78.951	1021.72	1.06215	23.0366	10.6418	21.8429
01-01-12 2:00		0.51529	28.4613	14.9052	79.14	1021.45	1.07431	23.0039	12.2355	24.929
01-01-12 3:00		0.67681	17.7943	14.9431	78.264	1021.25	1.05911	23.0552	8.4723	22.3998
01-01-12 4:00		0.71853	359.914	15.0319	77.567	1021.38	1.05668	22.9637	13.778	28.3332
01-01-12 5:00		0.451318	336.815	14.9737	77.457	1021.72	1.06884	23.1267	16.0281	32.1984
01-01-12 6:00		0.83612	316.111	14.9799	78.065	1022.32	1.07248	23.1128	21.2198	37.7753
01-01-12 7:00		0.65278	326.097	15.2315	78.083	1023	9.7957	22.9322	33.6441	43.7088
01-01-12 8:00		0.74671	324.4	15.5615	78.759	1023.65	36.6728	23.0064	31.7742	43.0478
01-01-12 9:00		0.346798	0.67724	15.9529	72.742	1023.66	48.6081	23.3782	25.3461	46.1956
01-01-12 10:00		0.55487	350.16	16.0206	71.523	1022.76	44.2951	23.4997	25.8319	48.6088
01-01-12 11:00		0.54441	7.6585	16.4554	68.31	1021.23	116.529	23.8215	19.3423	43.4601
01-01-12 12:00		1.04559	312.455	16.8138	67.066	1020.05	110.583	24.5243	10.5549	36.0215
01-01-12 13:00		0.98056	316.36	16.9592	67.095	1019.13	73.045	24.6973	16.6469	47.2323
01-01-12 14:00		0.54413	4.94018	17.0217	66.758	1018.58	60.762	24.7907	22.0938	53.122
01-01-12 15:00		0.41096	351.388	16.9693	67.968	1018.46	26.0015	24.9341	30.5399	58.451
01-01-12 16:00		0.372736	10.5371	16.7102	69.621	1018.71	1.83142	24.7489	46.1144	63.866
01-01-12 17:00		0.50241	343.213	16.6806	71.16	1018.88	1.002	24.7103	87.712	73.865
01-01-12 18:00		0.56768	26.2612	16.4181	72.753	1019.34	1.02509	24.6203	44.6663	53.559
01-01-12 19:00		0.361471	5.16	16.279	74.669	1019.68	0.98255	24.2493	40.3902	50.048
01-01-12 20:00		0.66209	343.147	16.2258	76.041	1019.64	0.98194	24.0652	43.437	50.144

Hình 11. Dữ liệu sau khi được tập hợp từ các file lưu trữ theo ngày

Dữ liệu sau khi tập hợp được lưu trữ dưới định dạng Excel hoặc *.CSV hỗ trợ dễ dàng cho các thao tác quản lý và phân tích dữ liệu. Đối với riêng tôi, các qui trình xử lý đề xuất, sử dụng ngôn ngữ R[(4)] làm công cụ thao tác chính với dữ liệu nên tất cả dữ liệu quan trắc trong phạm vi luận văn này được tôi lưu dưới định dạng *.CSV. Giữa hai định định dạng Excel và *.CSV đều có thể dễ dàng chuyển đổi qua lại cho nhau hay chuyển đổi sang những định dạng khác nên tùy từng đầu vào bài toán cụ thể mà ta có hướng chuyển đổi phù hợp.

Dữ liệu trong quá trình tập hợp được đưa về chuẩn cấu trúc cũng như qui ước về đơn vị đo như Bảng 12:

Cấu trúc dữ liệu:

[4] Xem tại mục 4.6.2

Bảng 12. Bảng qui ước chuẩn cấu trúc, định dạng và đơn vị đo cho các chỉ tiêu quan trắc môi trường tại Việt Nam.

Tên trường	Ý nghĩa	Đơn vị đo/ Định dạng
Time	Thời gian quan trắc	MM-DD-YY H:M
WindSpd	Tốc độ gió	m/s (mét/giây)
WindDir	Hướng gió	Degree
Temp	Nhiệt độ	°C
RH	Độ ẩm tương đối (Relative humidity)	%
Barometer	Khí áp	hPa
Radiation	Bức xạ	W/m^2
InnerTemp	Nhiệt độ cục bộ	°C
NO	Giá trị quan trắc NO	$\mu g/m^3$
NO2	Giá trị quan trắc NO_2	$\mu g/m^3$
NOx	Giá trị quan trắc NOx	$\mu g/m^3$
SO2	Giá trị quan trắc SO_2	$\mu g/m^3$
CO	Giá trị quan trắc CO	$\mu g/m^3$
O3	Giá trị quan trắc O_3	$\mu g/m^3$
PM10	Giá trị quan trắc PM10	$\mu g/m^3$
PM2.5	Giá trị quan trắc PM2.5	$\mu g/m^3$
PM1	Giá trị quan trắc PM1	$\mu g/m^3$
CH4	Giá trị quan trắc CH_4	ppm
NMHC	Giá trị quan trắc NMHC (Non-Methane Hydrocarbon)	ppm

THC	Giá trị quan trắc THC (Total Hydrocarbon)	ppm
BENZEN	Giá trị quan trắc Benzen	µg/m^3
TOLUEN	Giá trị quan trắc Toluen	µg/m^3
ETHYLBENZEN	Giá trị quan trắc Ethyl Benzen	µg/m^3
MPXYLEN	Giá trị quan trắc MPXYLEN	µg/m^3
OXYLEN	Giá trị quan trắc OXYLEN	µg/m^3

3.2.2 Đánh giá dữ liệu tổng quan

Tỉ lệ dữ liệu thiếu

Nhìn chung với bộ dữ liệu từ tháng 01/2011 tỉ lệ dữ liệu thiếu là không lớn. Đối với dữ liệu quan trắc bụi PM chỉ thiếu khoảng 2%. Đối với tất cả các chỉ tiêu quan trắc khắc được quan trắc đầy đủ 100% không có dữ liệu thiếu. Nguyên nhân thiếu thì có nhiều khả năng có thể do mất điện hoặc sửa chữa, bảo trì máy quan trắc….

Ngược lại theo thống kê quan trắc bụi PM và các chỉ tiêu quan trắc khác tháng 01/2012 số giờ thiếu là không có, 100% dữ liệu là hoàn chỉnh chỉ riêng 2 tỉ tiêu SO_2 thiếu 23% và O_3 là 37.4 % số giờ quan trắc dữ liệu.

Bảng 13. Bảng thống kê tỉ lệ dữ liệu thiếu theo từng tháng (tính theo số bản ghi thiếu / tổng số bản ghi cần quan trắc)

Chỉ tiêu/ Tháng	**01/2011**	**01/2012**
WindSpd	0	0
WindDir	0	0
Temp	0	0
RH	0	0
Barometer	0	0
Radiation	0	0

InnerTemp	0	0
NO	0	0
NO_2	0	0
NOx	0	0
SO_2	0	170/744
CO	0	0
O_3	0	278/744
PM10	15/744	0
PM25	15/744	0
PM1	15/744	0

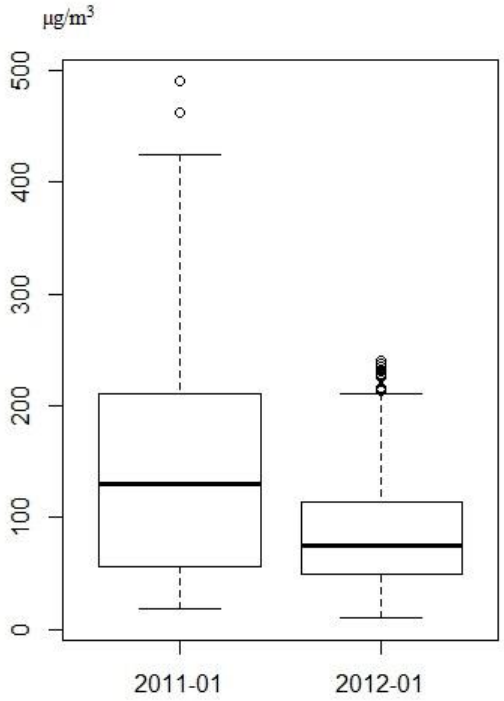
Độ tập trung và phân tán chỉ tiêu quan trắc PM10

Xét về độ tập trung và phân tán dữ liệu ta có biểu đồ boxplot PM10 theo từng tháng dựa vào biểu đồ boxplot như Hình 12.

Vạch nằm ngang giữa hộp là hàm lượng trung bình tháng (median), 2 đầu hộp là bách phân thứ 25 và 75, 2 cần vặn phía trên và dưới là giá trị lớn nhất và nhỏ nhất quan trắc được. Nhìn chung hàm lượng trung bình PM10 dao động trong khoảng từ 85-140 ug/m^3 gấn với tiêu chuẩn QCVN 05:2013/BTNMT đưa ra cho PM10 là 150 ug/m^3. Riêng tháng 01/2011 phạm vi dữ liệu khá lớn trong khoảng 10 đến gần 500 µg/m^3. Trái ngược hoàn toàn so với tháng 01/2012. Giá trị quan trắc khá cao cũng cũng là một vấn đề đáng lưu tâm bởi những giá trị quan trắc này có thể là những giá trị nhiễu.

Bảng 14. Bảng kết quả các chỉ số thống kê dữ liệu hai tháng 01/2011 và 01/2012.

Tháng/ Chỉ số	Mean	Median	Mode	Q1	Q3
01/2011	141.37	129.68	40.91	56.07	210.41
01/2012	87.18	75.39	97.22	49.61	113.61



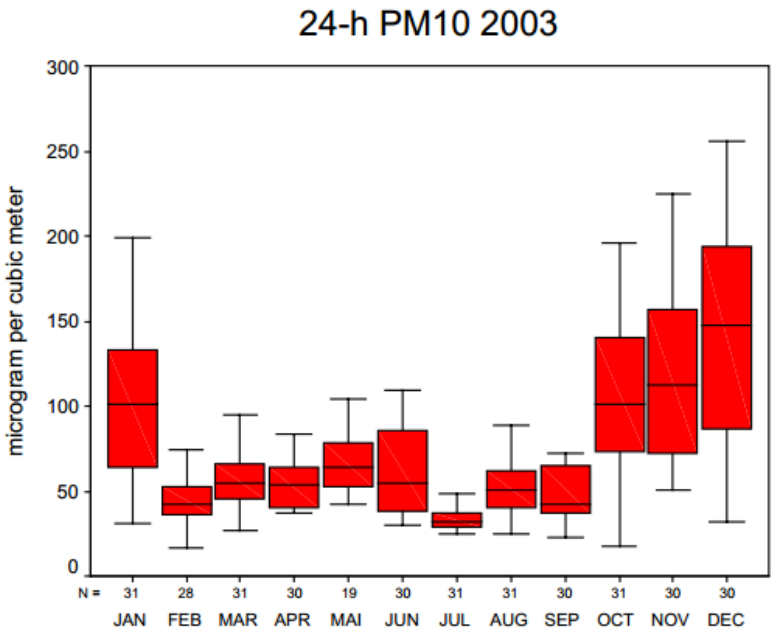
Hình 12. Biểu đồ boxplot mô tả dữ liệu hai tháng 01/2011 và 01/2012.

Hàm lượng trung bình PM10

So sánh với một nghiên cứu khác về chất lượng không khí năm 2003 tại Hà Nội [10] cho thấy hàm lượng trung bình các chỉ tiêu quan trắc thường cao về mua đông và thấp về mùa hè. Các giá trị cực đại thì thường xuyên qua sát được vào thời gian từ tháng 10 đến tháng 1 năm sau. Hàm lượng trung bình trong 4 tháng này với số liệu thực nghiệm cũng dao động trong khoảng từ 100-150 ug/m3 như Hình 13. Kết quả này cũng tương đồng với những số liệu thống kê bên trên (Bảng 14).

Nguyên nhân chính dẫn tới sự thay đổi xu hướng cao về mùa đông và thấp về mùa hè là do ảnh hưởng bởi khí hậu, khí tượng (Hình 13). Thời tiết mùa đông mang đặc trưng là khô, hanh kèm theo một số hiện tượng vật lý đặt trưng dẫn tới làm nồng độ ô nhiễm bụi tăng mạnh, từ tháng 2 trở đi gió mùa qua vịnh bắc bộ mang theo độ ẩm lớn làm giảm sự phát tán của bụi. Thêm nữa là về mùa đông do khả năng phát tán kém của khí quyển ngược với mùa hè, không khí bị đốt nóng, phát tán lên cao kèm theo mưa nhiều cũng góp phần ảnh hưởng tới xu hướng phát tán bụi theo mùa. “Quy luật

này phù hợp với kết qủa quan trắc tại nhiều nới khác ở các nước nhiệt đới Bắc bán cầu" [10].

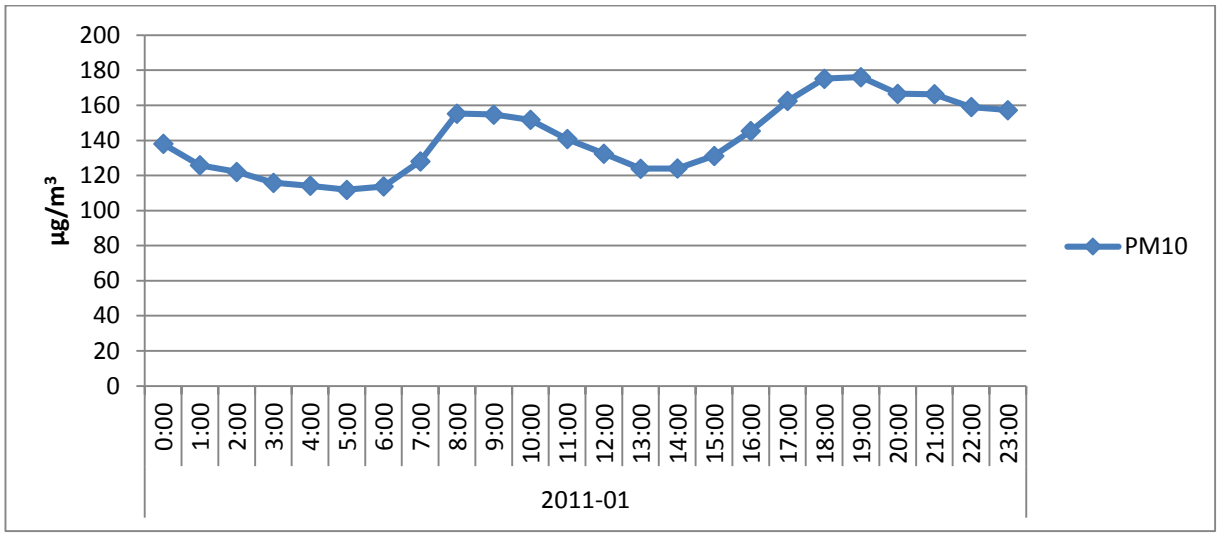


Hình 13. Diễn biến hàm lượng trung bình chỉ tiêu quan trắc PM10 năm 2003 [10]

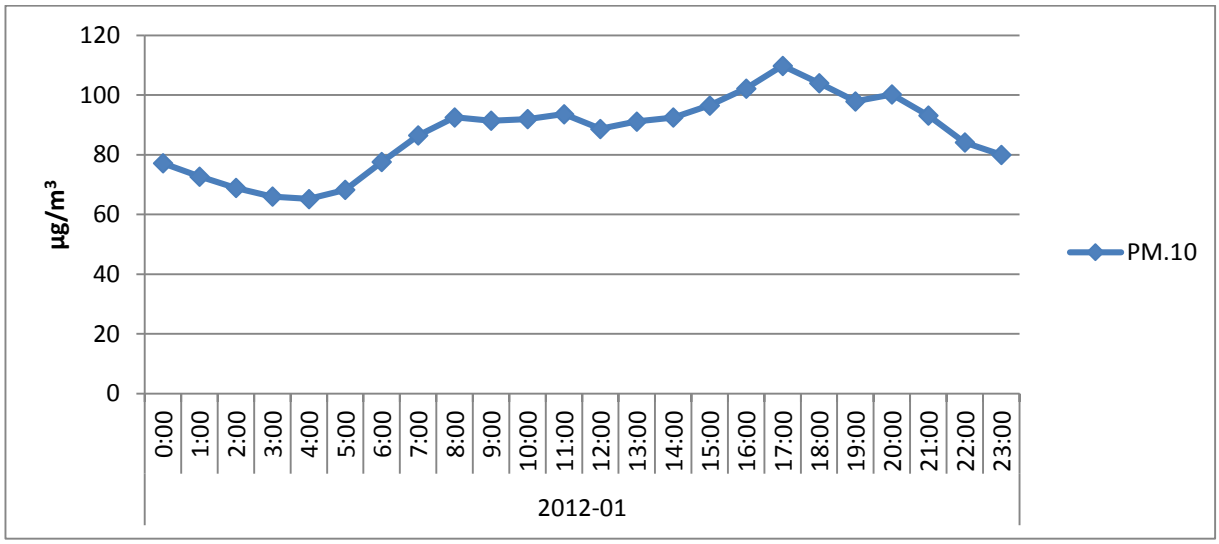
Quy luật diễn biến bụi PM10 theo ngày

Theo nghiên cứu [10] diễn biến hàm lượng ô nhiễm không khí theo ngày từ tháng 5 và tháng 9/2003 tại Hà Nội có xu hướng tăng cao vào giờ cao điểm hàng ngày từ 7-9h sáng và 18-20h tối, trong các tháng hè thì đỉnh cao nhất của buổi sáng và tối thường ngang nhau. Nhưng bắt đầu từ tháng 9 đến hết tháng 1 (mùa đông) các xu hướng này lại có những thay đổi bởi hiện tượng nghịch nhiệt do bức xạ về đêm (NRI - Nocturnal Radiation Inversion).

Áp dụng những qui luật hiện tượng đã nghiên cứu vào bộ dữ liệu. Tôi tiến hành tính trung bình giờ cho từng tháng với bộ dữ liệu ta có biểu đồ mô tả xu hướng PM10 theo 24 giờ trong 2 tháng như Hình14-15. Xu hướng theo ngày từ bộ dữ liệu tháng 01/2011 và 01/2012 tại trạm Nguyễn Văn Cừ có thể thấy vẫn tuân theo xu hướng chung là tăng vào giờ cao điểm (lượng phương tiện giao thông cao) từ 7-9h sáng và 18-20h đêm.



Hình 14. Diễn biến, xu hướng chỉ tiêu quan trắc PM10 theo 24h tháng 01/2011



Hình 15. Diễn biến, xu hướng chỉ tiêu quan trắc PM10 theo 24h tháng 01/2012

Từ những quy luật diễn biến phù hợp theo mùa và theo 24h cộng với so sánh kết quả từ những nghiên cứu trước có thể nói module quan trắc PM10 của trạm Nguyễn Văn Cừ cho giá trị tin cậy, có thể sử dụng được. Áp dụng phương pháp đánh giá dữ liệu tương tự với các chỉ tiêu quan trắc khác như NO_2, SO_2, CO… Kết quả cho thấy dữ liệu quan trắc của 2 tháng 01/2011 và 01/2011 đủ tin cậy để thực hiện các bước tiếp theo.

Với 2 bộ dữ liệu có được tôi quyết định lựa chọn bộ dữ liệu tháng 01/2011 làm bộ dữ liệu học (training) bởi lượng dữ liệu PM10 khá đầy đủ cộng thêm dữ liệu quan trắc của các chỉ tiêu khác có độ hoàn chỉnh cao, còn bộ dữ liệu 01/2012 làm bộ dữ liệu Test để thử nghiệm.

3.2.3 Xử lý dữ liệu nhiễu

Trong quan trắc dữ liệu nhiễu là những dữ liệu cho giá trị sai lệch bởi nhiều nguyên nhân như module quan trắc lỗi, hỏng, chịu ảnh hưởng của các hiện tượng tự nhiên, xã hội mang tính bất thường, đột ngột không tuân theo một quy luật nhất định…. Ví dụ như, quy luật của PM10 hàng ngày là cao vào giờ cao điểm, mà trong một ngày *x* nào đó dữ liệu quan trắc PM10 lại cho kết quả là rất cao và không tuân theo quy luật diễn biến theo ngày bình thường. Nhiệm vụ của nhà phân tích là phải phát hiện được những ngày có dữ liệu bất thường đó, tìm ra nguyên nhân gây nên hiện tượng, từ đó đưa ra phương án xử lý thích hợp. Những dữ liệu nhiễu như vậy nếu không được xử lý, đánh giá sẽ làm ảnh hưởng tới chất lượng của toàn bộ dữ liệu.

Trong phương pháp đề xuất của mình, tôi đưa ra hai hướng xử lý dữ liệu nhiễu đó là:

- Dựa vào khoảng dữ liệu tin cậy để loại bỏ những giá trị nhiễu
- Phát hiện dữ liệu bất thường bằng phép phân tích tương quan

Tùy theo mục đích của nhà phân tích mà có thể sử dụng riêng biệt từng hướng xử lý hoặc cũng có thể kết hợp đồng thời cả hai để giải quyết.

3.2.3.1 Loại bỏ giá trị nhiễu dựa vào khoảng tin cậy.

Có một phương pháp nhanh nhạy để xử lý nhiễu đó là lại bỏ nhiễu theo một khoảng giá trị xác định. Phương pháp này đòi hỏi nhà phân tích phải có khả năng và kinh nghiệm làm việc với dữ liệu quan trắc trong một thời gian dài, thường xuyên, am hiểu dữ liệu một cách chi tiết mới có thể đưa ra khoảng giá trị một cách chính xác.

Với khả năng giới hạn của mình cũng như thông qua các bài nghiên cứu, các báo cáo môi trường của các chuyên gia đi trước [2,16,20,21,22,23,11,12] tôi đề xuất khoảng giá trị tin cậy cho các chỉ tiêu quan trắc bụi như sau:

Bảng 15. Bảng kết quả xác định khoảng giá trị tin cậy đối với chỉ tiêu quan trắc bụi.

Chỉ tiêu quan trắc	Khoảng giá trị tin cậy (µg/m^3)
PM10	1 - 400
PM1	3 - 200
PM2.5	3 - 200

Áp dụng khoảng dữ liệu tin cậy từ 1 đếm 400 µg/m^3 vào bộ dữ liệu Training với chỉ tiêu quan trắc PM10. Kết quả cho thấy có 4 bản ghi có giá trị không phù hợp được loại bỏ khỏi tập dữ liệu. Cụ thể kết quả như Bảng 17

Bảng 16. Bảng thống kê danh sách bản ghi có giá trị nằm ngoài khoảng tin cậy từ bộ dữ liệu tháng 01/2011.

Ngày giờ quan trắc	Giá trị quan trắc
12/01/2011 10:00	490
17/01/2011 08:00	420.656
17/01/2011 17:00	462.044
17/11/2011 18:00	425.139

3.2.3.2 Loại bỏ giá trị nhiễu bằng phân tích tương quan.

Với cách này tôi so sánh tương quan theo 24h giữa tháng và các ngày trong tháng đó. Cụ thể, giá trị trung bình hàng giờ của từng tháng chính là những giá trị đại diện cho xu hướng biến đổi theo 24h của tháng đó. Vì vậy tôi đề xuất những ngày nào có giá trị quan trắc theo 24h có hệ số tương quan so với giá trị quan trắc trung bình của tháng theo 24h là thấp thì khả năng có dữ liệu nhiễu là rất lớn. Với những ngày có hệ

số tương quan thấp trong khoảng [-0.3; 0.3][5] sẽ được lọc ra để phân tích đánh giá thêm.

Bảng 17. Bảng kết quả thống kê danh sách những ngày có hệ số tương quan thấp so với giá trị trung bình tháng 01/2011

Ngày quan trắc	**Hệ số tương quan giữa số liệu 24h của ngày ngày và trung bình 24h giờ của tháng**
03/01/2011	-0.2829
04/01/2011	0.2108
09/01/2011	-0.0953
11/01/2011	0.1110
13/01/2011	0.1502
17/01/2011	0.2299
19/01/2011	-0.2411
23/01/2011	-0.0405

Kết quả áp dụng với bộ dữ liệu Training được thể hiện như Bảng 17. Muốn đánh giá được những ngày dữ liệu này có thực sự là những dữ liệu nhiễu hay không, cần có những kinh nghiệm chuyên môn về khí tượng, môi trường cộng thêm các đánh giá hiện trạng ô nhiễm từ các nguồn phát thải như giao thông, khu công nghiệp, sinh hoạt người dân… tại các khu vực xung quanh tại thời điểm quan trắc, đồng thời xem xét lại chĩnh thiết bị quan trắc vận hành tại thời điểm đó.

Bởi những giới hạn về kinh nghiệm, dữ liệu bổ trợ, cũng như kiến thức chuyên môn nên tại bước này tôi chưa thể xác định được những ngày phát hiện được liệu có chính xác là những dữ liệu sai lệch hay không. Vì vậy tôi quyết định giữ nguyên, không loại bỏ để đảm bảo tính toàn vẹn của bộ dữ liệu.

[5] Dựa nội dung 2.1.3.1

3.2.4 Xử lý dữ liệu thiếu.

Đặc trưng của những dữ liệu quan trắc khí tượng và môi trường là các chỉ tiêu có đều có một mối tương quan với nhau. Chính bởi vậy để điền dữ liệu thiếu một cách hoàn chỉnh tôi đưa ra phương án dựa vào mức độ tương quan giữa các chỉ tiêu quan trắc để xây dựng một hàm hồi quy tuyến tính. Từ hàm hồi quy này có thể dự đoán được những giá trị quan trắc bị thiếu.

Áp dụng với dữ liệu quan trắc PM10. Ví dụ, để có thể điền dữ liệu thiếu cho chỉ tiêu PM10 của tập Test ta cần có một bộ dữ liệu làm tập Training. Như đã đề cập bộ dữ liệu Training mà tôi sử dụng là bộ dữ liệu quan 01/2011. Bộ dữ liệu Test có thời gian quan trắc trong tháng 01/2012. Từ bộ dữ liệu Training, hàm hồi quy cho PM10 của tháng 01/2011 sẽ được xây dựng, trong bộ Test dữ liệu tháng nào sẽ được chạy mô hình hồi quy tương ứng với tháng đó, cụ thể mô hình hồi quy của tháng 01/2011 sẽ chạy trên bộ dữ liệu tháng 01/2012 để dự đoán và đánh giá kết quả giá trị quan trắc PM10.

Cụ thể phương pháp thực hiện trên ngôn ngữ R qua các bước như sau:

Input đầu vào:

- *Tập dữ liệu Training và chỉ tiêu quan trắc cần điền dữ liệu thiếu (tạm gọi chỉ tiêu này là X)*
- *Tập dữ liệu Test hay tập cần điền dữ liệu thiếu*

Quy trình:

1. *Tính tương quan giữa X và các chỉ tiêu quan trắc khác*
2. *Lựa chọn những chỉ tiêu quan chắc có tương quan lớn với X để dựng mô hình hồi quy*
3. *Dựng mô hình hồi quy tuyến tính từ tập Training*
4. *Chạy mô hình hồi quy trên tập Test để điền dữ liệu thiếu*
5. *Đánh giá mô hình. Nếu dữ liệu được xử lý không đúng với yêu cầu, chọn lại chỉ tiêu quan trắc để xây dựng mô hình tại bước 2*

Chi tiết nội dung từng bước được trình bày như sau:

a) Tính tương quan

Bảng thống kê tương quan giữa PM10 và các chỉ tiêu quan trắc khác từ tập dữ liệu tháng 01/2011 hay tập Training được thể hiện như sau:

Bảng 18. Bảng kết quả tương quan giữa PM10 với các chỉ tiêu quan trắc khác thời điểm tháng 01/2011

	PM10
WindSpd	0.04982
WindDir	0.03815
Temp	0.08365
RH	-0.34409
Barometer	0.03855
Radiation	-0.0124

	PM10
InnerTemp	0.02089
NO	0.23985
NO_2	0.59005
SO_2	0.53962
CO	0.44486
O_3	0.09338

b) Lựa chọn chỉ tiêu quan trắc để dựng mô hình hồi quy

Từ bảng tương quan ta có thể thấy được các chỉ tiêu quan trắc có hệ số tương quan lớn với PM10 có thể kể tới như NO_2, SO_2, CO. Kết quả này có thể tin cậy được bởi một số bài nghiên cứu quốc tế [24,25,26] cũng cho kết quả tương đồng về danh sách các chỉ tiêu có tương quan với PM10.

Như vậy có 3 chỉ tiêu được đưa vào danh sách xây dựng mô hình hồi quy để dự đoán giá trị PM10 đó là NO_2, SO_2, CO. Với 3 chỉ tiêu này ta có thể xây dựng được 7 mô hình hồi qui tuyến tính bằng cách thay đổi số lượng và vị trí các chỉ tiêu. Vậy từ 7 mô hình này ta sẽ chọn mô hình nào để tiên đoán PM10?

Để đánh giá một cách khách quan tôi sẽ thay đổi các tham số được lựa chọn, xây dựng và thử nghiệm đồng thời cả 7 mô hình này để đánh giá xem mô hình nào là tốt nhất.

c) Xây dựng mô hình hồi quy tuyến tính để điền dữ liệu thiếu

Để đảm bảo tính tương đồng về bộ dữ liệu xử lý khi chạy đồng loạt 7 mô hình cũng như để tiện so sánh kết quả tiên đoán với kết quả thực tế. Tôi tiến hành xóa bỏ tất cả những bản ghi có giá trị quan trắc thiếu dữ liệu PM10, NO_2, SO_2, CO từ tập Training. Nói dễ hiểu hơn là chỉ giữ lại những bản ghi có đồng thời cả 4 giá trị quan trắc PM10,

NO_2, SO_2, CO. Kết quả sau khi loại bỏ những bản ghi không phù hợp, có 725/744 bản ghi thỏa mãn yêu cầu. Tôi tạm gọi lập dữ liệu này là TrainingV1. Tập TrainingV1 được giả định là thiếu dữ liệu quan trắc PM10 100%, sau đó được dự đoán qua mô hình hồi quy và so sánh với giá trị quan trắc thực tế.

Dựa vào danh sách các tham số xây dựng mô hình hồi quy tuyến tính với 7 lần thử nghiệm thay đổi các tham số được lựa chọn. Kết quả so sánh khi chạy cả 7 mô hình hồi quy tuyến tính trên một tập dữ liệu TrainingV1 đạt được như Bảng 19.

Bảng 19. Bảng so sánh kết quả khi thử nghiệm 7 mô hình hồi quy tuyến tính.

Tham số mô hình	**Tổng số bản ghi**	**R^2 của mô hình**	**HS Tương quan***	**RMSE***	**MAPE***
SO_2	725	0.3	0.54	75.6	80
NO_2		0.35	0.6	72.6	74.7
CO		0.2	0.44	80.5	87.5
SO_2, NO_2		0.43	0.65	67.9	68.9
SO_2, CO		0.4	0.63	69.5	71.8
NO_2, CO		0.35	0.6	72.6	74.7
SO_2, NO_2, CO		0.43	0.66	67.6	68.8

** Bộ dữ liệu PM10 gốc và bộ dữ biệu PM10 dự đoán*

Nhìn chung trong 7 lần thay đổi các chỉ tiêu để xây dựng mô hình hồi quy cho PM10 thì mô hình được dựng bởi {SO_2, NO_2, CO} là mô hình có hệ số xác định cao nhất là 0.43 ứng với khả năng giải đoán được 43 % giá trị quan trắc. Sai số toàn phương trung bình sấp xỉ 68 µg/m^3, sai số trung bình tuyệt đối nằm ở ngưỡng 69%. Mô hình này khá tương đồng so với mô hình từ NO_2 và SO_2. Hệ số tương quan giữa bộ dữ liệu PM10 gốc và bộ dữ liệu dự đoán từ mô hình cũng nằm ở ngưỡng chấp nhận được với giá trị là gần 0.7. Từ đó 2 mô hình từ {NO_2, SO_2, CO} và {SO_2, NO_2} là 2 mô hình được lựa chọn ưu tiên số 1 khi xây dựng mô hình hồi quy tuyến tính.

Cho kết quả thấp hơn so với hai mô hình trên, mô hình được dựng bởi {SO_2, CO}, {NO_2, CO} và NO_2 thuộc nhóm có độ ưu tiên đứng thứ 2. Ba mô hình ứng với 3 nhóm chỉ tiêu cho kết quả khá tương đồng với nhau. Cụ thể, hệ số xác định của {NO_2, CO} và NO_2 cho kết quả sấp xỉ 0.35 ứng với hệ số tương quan là 0.6. Riêng hệ số xác định của {SO_2, CO} có hệ số xác định cao gần bằng 2 mô hình có độ ưu tiên thuộc nhóm số 1 với giá trị 0.4 chỉ nhỏ hơn 0.03 với mô hình thuộc nhóm số 1. Chính vì vậy trong 3 mô hình thuộc nhóm 2 thì mô hình được dựng bởi {SO_2, CO} có vẻ nhỉnh hơn về kết quả so với 2 mô hình kia. Kết quả đánh giá về sai số lỗi trung bình và phần trăm lỗi của 3 mô hình này cũng ngang như nhau là khoảng 70 µg/m^3 và 74%.

Nhóm thứ 3 có mô hình được dựng bởi SO_2 và mô hình được dựng bởi CO. Hai mô hình này được ghép vào nhóm cuối cùng bởi kết quả đạt được khá thấp về hệ số xác định cũng như tỉ lệ lỗi khá cao khoảng 80% với SO_2 và gần 90% với CO. Hai mô hình này được đưa vào danh sách mô hình lựa chọn cuối cùng.

Dựa trên so sánh các kết quả khi chạy từng mô hình. Tôi tiến hành đánh số phân cấp mức độ ưu tiên khi chạy các mô hình như sau:

Bảng 20. Bảng kết quả sắp xếp thứ tự các mô hình được đánh số tương ứng với mức độ ưu tiên.

Tham số mô hình	**Phương trình hồi quy**	**Đô ưu tiên/Đánh số mô hình**
SO_2, NO_2, CO	Y= -8.98 + 2.02*SO2 + 1.35*NO2 + 0.011*CO	**1**
SO_2, NO_2	Y= 0.79 + 1.87*SO2 + 1.80*NO2	**2**
SO_2, CO	Y= -1.95 + 2.59*SO2 + 0.028*CO	**3**
NO_2, CO	Y= 20.5 + 2.51*NO2 - 0.0004*CO	**4**
NO_2	Y= 20.2 + 2.5*NO2	**5**
SO_2	Y= 52.9 + 3.01*SO2	**6**
CO	Y= 42.5 + 0.04*CO	**7**

Kết hợp mô hình.

Theo lý thuyết ta sẽ lựa chọn mô hình này cho kết quả cao nhất để thực hiện điền dữ liệu thiếu nhưng trong thực tế dữ liệu lại không đơn giản như vậy. Trong 3 tham số SO_2, NO_2, CO, để chạy mô hình số 1 {SO_2, NO_2, CO} thì cần yêu cầu đồng thời 3 chỉ tiêu này phải có giá trị quan trắc. Nếu thiếu bất kì một trong 3 giá trị nào thì mô hình số 1 cũng không thể nào chạy được.

Đối với số lượng dữ liệu quan trắc theo giờ rất lớn. Cộng với các chỉ tiêu quan trắc có thể thiếu một cách ngẫu nhiên, với tỉ lệ thiếu dữ liệu ngẫu nhiên tại bất kì thời điểm quan trắc nào. Nếu chỉ chạy 1 mô mô hình hồi quy thì chắc chắn sẽ không đủ để điền dữ liệu thiếu cho bộ dữ liệu xử lý một cách hoàn thiện được. Chính vì vậy kết hợp các mô hình hồi quy là một giải pháp nhằm giải quyết được vấn đề này. Cụ thể với 7 mô hình hôi quy được xây dựng và đánh thứ tự ưu tiên mà tùy từng trường hợp bản ghi thiếu dữ liệu khác nhau sẽ chạy những mô hình hồi quy khác nhau:

Bảng 21. Bảng tổng kết các trường hợp thiếu dữ liệu và chạy mô hình hồi quy tuyến tính tương ứng.

Tình trạng bản ghi	**Chạy mô hình với độ ưu tiên số**
Bản ghi có đầy đủ dữ liệu SO_2, NO_2, CO	1
Bản ghi thiếu dữ liệu SO_2	4
Bản ghi thiếu dữ liệu NO_2	3
Bản ghi thiếu dữ liệu CO	2
Bản ghi thiếu dữ liệu SO_2, NO_2	7
Bản ghi thiếu dữ liệu SO_2, CO	5
Bản ghi thiếu dữ liệu NO_2, CO	6
Bản ghi thiếu tất cả dữ liệu SO_2, NO_2, CO	Không dự đoán được bằng mô hình hồi quy từ SO_2, NO_2, CO

3.2.5 Đánh giá kết quả.

Bảng 22. Tỉ lệ dữ liệu thiếu trước khi xử lý điền dữ liệu thiếu tháng 01/2012

Tháng 01/2012	SO_2	NO_2	CO	PM10
Tổng số bản ghi	744			
Số bản ghi thiếu	170	0	0	744

Bước đánh giá kết quả sẽ được thực hiện với bộ dữ liệu Test ứng với tháng 01/2012. Với bộ dữ liệu Test có tỉ lệ PM10 đầy đủ với mức 100% là một cơ sở tốt để so sánh kết quả của mô hình với kết quả thực tế, bởi vậy trong bước này tôi giả định xóa 100% dữ liệu PM10. Thêm nữa là dữ liệu quan trắc SO_2 thiếu khá lớn với 23% sẽ được thử nghiệm bước kết hợp các mô hình hồi quy. Như vậy đối với tập dữ liệu này sẽ kết hợp 2 mô hình hồi quy đó là mô hình được xây dựng từ {NO_2, SO_2, CO} và {NO_2, CO}

Bảng 23. Bảng kết quả dữ liệu tháng 01/2012 sau khi điền dữ liệu thiếu

Số bản ghi điền bởi mô hình từ NO_2, SO_2, CO	Số bản ghi điền bởi mô hình từ NO_2, CO	HS Tương quan*	RMSE*	MAPE*
574	170	0.56	51.4	45.3

**Giữa dữ liệu gốc và dự đoán từ mô hình kết hợp*

Như vậy với 100% số bản ghi PM10 thiếu kết quả cho thấy Hệ số tương quan giữa giá trị PM10 dự đoán và PM10 quan trắc được có độ tương đồng gần 0.6, sai số trung bình nằm ở mức 51 µg/m^3 và tỉ lệ lỗi nằm tại mức 45%. Kết quả này có thể chấp nhận được vì nó đảm bảo được tính hoàn thiển dữ liệu cũng như độ tương quan dữ liệu cũng đạt được ở mức trung bình.

Một thử nghiệm nhằm đánh giá ảnh hưởng của tỉ lệ thiếu dữ liệu khi chạy mô hình. Tôi tiến hành thử nghiệm mô hình với các bộ dữ liệu có tỉ lệ dữ liệu thiếu khác nhau. Từ với bộ dữ liệu Test, tôi tiền hành xóa dữ liệu PM10 ngẫu nhiên để tạo được các bộ dữ liệu PM10 thiếu tại các mức 10%, 20%, 30%, 40%, và 50%. Kết quả điền dữ liệu thiếu với mô hình hồi quy được kết hợp cho kết quả như bảng dưới.

Bảng 24. Bảng kết quả thử nghiệm bộ dữ liệu tháng 01/2012 với những tỉ lệ thiếu dữ liệu khác nhau.

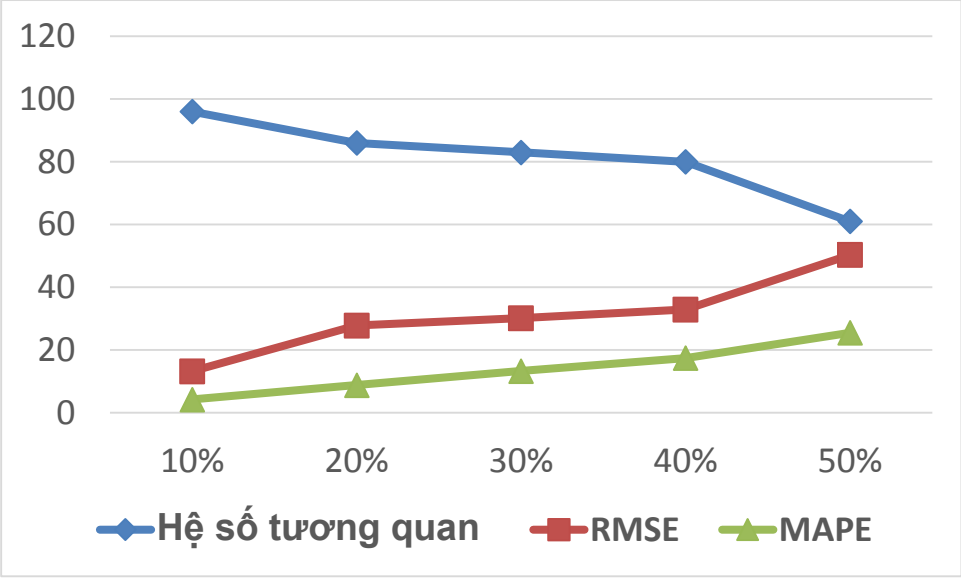
Tỉ lệ thiếu	**10%**	**20%**	**30%**	**40%**	**50%**
Tổng số bản ghi	744				
Số bản ghi PM10 thiếu	74	148	223	297	372
Số bản ghi điền bởi mô hình từ NO_2, SO_2, CO	63	119	168	236	319
Số bản ghi điền bởi mô hình từ NO_2, CO	11	29	55	61	53
Tỉ lệ bản ghi thiếu được điền dữ liệu	100%	100%	100%	100%	100%
Hệ số tương quan*	0.96	0.86	0.83	0.8	0.61
RMSE*	13.2	27.8	30.1	32.9	50.3
MAPE*	4.2	8.8	13.3	17.4	25.5

** Bộ dữ liệu PM10 gốc và bộ dữ liệu PM10 sau khi điền dữ liệu thiếu*

Đối với mỗi tập dữ liệu có tỉ lệ PM10 thiếu khác nhau đều cho một kết quả khác nhau. Kết quả này được mô tả trong Bảng 23 dựa trên tập dữ liệu đầy đủ với 744 bản ghi. Nhìn chung một cách tổng thể thì với những tỉ lệ thiếu dữ liệu khác nhau kết quả đánh giá cũng có những sai lệch đáng kể về sai số lỗi và phần trăm lỗi trung bình. Cụ thể, với bổ dữ liệu có tỉ lệ thiếu 10% cho kết quả tốt nhất với tỉ lệ lỗi trung binh chỉ ở mức 4.2% và sai số bình phương trung bình là 13 $\mu g/m^3$. Với tỉ lệ thiếu dữ liệu từ

10% đến 20% các chiri số đánh giá sai số lỗi và tỉ lệ lỗi có thể nói là cao gấp 2-3 lần so với trường hợp thiếu 10%, tỉ lệ lỗi trung từ 8-13% và sai số bình phương trung bình trong khoảng 27-30 µg/m^3. Với những tỉ lệ thiếu dữ liệu cao hơn từ 40% trở lên thì tỉ lệ lỗi trung binh bắt đầu tăng dần từ 17% đến 26% với trường hợp thiếu 50% dữ liệu và sai số bình phương trung bình khá lớn là 50 µg/m^3 ứng với bộ dữ liệu thiếu 50%. Biểu đồ thể hiện xu hướng chất lượng dữ liệu sau quá trình xử lý được mô tả như Hình 16. Trong biểu đồ, hệ số tương quan được nhân với 100 để dễ dàng vẽ và so sánh với các chỉ số khác.

Từ những quan sát, đánh giá thông qua các bộ dữ liệu thử nghiệm ứng với tỉ lệ thiếu dữ liệu khác nhau có thể nhận thấy được phương pháp hồi quy tuyến tính được sử dụng để điền giá trị thiếu có thể áp dụng tốt nhất với những bộ dữ liệu có tỉ lệ thiếu nhỏ hơn 10%. Với những bộ dữ liệu có tỉ lệ thiếu cao hơn nằm trong khoảng 20-30% vẫn có thể sử dụng được mô hình bởi tỉ lệ sai số có thể chấp nhận được. Với các trường hợp có tỉ lệ thiếu dữ liệu cao hơn thì phương pháp hồi quy tuyến tính dường như không cho kết quả tối ưu như người dùng mong muốn.



Hình 16. Biểu đồ mô tả kết quả ứng với từng tỉ lệ dữ liệu PM10 thiếu khác nhau.

3.3 Kết luận

Trong chương này, tôi đã tiến hành thực hiện chi tiết các bước đã đề xuất để chuẩn hóa dữ liệu quan trắc môi trường tại trạm Nguyễn Văn Cừ, Hà Nội với các tập

dữ liệu tháng 01 năm 2011 và 2012. Kết quả của quy trình là một bộ dữ liệu đã được chuẩn hóa về cấu trúc cũng như đảm bảo về chất lượng dữ liệu. Dựa trên các bước xử lý nhiễu, và xử lý thiếu dựa trên phương pháp hồi quy tuyến tính bộ dữ liệu cuối cùng được đảm bảo dữ liệu được toàn vẹn đạt được những tiêu chuẩn về dữ liệu nhất định và tính hoàn thiện dữ liệu sau quá trình xử lý.

Đối với việc xử lý dữ liệu thiếu, các kết quả thực nghiệm dựa trên tập dữ liệu Test với những tỉ lệ thiếu dữ liệu khác nhau từ 10%, 20%, 30%, 40%, 50% và 100% cho thấy với những tỉ lệ dữ liệu khác nhau thì kết quả có sự thay đổi khác nhau. Với bộ dữ liệu Test 01/2012 và mô hình hồi quy tuyến tính được xây dựng từ 01/2011 cho kết quả tốt với những bộ dữ liệu thiếu từ 10-30% dữ liệu quan trắc cần xử lý. Các kết quả này cho thấy tiềm năng của việc sử dụng phương pháp vào vào các bài toán thực tế. Tuy nhiên việc sử dụng mô hình hồi quy mới chỉ giải quyết được một phần nào đó những dữ liệu quan trắc thiếu bởi đối với những bản ghi mà các giái trị quan trắc được sử dụng để xây dựng mô hình mà không có dữ liệu thì không thể sử dụng mô hình hồi quy tuyến tính để dự đoán. Bởi vậy cần kết hợp thêm các phương pháp khác, qui trình xử lý khác để đảm bảo độ hoàn thiện dữ liệu

CHƯƠNG 4. NGHIÊN CỨU, PHÁT TRIỂN CÔNG CỤ HỖ TRỢ XỬ LÝ DỮ LIỆU QUAN TRẮC MÔI TRƯỜNG TẠI VIỆT NAM (ENVPRO).

4.1 Phát biểu bài toán

Cho đến thời điểm hiện tại việc xử lý và phân tích dữ liệu, khai thác thông tin từ dữ liệu quan trắc môi trường còn khá thô sơ. Chủ yếu các nhà phân tích sử dụng những công cụ như Word/Excel để mô tả, tính toán, thống kê, mô hình.. dữ liệu quan trắc. Giả sử trạm quan trắc không khí đo giá trị theo giờ với 10 tham số ô nhiễm vậy tính sơ 1 năm ta có khoảng 365 ngày*24 giờ = 8760 bản ghi. Kết quả này mới chỉ tính ứng với một chỉ tiêu quan trắc, vậy nếu nhà phân tích phải tính cả 10 chỉ tiêu quan trắc hoặc là dữ liệu trong 2 năm, 3 năm…thì đó thực sự là một công việc lớn sẽ gây ra lãng phí thời gian, công sức và tiền của. Thêm nữa là nó sẽ ảnh hưởng tới những quyết định mang tính kịp thời của nhà quản lý nhằm đưa ra phương hướng bảo vệ nguồn tài nguyên môi trường.

Dựa trên quy trình chuẩn hóa dữ liệu quan trắc đã đề xuất, kết quả thực nghiệm, thực tế dữ liệu quan trăc hiện có ở Việt Nam thì nhu cầu cấp thiết là cần có một công cụ trợ giúp nhà phân tích xử lý các công tác chuyên môn. Từ đó tôi đề xuất xây dựng công cụ có khả năng cung cấp thông tin và hỗ trợ phân tích và xử lý những dữ liệu quan trắc môi trường môt cách tự động. Với hệ thống này tôi gọi tên là EnvPro. Ý nghĩa của tên hệ thống được ghép từ Environment và Processing/Professional thể hiện cho những đặc tính mà hệ thống xây dựng đó là xử lý dữ liệu quan trắc môi trường một cách chuyên nghiệp.

4.2 Yêu cầu hệ thống

Hệ thống được xây dựng cần đảm bảo các yêu cầu sau:

- *Có khả năng lưu trữ dữ liệu lớn*: Dữ liệu quan trắc môi trường được lấy tại 2 mạng lưới. Thứ nhất là mạng lưới quan trắc khí tượng thủy văn và môi trường quốc gia. Thứ hai là mạng lưới quan trắc môi trường quốc gia do Tổng cục Môi trường quản lý.
- *Giao diện trực quan thân thiện với người dùng*: Hệ thống có giao diện thân thiện dễ sử dụng, cung cấp cái nhìn trực quan hỗ trợ người sử dụng trong quá trình phân tích và xử lý dữ liệu. Hệ thống không đòi hỏi người dùng phải có kiến thức chuyên môn về nghiên cứu chất lượng không khí cũng như những

khả năng về công nghệ. Nói chung người dùng có thể dễ dàng sử dụng và hiểu được những thông tin mà hệ thống cung cấp.

- *Cung cấp khả năng phân tích thống kê và biểu diễn dữ liệu trên biểu đồ*: Tính năng biểu diễn dữ liệu trên biểu đồ là rất quan trọng. Một biểu đồ được xây dựng mang rất nhiều giá trị hơn cả vạn chữ viết. Mọi thông tin cần thiết đều có thể trích xuất từ biểu đồ ra.
- *Cung cấp khả năng làm sạch dữ liệu*: Với yêu cầu này hệ thống sẽ giúp người dùng phát hiện ra những ngày tiềm tàng quan trắc sai hoặc không chính xác từ đó người dùng có thể đưa ra những hướng xử lý phù hợp. Ngoài ra khả năng này còn giúp xây dựng một bộ dữ liệu hoàn chỉnh bằng cách dựa vào thuật toán cụ thể để điền dữ liệu thiếu ứng với yêu cầu của người dùng.
- *Cho phép người dùng được tải về các file dữ liệu đã xử lý*: Các file dữ liệu được cho phép tải về không chỉ hỗ trợ cho công tác quản lý, sao lưu mà còn có thể là file dữ liệu nguồn cho một công trình nghiên cứu, phân tích khác tùy thuộc bài toán khác nhau.

4.3 Tổng quan hệ thống EnvPro.

Hệ thống hỗ trợ xử lý dữ liệu quan trắc môi trường EnvPro được chia làm 3 tầng bao gồm:

- Tầng quản lý dữ liệu: Thực hiện các nhiệm vụ liên quan đến lưu trữ, truy xuất dữ liệu của ứng dụng. Dữ liệu chính được sử dụng trong hệ thống đó là những dữ liệu quan trắc môi trường của các trạm quan trắc và thông tin người dùng cũng như phân quyền truy cập của hệ thống. Dữ liệu quan trắc môi trường được lấy tại 2 mạng lưới. Thứ nhất là mạng lưới quan trắc khí tượng thủy văn và môi trường quốc gia. Thứ hai là mạng lưới quan trắc môi trường quốc gia do Tổng cục Môi trường quản lý.

Bảng 25. Bảng thông tin các trạm quan trắc hiện có trên toàn lãnh thổ Việt Nam

Mạng lưới	Dạng	Khu vực
Mạng lưới quan trắc khí tượng thủy văn và	Trạm quan trắc chất lượng không khí tự	Hà Nội
		Hải Phòng

môi trường quốc gia	động	Ninh Bình
		Vinh
		Đà Nẵng
		Hỗ Chí Minh
		Pleiku
		Cần Thơ
		Sơn La
Mạng lưới quan trắc môi trường quốc gia - Tổng cục Môi trường	Trạm quan trắc chất lượng không khí tự động	556 Nguyễn Văn Cừ (Hà Nội)
		Lăng Chủ tịch Hồ Chí Minh (Hà Nội)
		Đà Nẵng
		Khánh Hòa
		Huế
		Phú Thọ
		Quảng Ninh

Dữ liệu quan trắc được lưu lại theo thời gian thực tức sau mỗi giờ trạm quan trắc sẽ gửi gữ liệu về server và lưu vào cơ sở dữ liệu.Trước khi lưu vào CSDL các dữ liệu chi tiết được đặt lại về chuẩn theo đúng qui ước đảm bảo sự thống nhất về cấu trúc của toàn bộ dữ liệu với nhau cũng như giữa các trạm, các mạng lưới với nhau.

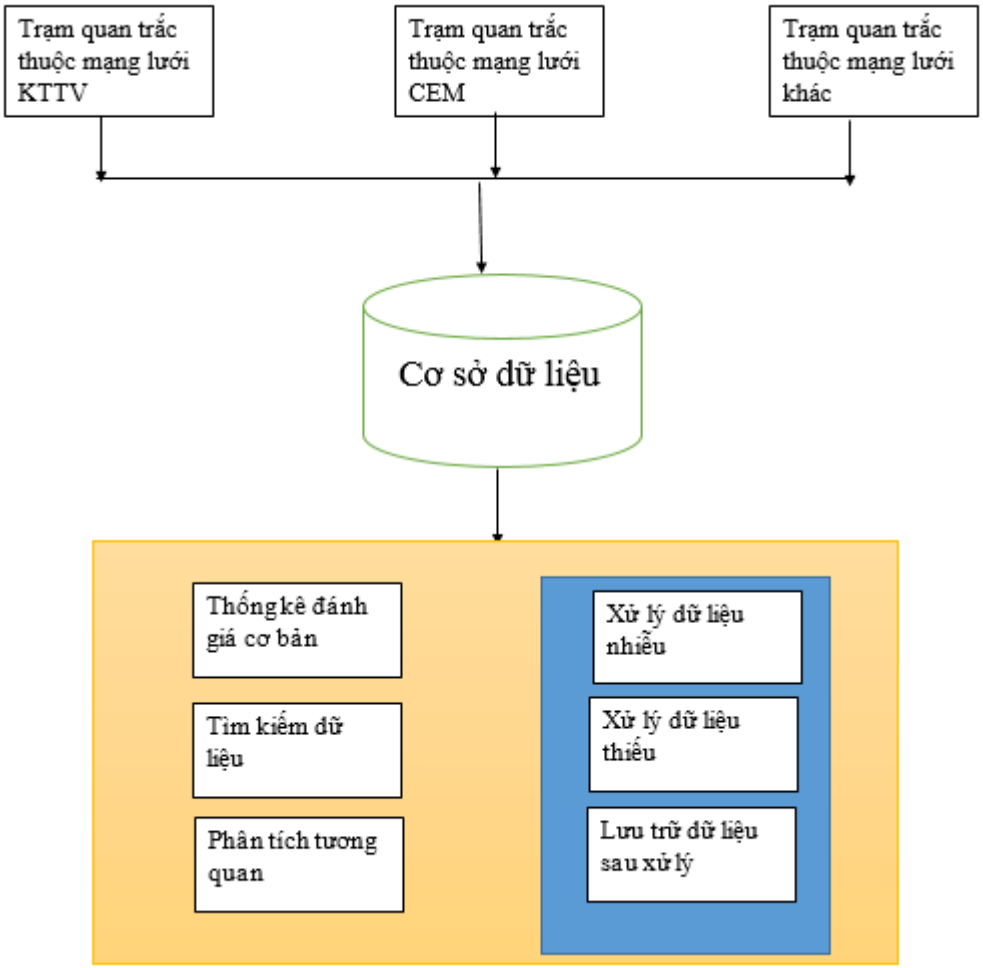
- Tầng xử lý dữ liệu và mô hình hóa: Thực hiện các nghiệp vụ liên quan đến xử lý các thuật toán dựa trên dữ liệu của hệ thống. Nó sử dụng dữ liệu truy xuất từ tầng quản lý dữ liệu là cung cấp dữ liệu cho người dùng thông qua cổng dịch vụ web. Ở tầng xử lý dữ liệu và mô hình hóa các mô đun phần mềm sẽ chạy các thuật toán xử lý dữ liệu như tính toán thống kê, phát hiện dữ liệu bất thường, điền dữ liệu thiếu… dữ liệu được xử lý sau đó trả về cho người dùng thông qua các biểu đồ được mô hình hóa để mô tả dữ liệu quan trắc.

- Tầng giao diện người dùng: Làm nhiệm vụ cung cấp các thành phần giúp người dùng tương tác với hệ thống, hiển thị kết quả/dữ liệu trực quan ra màn hình.

4.4 Phân rã chức năng và người dùng

4.4.1 Phân rã chức năng

- Nhóm chức năng tìm kiếm dữ liệu:
  - Tìm kiếm theo mạng lưới/nguồn dữ liệu
  - Tìm kiếm theo vị trí quan trắc
  - Tìm kiếm theo chỉ tiêu quan trắc
  - Tìm kiếm theo khoảng thời gian
- Nhóm chức năng thống kê, đánh giá dữ liêu cơ bản
  - Thống kê dữ liệu thiếu.
  - Thống kê cơ bản: Tính toán chỉ số thống kê Min, Max, Median, Mean, Mode, Quartile, Range, Variance, Standard Deviation
  - Tính trung bình từng giờ với toàn bộ dữ liệu trong tháng
  - Thống kê ngày vượt tiêu chuẩn an toàn: Theo tiêu chuẩn chất lượng không khí Việt Nam, WHO, USA. Thời gian thống kê theo tháng/năm
  - Vẽ biểu đồ: Dạng Line, Column, Boxplot



Hình 17. Sơ đồ tổng quan về cấu trúc các chức năng hệ thống EnvPro

- Nhóm chức năng xử lý dữ liệu nhiễu
  - Phát hiện dữ liệu bất thường
  - Loại bỏ dữ liệu bất thường
  - Loại bỏ dữ liệu ngoài khoảng giá trị mong muốn
  - Vẽ biểu đồ thể hiện dữ liệu trước và sau khi xử lý
- Nhóm chức năng xác định tương quan
  - Tính hệ số tương quan giữa các chỉ tiêu quan trắc trong 1 trạm
  - Tinh hệ số tương quan giữa các trạm với nhau
  - Vẽ hiểu đồ mô tả tương quan dữ liệu
- Nhóm chức năng điền dữ liệu thiếu
  - Sử dụng thuật toán hồi quy tuyến tính

- Sử dụng bộ dữ liệu khác
- Biểu đồ minh họa trước và sau khi xử lý

- Chức năng lưu trữ dữ liệu sau xử lý: Chức năng cho phép người dùng tải về các file dữ liệu mong muốn. Dữ liệu được tải và lưu với định dạng *.CSV

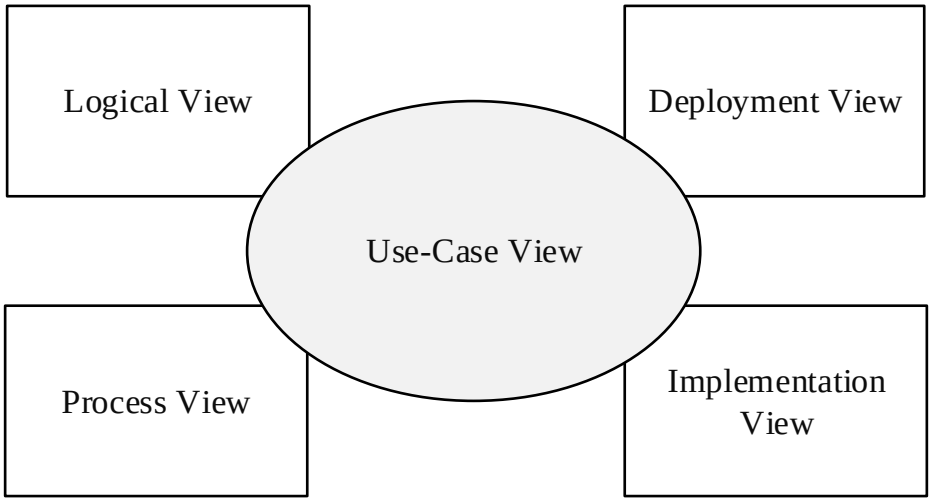
4.4.2 Nhóm người dùng

- Người dùng thông thường: Là toàn bộ người dùng có nhu cầu sử dụng các tính năng của hệ thống để hỗ trợ xử lý dữ liệu quan trắc môi trường. Người dùng thông thường được phép sử dụng hết tất cả các chức năng hệ thống ngoại trừ chức năng tìm kiếm dữ liệu. Hệ thống cho phép người dùng tải lên file dữ liệu quan trắc cá nhân để xử lý. Các file này phải tuân theo qui ước của hệ thống đã đề ra.
- Nhà nghiên cứu: Bao gồm các chuyên gia của các đơn vị sử dụng hệ thông. Truy xuất dữ liệu từ hệ thống và sử dụng các chức năng của hệ thống để xử lý, phân tích đưa ra các báo cáo về môi trường, sức khỏe cộng đồng.... Nhóm người dùng này có quyền sử dụng toàn bộ chức năng của hệ thống với điều kiện phải Đăng nhập
- Nhà quản trị hệ thống: Có trách nhiệm quản lý dữ liệu, thực hiện các công việc thiết lập của hệ thống, phân quyền người dùng, theo dõi và kiểm soát việc vận hành hệ thống

4.5 Nguyên tắc và ràng buộc thiết kế

4.5.1 Nguyên tắc thiết kế

Dựa vào các yêu cầu bài toán, hệ thống chia sẻ thông tin cảo báo ô nhiễm không khí được phân tích, thiết kế và trình bày theo mô hình 4+1, mô tả hệ thống dưới nhiều khung nhìn để miêu tả các khía cạnh kiến trúc của hệ thống. Mô hình này gồm 5 thành phần chính (gọi là view):



Hình 18. Mô hình kiến trúc 4+1.

- Use-Case View: Mô tả các kịch bản hoặc các ca sử dụng đại diện cho các chức năng chính mà hệ thống cung cấp như đã nêu ở phần …
- Logical View: Mô tả các các chức năng được thiết kế bên trong hệ thống chia sẻ thông tin cảnh báo ô nhiễm không khí.
- Process View: Mô tả quy trình xử lý và giao tiếp trong hệ thống, các thao tác dữ liệu giữa client và server.
- Deployment View: Mô tả cách hệ thống được triển khai trên server của trung tâm công nghệ tích hợp liên ngành và giám sát hiện trường – FIMO thuộc Đại Học Công Nghệ - Đại Học Quốc Gia Hà Nội.
- Implementation View: Mô tả các mô đun được tích hợp vào hệ thống chia sẻ thông tin cảnh báo ô nhiễm không khí.

4.5.2 Ràng buộc thiết kế

Chuẩn dữ liệu giao tiếp:

Hệ thống sử dụng ngôn ngữ thống kê R hỗ trợ cho quá trình xử lý. Mọi dữ liệu được chia sẻ từ hệ thống tới R và ngược lại được áp dụng với kiểu dữ liệu là JSON. Ngoài ra các phương thức giao tiếp dữ liệu giữa Client và Server cũng sử dụng định dạng JSON theo chuẩn RESTful, dịch vụ web RESTful được xây dựng bằng Yii2 – một framework của PHP. Lợi ích của việc thiết kế theo chuẩn giao tiếp này là hệ thống có thể dễ dàng triển khai trên nhiều nền tảng phần cứng như ứng dụng web, ứng dụng mobile, ứng dụng desktop. Các ngôn ngữ nổi bật hiện nay như PHP, C#, JAVA… đề có phương thức hỗ trợ chuyển đổi dễ dàng dữ liệu sang JSON và ngược lại giúp thao tác với dữ liệu dễ dàng hơn.

Tích hợp Apache Server:

Apache là 1 open-source Web server và miễn phí hoàn toàn được hỗ trợ bởi Apache Software Foundation (http://httpd.apache.org) Mặc dù miễn phí và Open-Source nhưng Apache có sức mạnh và tính ổn định của nó được đánh giá ở tầm cỡ thương mại. Đến nay đây vẫn là lựa chọn tốt nhất cho giải pháp máy chủ Web. Với hệ thống EnvPro, Apache được sử dụng để triển khai ứng dụng trên nền web của hệ bao gồm client, server và các api cung cấp dữ liệu sử dụng ngôn ngữ php.

Xác thực (Authentication):

Với hệ thống EnvPro người dùng muốn sử dụng chức năng tìm kiếm dữ liệu bắt buộc phải Đăng nhập. Việc xác thực danh tính của người dùng được bắt đầu khi người dùng khởi tạo một session bằng việc truy vập vào trang web. Khi người đó muốn sử dụng chức năng tìm kiếm hệ thống sẽ đưa ra yêu cầu đăng nhập. Ta có thể dử dụng UserApi Service cho phép người dùng truy xuất thông tin chi tiết của họ thông qua web service của hệ thống.

Phân quyền (Authorization):

Hệ thống EnvPro sử dụng phương pháp phân quyền dựa trên vai trò. Vì vậy một người dùng chỉ được phép truy cập các chức năng mà họ được gán. Việc phân quyền có thể được thực hiện tại nhiều phạm vi từ module tới từng hành động xem/sửa/xóa/thêm. Với mỗi phạm vi cơ chế phân quyền sẽ kiểm tra quyền của người dùng và quyết định có cho người dùng kích hoạt hành động hay không.

4.6 Công nghệ sử dụng

Dựa trên những thông tin về yêu cầu hệ thống, dựa trên những mô hình kiến trúc và công nghệ đã tìm hiểu. Tôi quyết định xây dựng hệ thống hỗ trợ xử lý dữ liệu quan trắc môi trường EnvPro dựa trên công nghệ mã nguồn mở. Nhìn chung lợi ích của việc sử dụng mã nguồn mở có thể kể tới như:

- Tính kinh tế: Được cung cấp miễn phí, giúp nhà phát triển thu được lợi nhuận lớn. Đặc biệt giúp hạn chế vấn đề vi phạm bản quyền, quyền sở hữu trí tuệ.
- Tính an toàn: Mã nguồn được phổ biến rộng rãi vì vậy việc mã nguồn được phổ biến rộng rãi giúp người lập trình và người sử dụng dễ phát hiện, khắc phục các lỗ hỏng an toàn trước khi chúng bị lợi dụng
- Tính phổ biến: Được cộng đồng sử dụng đông đảo, dễ dàng hỗ trợ người dùng mới trong quá trình phát triển phần mềm

- Tính tương thích: Dễ dàng tích hợp với những ngôn ngữ, hệ thống mã nguồn mở khác nhau
- Dễ dàng tùy biến: Các doanh nghiệp có thể biến đổi một phần của gói phần mềm mã nguồn mở để biến chúng phù hợp với những nhu cầu của mình.
- Cập nhật thường xuyên: Những phiên bản mã nguồn mở của các phần mềm được cập nhật thường xuyên và liên tục bao gồm các bản vá lỗi cũng như mở rộng các chức năng.

Những công nghệ mà tôi quyết định sử dụng luôn hướng tới mục tiêu đó là khả năng cung cấp và hỗ trợ đầy đủ, đa dạng, cộng đồng người sử dụng lớn, tính phổ biến cũng như được cập nhật thường xuyên. Cu thể các ngôn ngữ lập trình/phần mềm tôi sử dụng để phát triển hệ thống EnvPro là PHP, R, Jquery và PostgreSQL.

4.6.1 PHP – Yii 2.0 framework

PHP là 1 ngôn ngữ script rất phổ biến hiện nay bởi những lý do: linh hoạt, dễ sử dụng, dễ học, vv…. nhưng đôi khi việc viết mã PHP, hay bất cứ ngôn ngữ (lập trình) nào khác, có thể trở nên đơn điệu và lủng củng. Đó là lúc PHP framework có thể giúp ta.

Có rất nhiều lý do khác nhau để các lập trình viên sử dụng PHP framework, nhưng 1 trong những lý do chính vẫn là khả năng giúp các lập trình viền tăng tốc quá trình phát triển ứng dụng. Việc sử dụng lại các mã lệnh giống nhau trong nhiều project sẽ giúp tiết kiệm được thời gian và công sức 1 cách đáng kể. Một framework sẽ cung cấp sẵn các module nền tảng cần thiết để xây dựng 1 project, vì thế, các lập trình viên có thể tận dụng được thời gian để phát triển các ứng dụng thực tế, hơn là mất thời gian để xây dựng lại nền tảng trên mỗi project.

Sự ổn định là 1 lý do lớn đối với các lập trình viên đang sử dụng Framework. Tính đơn giản là 1 điểm mạnh của PHP, đó là lý do tại sao lại có nhiều người thích sử dụng nó. PHP thì khá dễ học và sử dụng, đặc biệt là đối với những người mới làm quen với lập trình, tuy nhiên, họ có thể thường xuyên viết mã 1 cách không khoa học và thậm chí không hề nhận thức được điều này, với PHP, trong nhiều trường hợp các ứng dụng vẫn sẽ làm việc được, nhưng vô tình họ có thể tạo ra các lỗ hổng bảo mật lớn trong mã lệnh của mình, và bị hacker khai thác [42].

Chính bởi vậy, xét về hướng giao diện người dùng, server và các API giao tiếp dữ liệu của hệ thống EnvPro tôi sử dụng framework Yii 2.0 của PHP. Yii2 hỗ trợ xây dựng nhanh các dịch vụ web theo chuẩn RESTful với đầu ra dữ liệu là json hoặc xml

kết hợp với R và Jquery để xử lý và mô hình hóa dữ liệu trực quan. Việc sử dụng RESTful tạo luồng lưu thông dữ liệu hỗ trợ cung cấp dữ liệu đa nền tảng từ desktop cho tới mobile. Với những ứng dụng cụ thể chỉ cần có dịch vụ cung cấp dữ liệu từ server thì đều có thể xử lý và tương tác với server.

4.6.2 Ngông ngữ thống kê R

Trong hệ thống EnvPro mà tôi xây dựng thì các nhiệm vụ thống thống kê và xử lý dữ liệu quan trắc môi trường là những nhiệm vụ cốt lõi. Với số lượng các bản ghi đầu vào rất lớn kèm theo việc phân tích chi tiết từng chỉ tiêu quan trắc là khá tốn thời gian. Vì vậy hệ thống cần có một qui trình xử lý nhanh nhạy và đơn giản. Với những thông tin thông qua nghiên cứu và sử dụng thực tế tôi tích hợp R, một ngôn ngữ hỗ trợ công tác thống kê và phân tích dữ liệu với hệ thống của mình.

Nói một cách ngắn gọn, R là một phần mềm sử dụng cho phân tích thống kê và đồ thị. Thật ra, về bản chất, R là ngôn ngữ máy tính đa năng, có thể sử dụng cho nhiều mục tiêu khác nhau, từ tính toán đơn giản, toán học giải trí (recreational mathematics), tính toán ma trận (matrix), đến các phân tích thống kê phức tạp. Vì là một ngôn ngữ, cho nênta có thể sử dụng R để phát triển thành các phần mềm chuyên môn cho một vấn đề tính toán cá biệt.

Hai người sáng tạo ra R là hai nhà thống kê học tên là Ross Ihaka và Robert Gentleman. Kể từ khi R ra đời, rất nhiều nhà nghiên cứu thống kê và toán học trên thế giới ủng hộ và tham gia vào việc phát triển R. Chủ trương của những người sáng tạo ra R là theo định hướng mở rộng (Open Access). Cũng một phần vì chủ trương này mà R hoàn toàn miễn phí. Bất cứ ai ở bất cứ nơi nào trên thế giới đều có thể truy nhập và tải toàn bộ mã nguồn của R về máy tính của mình để sử dụng. Cho đến nay, chỉ qua chưa đầy 5 năm phát triển, càng ngày càng có nhiều các nhà thống kê học, toán học, nghiên cứu trong mọi lĩnh vực đã chuyển sang sử dụng R để phân tích dữ liệu khoa học. Trên toàn cầu, đã có một mạng lưới gần một triệu người sử dụng R, và con số này đang tăng theo cấp số nhân. Có thể nói trong vòng 10 năm nữa, chúng ta sẽ không cần đến các phần mềm thống kê đắt tiến như SAS, SPSS hay Stata (các phần mềm này rất đắt tiền, có thể lên đến 100.000 USD một năm) để phân tích thống kê nữa, vì tất cả các phân tích đó có thể tiến hành bằng R . Trên thế giới hiện có rất nhiều các công ty lớn sử

dụng R trong các sản phẩm của mình có thể nói tới như Google[6]. Họ đã sử dụng R kết hợp với công cụ Google Analytics thực hiện các phân tích thống kê, mô hình hóa dữ liệu để hiểu rõ và cải thiện vấn đề kinh doanh của mình.

R là một phần mềm miễn phí, là một phần mềm mã mở [7], nhưng khả năng xử lý số liệu không thua kém bất cứ một phần mềm phân tích số liệu thương mại đắt tiền hiện hành nào khác. Phần hấp dẫn khác của R, như bản chất mã nguồn mở là người sử dụng có thể tự viết chương trình theo ý của riêng mình một khi đã nắm vững ngôn ngữ R. Các tài liệu tham khảo luôn được cập nhật từ khắp nơi, thuận tiện cho việc tham khảo. R đang tạo nên sức mạnh cho cuộc cách mạng của dữ liệu lớn, đối với bất cứ ai có nhu cầu phân tích dữ liệu nghiêm túc. Từ khoa học và kinh doanh cho đến giải trí và truyền thông xã hội, R là ngôn ngữ dùng để phân tích thống kê trên gần như mọi lĩnh vực cuộc sống. Với những đặc điểm vừa mô tả R là một công cụ thích hợp, tiện dụng, miễn phí mà có thể sử dụng rộng rãi để phục vụ công tác phát triển và mở rộng nghiên cứu khoa học ở Việt Nam. [7]

4.6.3 Jquery

Với sự phát triển rất mau lẹ của Internet, người dùng ngày càng quan tâm hơn đến hình thức của một trang web. Trước đây một trang web chỉ cần có banner, nội dung và ít footer là đã được cho là một trang web hoàn chỉnh. Nhưng bây giờ trang web đó phải có banner bắt mắt, nội dung hay và còn nhiều hiệu ứng lạ mắt khác nữa thì mới có thể thu hút được người đọc.

Chính vì thế những nhà thiết kế web bắt đầu chú ý đến các thư viện JavaScript mở như jQuery để tạo ra các hiệu ứng có thể tương tác trực tiếp với người đọc một cách nhanh chóng và dễ dàng hơn rất nhiều là sử dụng thuần JavaScript. Cụ thể Jquery có thể làm được:

- Hướng tới các thành phần trong tài liệu HTML. Nếu không sử dụng thư viện JavaScript này, bạn phải viết rất nhiều dòng code mới có thể đạt được mục tiêu là di chuyển trong cấu trúc cây (hay còn gọi là DOM hay Document Object Model) của một tài liệu HTML và chọn ra các thành

[6] http://www.r-bloggers.com/r-at-google/

[7] https://cran.r-project.org/

phần liên quan. Jquery cho phép bạn chọn bất cứ thành phần nào của tài liệu để "tác động" một cách dễ dàng như sử dụng CSS.

- Thay đổi giao diện của một trang web. CSS là công cụ rất mạnh để định dạng một trang web nhưng nó có một nhược điểm là không phải tất cả các trình duyệt đều hiển thị giống nhau. Cho nên jQuery ra đời để lấp chỗ trống này, vì vậy các bạn có thể sử dụng nó để giúp trang web có thể hiển thị tốt trên hầu hết các trình duyệt. Hơn nữa jQuery cũng có thể thay đổi class hoặc những định dạng CSS đã được áp dụng lên bất cứ thành phần nào của tài liệu HTML ngay cả khi trang web đó đã được trình duyệt load thành công. Thay đổi nội dung của tài liệu. Jquery không phải chỉ có thể thay đổi bề ngoài của trang web, nó cũng có thể thay đổi nội dung của chính tài liệu đó chỉ với vài dòng code. Nó có thể thêm hoặc bớt nội dung trên trang, hình ảnh có thể được thêm vào hoặc đổi sang hình khác, danh sách có thể được sắp xếp lại hoặc thậm chí cả cấu trúc HTML của một trang web cũng có thể được viết lại và mở rộng. Tất cả những điều này bạn hoàn toàn có thể làm được nhờ sự giúp đỡ của API (Application Programming Interface).
- Tương tác với người dùng. Cho dù công cụ bạn dùng có mạnh mẽ đến mấy, nhưng nếu bạn không có quyền quyết định khi nào nó được sử dụng thì công cụ đó cũng coi như bỏ. Với thư viện javaScript như jQuery, nó cho bạn nhiều cách để tương tác với người dùng ví dụ như khi người dùng nhấp chuột vào đường link thì sẽ có gì xảy ra. Nhưng cái hay của nó là không làm cho code HTML của bạn rối tung lên chính là nhờ các Event Handlers. Hơn nữa Event Handler API sẽ bảo đảm rằng trang web của bạn tương thích hầu hết với các trình duyệt, điều này đã và đang làm đau đầu rất nhiều các web designer.
- Tạo hiệu ứng động cho những thay đổi của tài liệu. Để tương tác tốt với người dùng, các web designer phải cho người dùng thấy được hiệu ứng gì sẽ xảy ra khi họ làm một tác vụ nào đó. Jquery cho phép bạn sử dụng rất nhiều hiệu ứng động như mờ dần, chạy dọc chạy ngang v.v.. và nếu vẫn chưa đủ, nó còn cho phép bạn tự tạo ra các hiệu ứng của riêng mình.
- Lấy thông tin từ server mà không cần tải lại trang web. Đây chính là công nghệ ngày càng trở nên phổ biến Asynchronous JavaScript And XML (AJAX), nó giúp người thiết kế web tạo ra những trang web tương tác cực tốt và nhiều tính năng. Thư viện jQuery loại bỏ sự phức tạp của trình duyệt trong quá trình này và cho phép người phát triển web có thể tập

trung vào các tính năng đầu cuối. Đơn giản hoá các tác vụ javaScript. Ngoài những tính năng như đã nêu ở trên, jQuery còn cho phép bạn viết code javaScript đơn giản hơn nhiều so với cách truyền thống như là các vòng lặp và điều khiển mảng.

Từ những ưu điểm của Jquery cũng như các đặc điểm của hệ thống EnvPro đó là dữ liệu được mô hình hóa một cách trực quan thông qua giao diện biểu đồ trên trình duyệt web. Vì vậy tôi lựa chọn Jquery để hỗ trợ công việc xây dựng biểu đồ, tăng tính tương tác giữa người dùng và hệ thống EnvPro.

4.6.4 PostgreSQL

PostgreSQL là một hệ quản trị cơ sở dữ liệu quan hệ và đối tượng dựa trên POSTGRES, bản 4.2, được khoa điện toán của đại học California tại Berkeley phát triển. POSTGRES mở đường cho nhiều khái niệm quan trọng mà các hệ quản trị dữ liệu. Hệ quản trị CSDL này là một chương trình mã nguồn mở xây dựng trên mã nguồn ban đầu của đại học Berkeley. Nó theo chuẩn SQL99 và có nhiều đặc điểm hiện đại:

- Câu truy vấn phức hợp (complex query)
- Khóa ngoại (foreign key)
- Thủ tục sự kiện (trigger)
- Các khung nhìn (view)
- Tính toàn vẹn của các giao dịch (integrity transactions)
- Việc kiểm tra truy cập đồng thời đa phiên bản (multiversion concurrency control)

Hơn nữa, PostgreSQL có thể dùng trong nhiều trường hợp khác, chẳng hạn như tạo ra các khả năng mới như:

- Kiểu dữ liệu
- Hàm
- Toán tử
- Hàm tập hợp
- Phương pháp liệt kê
- Ngôn ngữ theo thủ tục

PostgreSQL hiện đang được dùng phổ biến ở nhiều nơi. Nó không quy định những hạn chế trong việc sử dụng mã nguồn của phần mềm. Bởi vậy PostgreSQL có thể được dùng, sửa đổi và phổ biến bởi bất kỳ ai cho bất kỳ mục đích nào.

Đây cũng là hệ quản trị cơ sở dữ liệu hỗ trợ mạnh trong việc lưu trữ dữ liệu không gian. PostgreSQL kết hợp với module Postgis cho phép người dùng lưu trữ các lớp dữ liệu không gian. Khi sử dụng PostgreSQL, Postgis kết hợp với các phần mềm GIS hỗ trợ hiển thị, truy vấn, thống kê hoặc xử lý dữ liệu không gian. thương mại rất lâu sau mới có.

Chính vì những ưu điểm như vậy mà PostgreSQL đã đạt được sự hài lòng của những người sử dụng và cả chuyên gia về công nghệ thông qua các giải thưởng như Linux New Media dành cho hệ điều hành tốt nhất và năm lần chiến thắng giải do tạp chí Linux Journal Editors bình chọn về DBMS tốt nhất.

4.7 Môi trường phát triển và thực thi

Bảng 26, 27 mô tả các thành phần của môi trường phát triển và thực thì gồm các thành phân là phần cứng, phần mềm, hạ tầng mạng, cơ sở dữ liệu. Các tài nguyên này được cài đặt và thiết lập trên máy tính cá hân và máy chủ thuộc trung tâm công nghệ tích hợp liên ngành và giám sát hiện trường – FIMO thuộc Đại Học Công Nghệ.

Bảng 26. Bảng mô tả môi trường phát triển hệ thống EnvPro

STT	***Tên thành phần***	***Mô tả***
1	Hệ điều hành	- Lập trình viên: Microsoft Windows 7, 8.1 (English) - Máy chủ CSDL: Linux Cent OS 7.0 - Máy chủ ảo chạy ứng dụng: Linux Cent OS 7.0 - Máy chủ thật host các máy ảo: Cent OS 6.5
2	Tầng trung gian bao gồm cơ sở dữ liệu, máy chủ web	- Hệ quản trị CSDL: PostgreSQL 9.4 - Máy chủ Web: Apache Web Server 2
3	Phần mềm	- R 3.2.3 - PHP 5.5 - PHP Yii2 Framework 2.0 - Microsoft Office 2007 - Adobe Flash Player 10+ - Google Chrome 50.0.2661.87

4	Phần cứng	- Lập trình viên: Intel Core i5 – M450 2.40 Ghz/4GB RAM. - Máy chủ CSDL: Intel Xeon 4 core 2.5GHz+/8GB RAM - Máy chủ ảo chạy ứng dụng: Intel Xeon 4 core 2.5GHz+/8GB - Máy chủ thật host các máy ảo: Intel Xeon 80 core 2.5GHz+/128GB
5	Ngôn ngữ lập trình và các công cụ phát triển hệ thống	*Ngôn ngữ phát triển hệ thống:* - PHP Script 5.5 - R 3.2.3 - Jquery *Công cụ phát triển hệ thống:* - JetBrains PhpStorm 10.0.3, PHP Composer - Yii2 Framework 2.0 - Rstudio 0.98.1103 - Google Chrome 50.0.2661.87

Môi trường thực thi

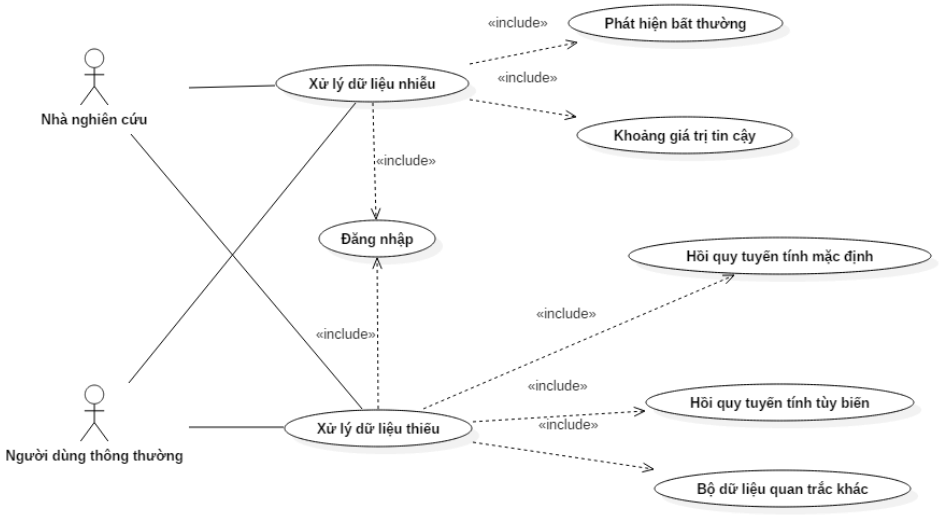
Bảng 27. Bảng mô tả môi trường thực thi hệ thống EnvPro

STT	***Tên thành phần***	***Mô tả***
1	*Hệ điều hành*	- Máy chủ CSDL: Linux Cent OS 7.0 - Máy chủ ảo chạy ứng dụng: Linux Cent OS 7.0 - Máy chủ thật host các máy ảo: Cent OS 6.5
2	*Tầng trung gian*	- Hệ quản trị CSDL: PostgreSQL 9.4 9.1 - Máy chủ Web: Apache Web Server 2
3	*Phần mềm*	- R 3.2.3 - PHP 5.5 - PHP Yii2 Framework 2.0 - Microsoft Office 2007 - Adobe Flash Player 10+ - Google Chrome 50.0.2661.87

4	*Phần cứng*	- Máy chủ CSDL: Intel Xeon 4 core 2.5GHz+/8GB RAM - Máy chủ ảo chạy ứng dụng: Intel Xeon 4 core - 2.5GHz+/8GB - Máy chủ thật host các máy ảo: Intel Xeon 80 core - 2.5GHz+/128GB
5	*Đường truyền mạng LAN, Internet*	- Tốc độ mạng LAN: 1 Gigabit/giây. - Tốc độ mạng Internet: 8 Megabit/giây.

4.8 Phân tích thiết kế ca sử dụng

Trong phần này, danh sách các use-case mà tôi xây dựng trong hệ thống EnvPro sẽ được mô tả. Các nhóm chức năng này là 2 nhóm chức năng chính được xây dựng nhằm hỗ trợ xử lý đối với dữ liệu quan trắc bị nhiễu và thiếu. Các nhóm chức năng này nằm về phía Client của hệ thống cho 2 đối tượng chính sử dụng là người dùng thông thường và nhà nghiên cứu.



Hình 19. Biểu đồ User-case tổng quát hai nhóm chức năng xử lý dữ liệu nhiễu và thiếu của hệ thống EnvPro

4.8.1 Nhóm chức năng xử lý dữ liệu nhiễu

Nhóm chức năng này cho phép người dùng phát hiện và loại bỏ những giá trị quan trắc sai lệch dựa vào việc phân tích tương quan giá trị quan trắc của ngày và

tháng hoặc thông qua khoảng giá trị cho phép mà người dùng mong muốn với bất kì chỉ tiêu quan trắc nào.

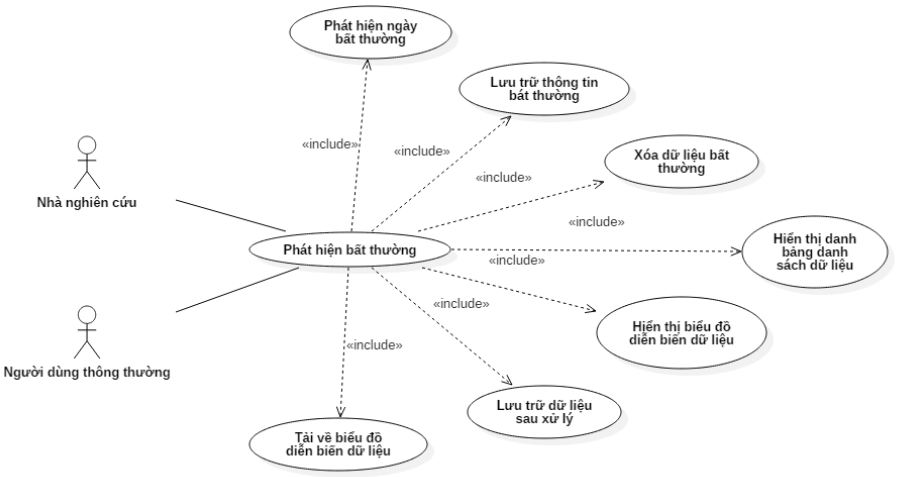
Nhóm các chức năng bắt đầu khi người dùng nhập vào file dữ liệu quan trắc hoặc tìm kiếm dữ liệu quan trắc từ hệ thống cần xử lý nhiễu và kết thúc khi chỉ tiêu quan trắc mong muốn được xử lý nhiễu và hiển thị biểu đồ diễn biến chi tiết những giá trị nhiễu được lược bỏ lên màn hình.

4.8.1.1 Chức năng phát hiện bất thường

Mô tả chung.

Chức năng cho phép phát hiện, loại bỏ, lưu trữ dữ liệu nhiễu và hiển thị biểu đồ diễn biến biến dữ liệu quan trắc trước và sau quá trình xử lý

Biểu đồ ca sử dụng.



Hình 20. Biểu đô use-case chức năng phát hiện bất thường

Luồng cơ bản

Người dùng click chọn chức năng phát hiện bất thường. Hệ thống hiển thị màn hình truyền vào file dữ liệu cần xử lý hoặc tìm kiếm dữ liệu ngay trên hệ thống.

Người dùng có thể tìm kiếm dữ liệu từ hệ thống hoặc truyền vào file dữ liệu quan trắc với định dạng là .CSV cần xử lý.

Hệ thống tự động load dữ liệu và hiển thị danh sách các chỉ tiêu quan trắc nhận được vào dropdown list.

Người dùng click chọn dropdown list và chọn chỉ tiêu quan trắc cần xử lý. Click chọn Xử lý để phát hiện dữ liệu nhiễu

Hệ thống tự động gọi tới công cụ R với những thông tin dữ liệu được truyền vào và tự động lấy kết quả trả về hiện lên màn hình là danh sách những ngày phát hiện được có hệ số tương quan thấp.

Với những ngày phát hiện được, người dùng có thể chọn chức năng ghi chú để ghi lại thông tin về những ngày bất thường đó và lưu lại trên máy tính cá nhân. Với những ngày phát hiện là bất thường người dùng check chọn vào những ngày mong muốn hoặc chọn Check All và click Xóa để xóa dữ liệu thuộc những ngày này trên bộ dữ liệu đầu vào.

Hệ thống nhận được thông tin xóa dữ liệu tiến hành xóa tự động và hiển thị bảng kết quả cùng biểu đồ mô tả diễn biến số liệu trước và sau khi xóa lên màn hình.

Tại đây nếu thấy kết quả đã đúng như yêu cầu, người dùng chọn chức năng Tải Về để lưu lại file dữ liệu sau khi xử lý. Nếu không đúng như yêu cầu mong muốn người dùng chọn button Quay lại để tiến hành xử lý lại dữ liệu. Người dùng chọn Kết thúc để hoàn thành chức năng. Hệ thống tự động đưa người dùng đến trang hiển thị các chức năng hệ thống.

Luồng rẽ nhánh

- Lỗi truy xuất cơ sở dữ liệu
- Lỗi truyền vào file dữ liệu sai về định dạng và cấu trúc
- Lỗi về hiển thị biểu đồ diễn biến chỉ tiêu quan trắc mong muốn trước và sau quá trình xử lý.

Tiền điều kiện

- Đối với đối tượng người dùng là Nhà nghiên cứu thì cần phải login vào hệ thống trước khi sử dụng để có thể tìm kiếm được dữ liệu.
- Hệ thống được kết nối cơ sở dữ liệu hoặc truyền vào file dữ liệu quan trắc đúng theo yêu cầu để xử lý.
- Hệ thống phải kết ngối được với công cụ hỗ trợ xử lý R, để tiến hành xử lý nhiễu

Hậu điều kiện

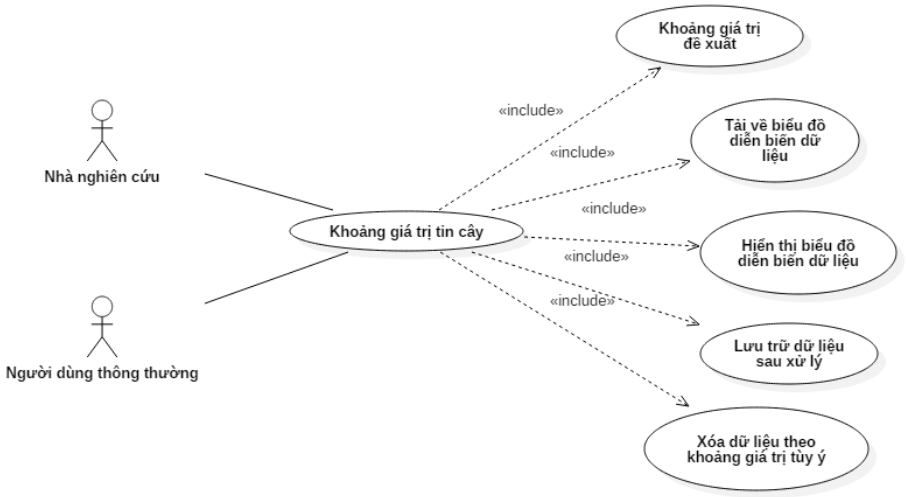
- Nếu use-case thành công, danh sách những ngày phát hiện được là bất thường được hiển thị lên màn hình. Dữ liệu sau khi xử lý bất thường được hiển thị thông qua biểu đồ trực quan và cho phép người dùng tải về.
- Nếu use-case không thành công, hệ thống sẽ thông báo lỗi tới màn hình cho người dùng nắm được thông tin.

4.8.1.2 Chức năng loại bỏ giá trị không phù hợp theo khoảng tin cậy

Mô tả chung.

Chức năng cho phép phát hiện, loại bỏ, lưu trữ dữ liệu nằm ngoài khoảng giá trị mà người dùng mong muốn và hiển thị biểu đồ diễn biến biến dữ liệu quan trắc trước và sau quá trình xử lý

Biểu đồ ca sử dụng.



Hình 21. Biểu đồ use-case loại bỏ giá trị nhiễu dựa vào khoảng tin cậy

Luồng cơ bản

Người dùng click chọn chức năng Khoảng tin cậy. Hệ thống hiển thị màn hình truyền vào file dữ liệu cần xử lý hoặc tìm kiếm dữ liệu ngay trên hệ thống.

Người dùng có thể tìm kiếm dữ liệu từ hệ thống hoặc truyền vào file dữ liệu quan trắc với định dạng là .CSV cần xử lý.

Hệ thống tự động load dữ liệu và hiển thị danh sách các chỉ tiêu quan trắc nhận được vào dropdown list và 2 textbox cho phép người dùng nhập vào giá trị Min và Max.

Người dùng click chọn dropdown list và chọn chỉ tiêu quan trắc cần xử lý. Nhập vào khoảng giá trị Min, Max. Với một số chỉ tiêu quan trắc hệ thống sẽ tự động đề xuất khoảng giá trị phù hợp. Nếu người dùng nhận thấy khoảng giá trị đó là ổn sẽ chọn radio button Sử dụng khoảng dư liệu mặc định. Sau đó click chọn Xử lý để phát hiện và loại bỏ dữ liệu.

Hệ thống tự động gọi tới công cụ R với những thông tin dữ liệu được truyền vào và tự động lấy kết quả trả về hiện lên màn hình là danh sách là bộ dữ liệu sau khi được loại bỏ các giá trị không phù hợp và hiển thị biểu đồ mô tả diễn biến số liệu trước và sau khi xóa dữ liệu lên màn hình.

Tại đây nếu thấy kết quả đã đúng như yêu cầu, người dùng chọn chức năng Tải Về để lưu lại file dữ liệu sau khi xử lý. Nếu không đúng như yêu cầu mong muốn người dùng chọn button Quay lại để tiến hành xử lý lại dữ liệu. Người dùng chọn Kết thúc để hoàn thành chức năng. Hệ thống tự động đưa người dùng đến trang hiển thị các chức năng hệ thống.

Luồng rẽ nhánh

- Lỗi truy xuất cơ sở dữ liệu
- Lỗi truyền vào file dữ liệu sai về định dạng và cấu trúc
- Lỗi về hiển thị biểu đồ diễn biến chỉ tiêu quan trắc mong muốn trước và sau quá trình xử lý.

Tiền điều kiện

- Đối với đối tượng người dùng là Nhà nghiên cứu thì cần phải login vào hệ thống trước khi sử dụng để có thể tìm kiếm được dữ liệu.
- Hệ thống được kết nối cơ sở dữ liệu hoặc truyền vào file dữ liệu quan trắc đúng theo yêu cầu để xử lý.
- Hệ thống phải kết ngối được với công cụ hỗ trợ xử lý R, để tiến hành xử lý nhiễu

Hậu điều kiện

- Nếu use-case thành công, bảng số liệu trước và sau quá trình xử lý của chỉ tiêu quan trắc được được thị lên màn hình. Dữ liệu sau khi xử lý được hiển thị thông qua biểu đồ trực quan và cho phép người dùng tải về.
- Nếu use-case không thành công, hệ thống sẽ thông báo lỗi tới màn hình cho người dùng nắm được thông tin.

4.8.2 Nhóm chức năng xử lý dữ liệu thiếu

Nhóm chức năng này cho phép người dùng phát hiện và điền giá trị cho những bản ghi thiếu của giá trị quan trắc chỉ tiêu mong muốn. Người dùng có thể tùy chọn một trong 3 chức năng đó là:

- Sử dụng phép hồi quy tuyến tích măc định: Nghĩa là sử dụng hàm hồi quy từ tập dữ liệu hoạc điền vào bộ dữ liệu cần xử lý
- Sử dụng phép hồi quy tuyến tính với các tham số hồi quy tùy biến: Người dùng tùy ý chọn các chỉ tiêu quan trắc để dựng mô hình hồi quy và điền tham số hồi quy phù hợp cho từng giá trị quan trắc đó, để dựng mô hình hồi quy.
- Sử dụng bộ dữ liệu khác để điền dữ liệu thiếu: Nghĩa là người dùng có thể chọn các trạm quan trắc tương đồng trong khu vực cần đánh giá hoặc lấy dữ liệu từ những năm trước…. để lấy giá trị điền vào những bản ghi thiếu dữ liệu.

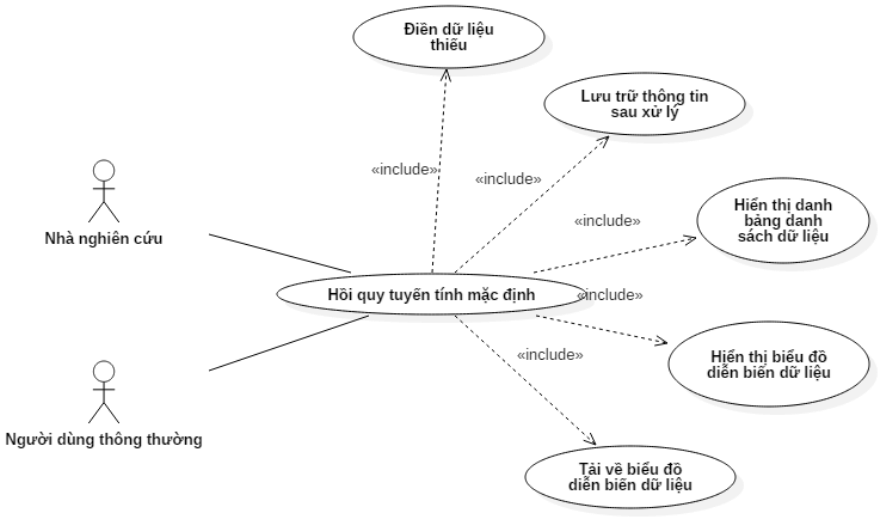
Nhóm các chức năng bắt đầu khi người dùng nhập vào file dữ liệu quan trắc hoặc tìm kiếm dữ liệu quan trắc từ hệ thống cần xử lý nhiễu. Hệ thống tự động chạy xử lý đối với từng chức năng khác nhau và kết thúc khi chỉ tiêu quan trắc mong muốn được xử lý thiếu và hiển thị biểu đồ diễn biến chi tiết chỉ tiêu quan trắc trước và sau khi xử lý lên màn hình.

4.8.2.1 Xử lý dữ liệu thiếu sử dụng phép hồi quy tuyến tính mặc định

Mô tả chung.

Chức năng cho phép phát hiện và điền vào những giá trị tiên đoán phù hợp cho một chỉ tiêu quan trắc mà người dùng mong muốn dựa trên phương pháp hồi quy tuyến tính. Phương pháp này được xây dựng tự động từ nhưng chỉ tiêu quan trắc mà người dùng lựa chọn thông qua bộ dữ liệu học.

Biểu đồ ca sử dụng.



Hình 22. Biểu đô use-case diền dữ liệu thiếu dự vào phương tình hồi quy tuyến tính tự động

Luồng cơ bản

Người dùng click chọn chức năng Điền dữ liệu thiếu. Hệ thống hiển thị màn hình truyền vào file dữ liệu cần xử lý hoặc tìm kiếm dữ liệu ngay trên hệ thống.

Người dùng có thể tìm kiếm dữ liệu từ hệ thống hoặc truyền vào file dữ liệu quan trắc với định dạng là .CSV cần xử lý.

Hệ thống tự động load dữ liệu và hiển thị danh sách các chỉ tiêu quan trắc nhận được từ dánh ách dữ liệu cần xử lý vào dropdown list

Người dùng click chọn dropdown list và chọn chỉ tiêu quan trắc cần xử lý. Sau đó người dùng truyền vào file dữ liệu học.

Hệ thống tự động load file dữ liệu học và hiển thị danh sách các chỉ tiêu quan trắc nhận được từ dữ liệu học vào dropdown list

Người dùng lựa chọn một hoặc nhiều chỉ tiêu quan trắc từ bộ dữ liệu học để xây dựng mô hình hồi quy tuyến tính tự động và click chọn Xử lý

Hệ thống tự động gọi tới công cụ R với những thông tin dữ liệu được truyền vào và tự động lấy kết quả trả về hiện lên màn hình là danh sách là bộ dữ liệu với chỉ tiêu quan trắc sau khi được điền dữ liệu thiếu đồng thời hiển thị biểu đồ mô tả diễn biến số liệu trước và sau khi điền dữ liệu thiếu lên màn hình.

Tại đây nếu thấy kết quả đã đúng như yêu cầu, người dùng chọn chức năng Tải Về để lưu lại file dữ liệu sau khi xử lý. Nếu không đúng như yêu cầu mong muốn người dùng chọn button Quay lại để lựa chọn lại các chỉ tiêu xây dựng mô hình hồi quy. Người dùng chọn Kết thúc để hoàn thành chức năng. Hệ thống tự động đưa người dùng đến trang hiển thị các chức năng hệ thống.

Luồng rẽ nhánh

- Lỗi truy xuất cơ sở dữ liệu
- Lỗi truyền vào file dữ liệu xử lý, file dữ liệu học sai về định dạng và cấu trúc
- Lỗi về hiển thị biểu đồ diễn biến chỉ tiêu quan trắc mong muốn trước và sau quá trình xử lý.

Tiền điều kiện

- Đối với đối tượng người dùng là Nhà nghiên cứu thì cần phải login vào hệ thống trước khi sử dụng để có thể tìm kiếm được dữ liệu.
- Hệ thống được kết nối cơ sở dữ liệu hoặc truyền vào file dữ liệu quan trắc đúng theo yêu cầu để xử lý.
- Hệ thống phải kết ngối được với công cụ hỗ trợ xử lý R, để tiến hành xử lý nhiễu

Hậu điều kiện

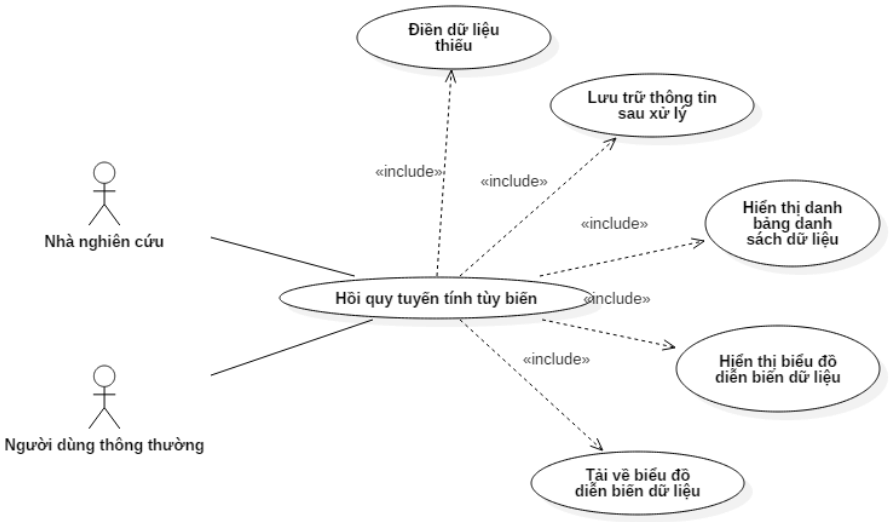
- Nếu use-case thành công, bảng số liệu trước và sau quá trình xử lý của chỉ tiêu quan trắc được được thị lên màn hình. Dữ liệu sau khi xử lý được hiển thị thông qua biểu đồ trực quan và cho phép người dùng tải về.
- Nếu use-case không thành công, hệ thống sẽ thông báo lỗi tới màn hình cho người dùng nắm được thông tin.

4.8.2.2 Xử lý dữ liệu thiếu sử dụng hồi quy tuyến tính với tham số tùy ý

Mô tả chung.

Chức năng cho phép phát hiện và điền vào những giá trị tiên đoán phù hợp cho một chỉ tiêu quan trắc mà người dùng mong muốn dựa trên phương pháp hồi quy tuyến tính. Phương pháp này được xây dựng tự động bằng cách yêu cầu người dùng nhập vào danh sách chỉ tiêu quan trắc cùng các tham số hồi quy tương ứng với từng chỉ tiêu quan trắc đó.

Biểu đồ ca sử dụng.



Hình 23. Biểu đồ use-case điền dữ liệu thiếu dựa vào phép hồi quy tuyến tính tùy biến.

Luồng cơ bản

Người dùng click chọn chức năng Điền dữ liệu thiếu. Hệ thống hiển thị màn hình truyền vào file dữ liệu cần xử lý hoặc tìm kiếm dữ liệu ngay trên hệ thống.

Người dùng có thể tìm kiếm dữ liệu từ hệ thống hoặc truyền vào file dữ liệu quan trắc với định dạng là .CSV cần xử lý.

Hệ thống tự động load dữ liệu và hiển thị danh sách các chỉ tiêu quan trắc nhận được từ dánh ách dữ liệu cần xử lý vào dropdown list

Người dùng click chọn dropdown list và chọn chỉ tiêu quan trắc cần xử lý.

Hệ thống hiển thị dropdownlist danh sách các chỉ tiêu quan trắc thu được từ tập dữ liệu và textbox tương ứng cho phép người dùng nhập vào các tham số để xây dựng hàm hồi quy. Người dùng có thể xây dựng hàm hồi quy với một hoặc nhiều tham số bằng cách lựa chọn button Add hoặc Remove để loại bỏ. Sau khi lựa chọn thông tin đầy đủ, người dùng click chọn Xử lý

Hệ thống tự động gọi tới công cụ R với những thông tin dữ liệu được truyền vào và tự động lấy kết quả trả về hiện lên màn hình là danh sách là bộ dữ liệu với chỉ tiêu quan trắc sau khi được điền dữ liệu thiếu đồng thời hiển thị biểu đồ mô tả diễn biến số liệu trước và sau khi điền dữ liệu thiếu lên màn hình.

Tại đây nếu thấy kết quả đã đúng như yêu cầu mong muốn, người dùng chọn chức năng Tải Về để lưu lại file dữ liệu sau khi xử lý. Nếu không đúng như yêu cầu

mong muốn người dùng chọn button Quay lại để lựa chọn lại các chỉ tiêu xây dựng mô hình hồi quy. Người dùng chọn Kết thúc để hoàn thành chức năng. Hệ thống tự động đưa người dùng đến trang hiển thị các chức năng hệ thống.

Luồng rẽ nhánh

- Lỗi truy xuất cơ sở dữ liệu
- Lỗi truyền vào file dữ liệu xử lý sai về định dạng và cấu trúc
- Lỗi về hiển thị biểu đồ diễn biến chỉ tiêu quan trắc mong muốn trước và sau quá trình xử lý.

Tiền điều kiện

- Đối với đối tượng người dùng là Nhà nghiên cứu thì cần phải login vào hệ thống trước khi sử dụng để có thể tìm kiếm được dữ liệu.
- Hệ thống được kết nối cơ sở dữ liệu hoặc truyền vào file dữ liệu quan trắc đúng theo yêu cầu để xử lý.
- Hệ thống phải kết ngối được với công cụ hỗ trợ xử lý R, để tiến hành xử lý nhiễu

Hậu điều kiện

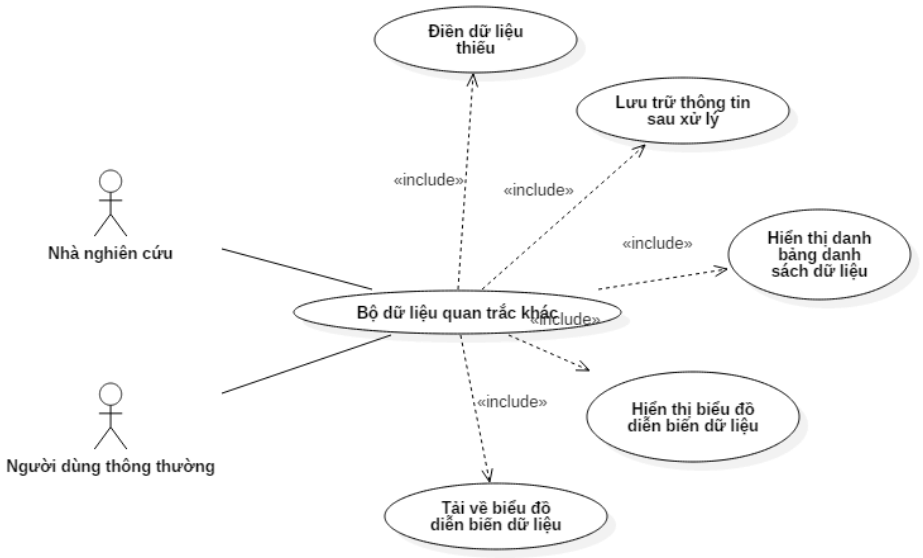
- Nếu use-case thành công, bảng số liệu trước và sau quá trình xử lý của chỉ tiêu quan trắc được được thị lên màn hình. Dữ liệu sau khi xử lý được hiển thị thông qua biểu đồ trực quan và cho phép người dùng tải về.
- Nếu use-case không thành công, hệ thống sẽ thông báo lỗi tới màn hình cho người dùng nắm được thông tin.

4.8.2.3 Xử lý dữ liệu thiếu từ bộ dữ liệu quan trắc khác

Mô tả chung.

Chức năng cho phép người dùng tích hợp dữ liệu giữa bộ dữ liệu cần xử lý (A) và bộ dữ liệu quan trắc khác (B). Sau quá trình xử lý những chỉ tiêu quan trắc có bản ghi thiếu dữ liệu của A được tự đồng điền từ bộ dữ liệu B tương ứng đồng thời về thời gian quan trắc. Chức năng kết thúc khi bộ dữ liệu A được điền dữ liệu từ B và hiển thị danh sách dữ liệu cùng biểu đồ trực quan trước và sau quá trình xử lý lên màn hình.

Biểu đồ ca sử dụng.



Hình 24. Biểu đô use-case điền dữ liệu thiếu dựa vào bộ dữ liệu quan trắc môi trường khác

Luồng cơ bản

Người dùng click chọn chức năng Điền dữ liệu thiếu. Hệ thống hiển thị màn hình truyền vào file dữ liệu cần xử lý hoặc tìm kiếm dữ liệu ngay trên hệ thống.

Người dùng có thể tìm kiếm dữ liệu từ hệ thống hoặc truyền vào file dữ liệu quan trắc A với định dạng là .CSV cần xử lý. Đồng thời truyền file dữ liệu quan trắc B vào form yêu cầu.

Hệ thống tự động load dữ liệu và hiển thị danh sách các chỉ tiêu quan trắc có mặt đồng thời nhận được từ 2 tập dữ liệu A và B vào dropdown list

Người dùng click chọn dropdown list và chọn chỉ tiêu quan trắc cần xử lý. Click chọn Xử lý.

Hệ thống tự động gọi tới công cụ R với những thông tin dữ liệu được truyền vào và tự động lấy kết quả trả về hiện lên màn hình là danh sách là bộ dữ liệu với chỉ tiêu quan trắc sau khi được điền dữ liệu thiếu đồng thời hiển thị biểu đồ mô tả diễn biến số liệu trước và sau khi điền dữ liệu thiếu lên màn hình.

Tại đây nếu thấy kết quả đã đúng như yêu cầu mong muốn, người dùng chọn chức năng Tải Về để lưu lại file dữ liệu sau khi xử lý. Nếu không đúng như yêu cầu mong muốn người dùng chọn button Quay lại để lựa chọn lại dữ liệu quan trắc B khác. Người dùng chọn Kết thúc để hoàn thành chức năng. Hệ thống tự động đưa người dùng đến trang hiển thị các chức năng hệ thống.

Luồng rẽ nhánh

- Lỗi truy xuất cơ sở dữ liệu

- Lỗi truyền vào file dữ liệu xử lý A và B sai về định dạng và cấu trúc
- Lỗi về hiển thị biểu đồ diễn biến chỉ tiêu quan trắc mong muốn trước và sau quá trình xử lý.

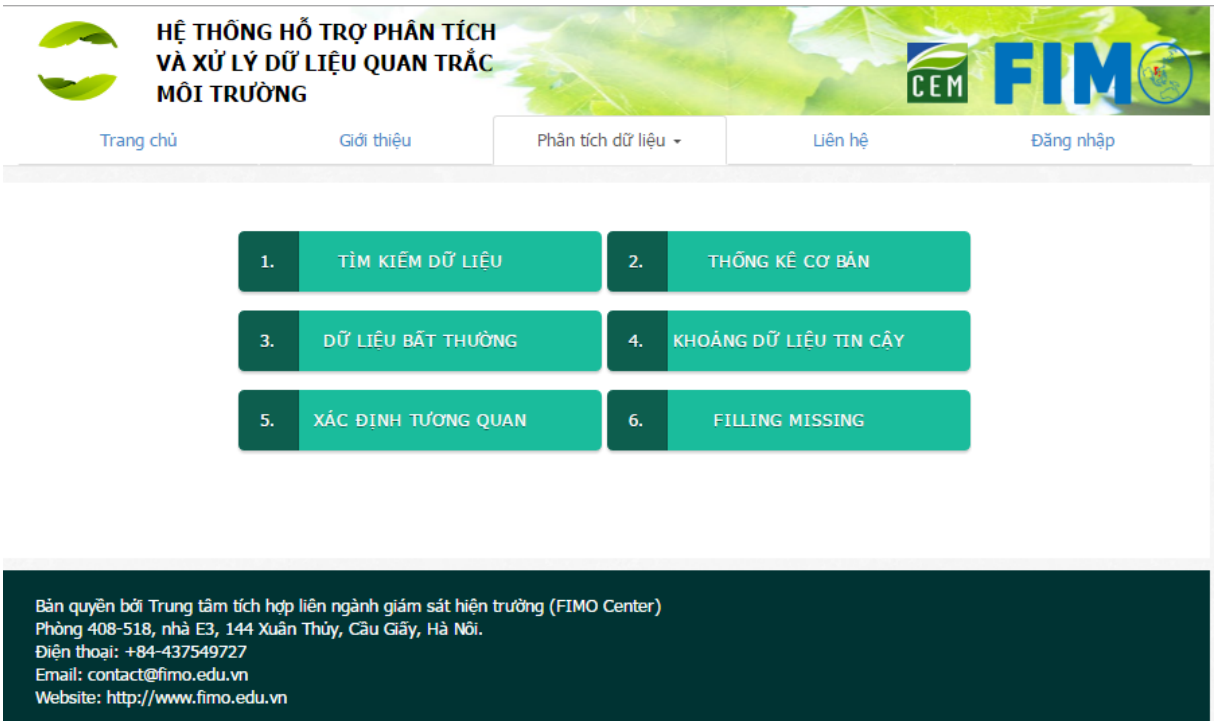
Tiền điều kiện

- Đối với đối tượng người dùng là Nhà nghiên cứu thì cần phải login vào hệ thống trước khi sử dụng để có thể tìm kiếm được dữ liệu.
- Hệ thống được kết nối cơ sở dữ liệu hoặc truyền vào file dữ liệu quan trắc đúng theo yêu cầu để xử lý.
- Hệ thống phải kết ngối được với công cụ hỗ trợ xử lý R, để tiến hành xử lý nhiễu

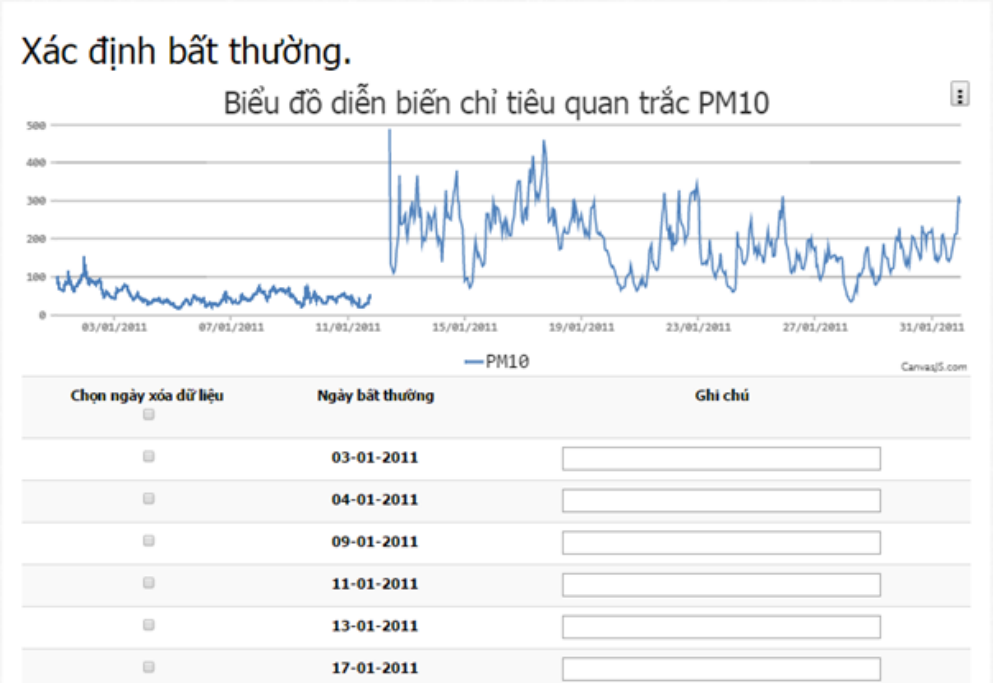
Hậu điều kiện

- Nếu use-case thành công, bảng số liệu trước và sau quá trình xử lý của chỉ tiêu quan trắc được được thị lên màn hình. Dữ liệu sau khi xử lý được hiển thị thông qua biểu đồ trực quan và cho phép người dùng tải về.
- Nếu use-case không thành công, hệ thống sẽ thông báo lỗi tới màn hình cho người dùng nắm được thông tin.

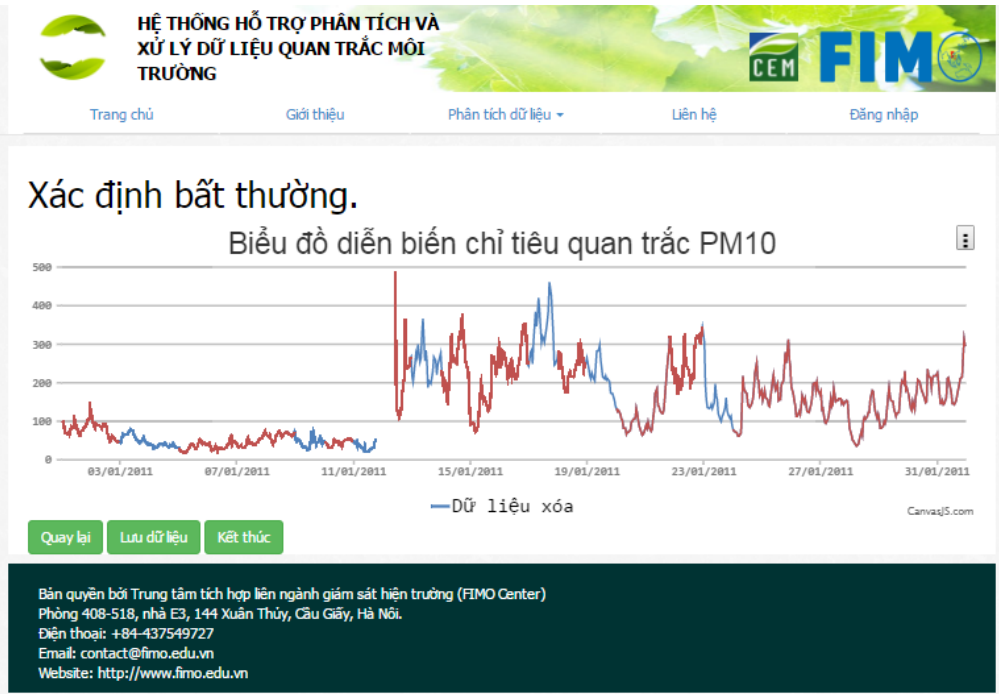
4.9 Kết quả đạt được



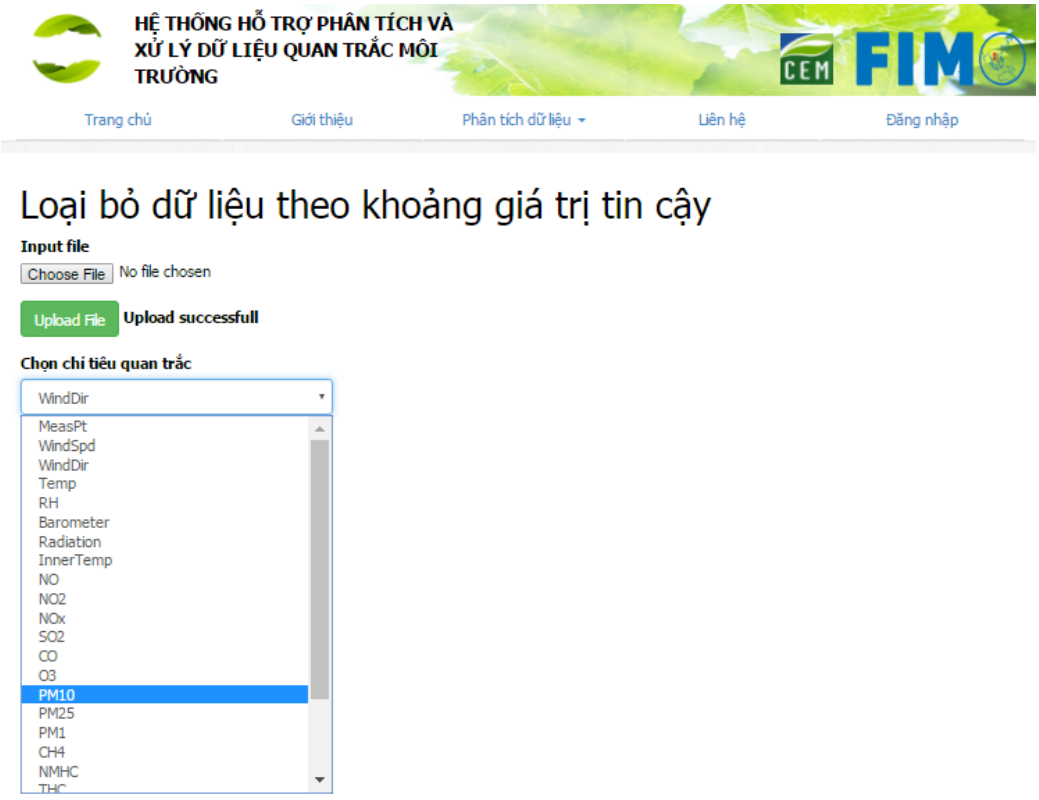
Hình 25. Giao diện tổng quan hệ thống.



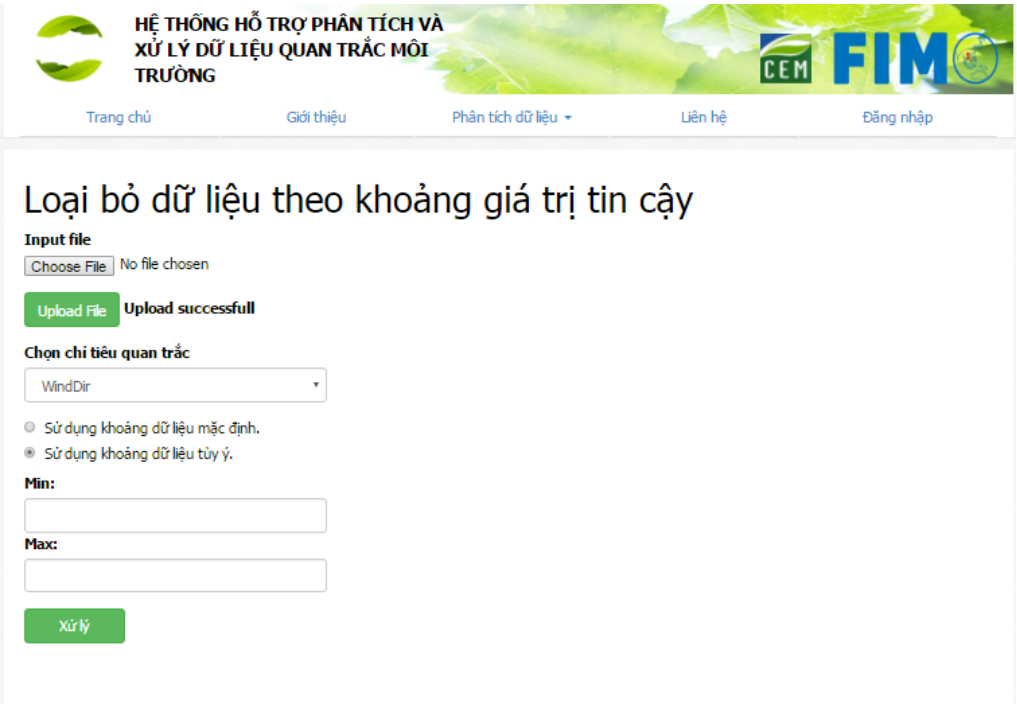
Hình 26. Giao diện kết quả xác định dữ liệu bất thường



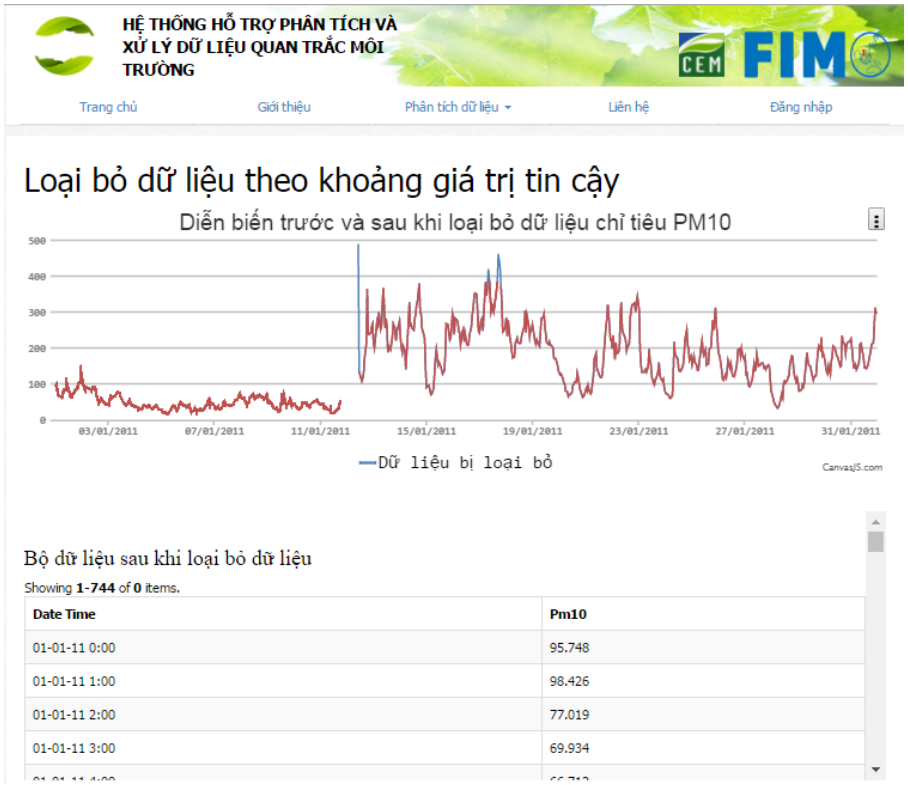
Hình 27. Giao diện biểu đồ diễn biến PM10 sau khi xử lý dữ liệu bất thường



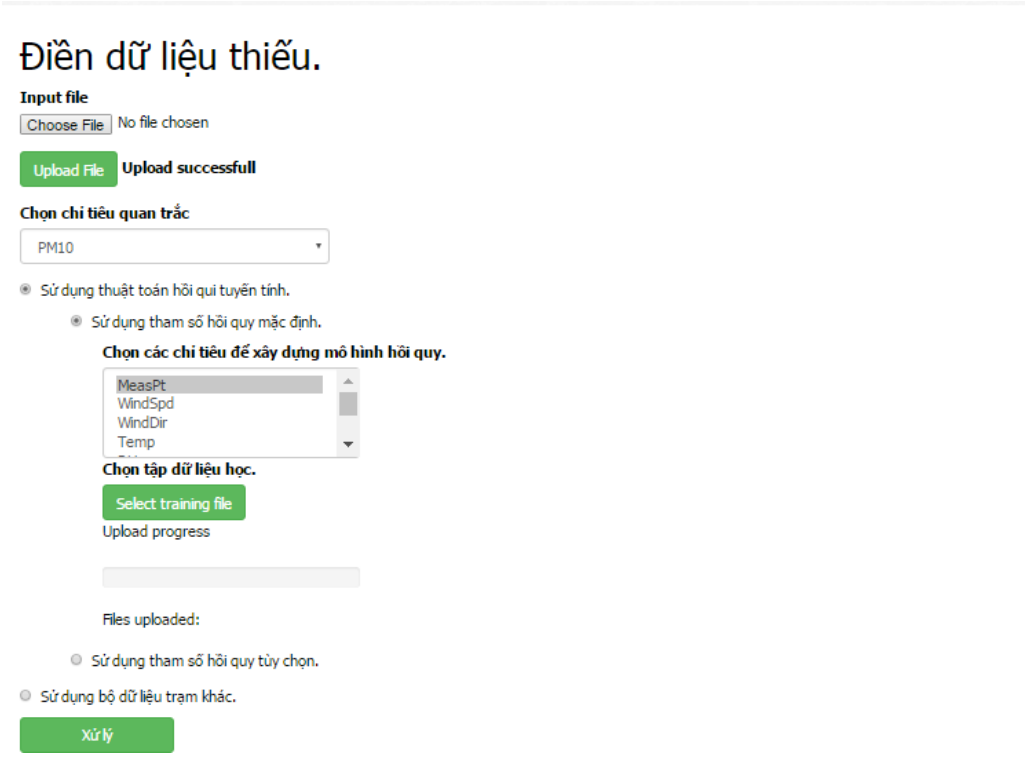
Hình 28. Giao diện hệ thống hiển thị danh sách các chỉ tiêu quan trắc hiện cho cho người dùng lựa chọn



Hình 29. Giao diện chức năng loại bỏ dữ liệu theo khoảng giá trị tin cậy



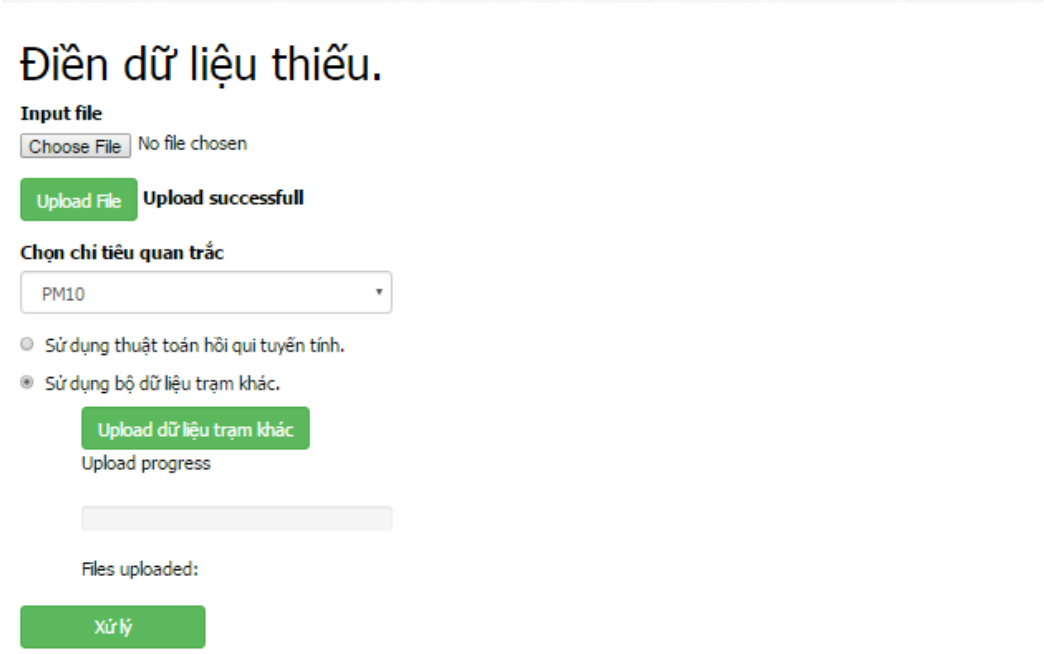
Hình 30. Giao diện kết quả bước loại bỏ dữ liệu nhiễu theo khoảng giá trị tin cậy



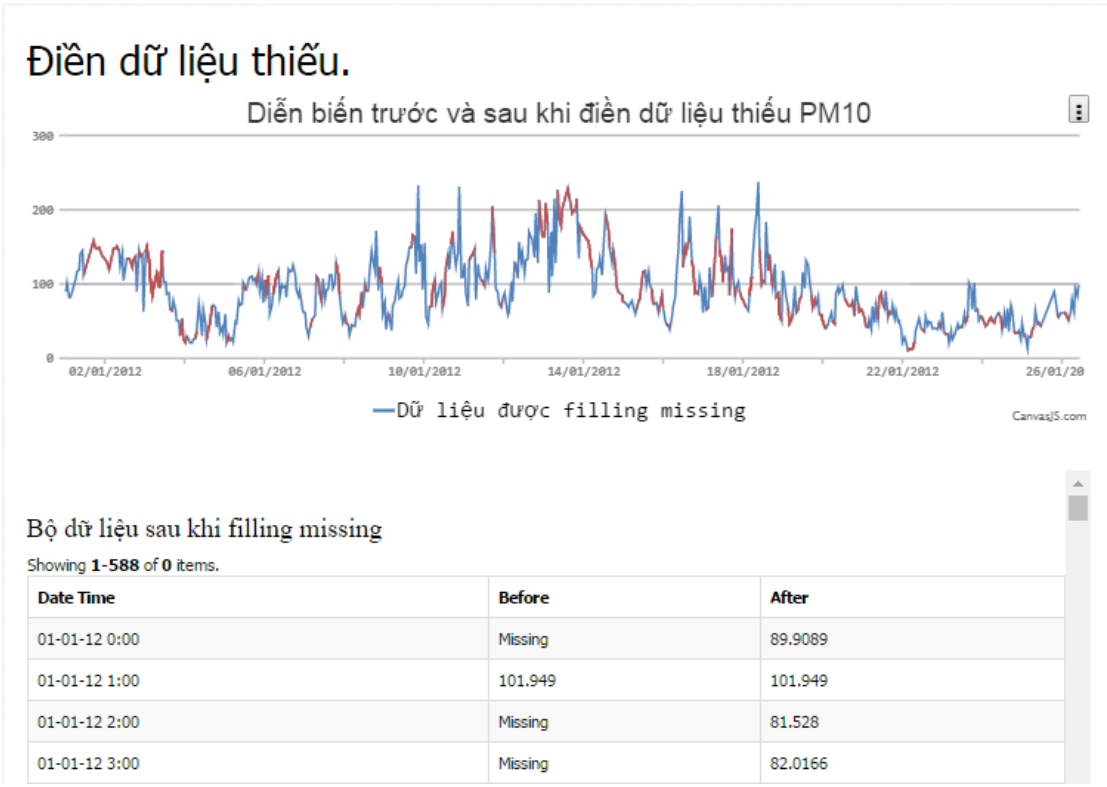
Hình 31. Giao diện điền dữ liệu thiếu sử dụng phép hồi quy tuyến tính tự động

Điền dữ liệu thiếu.
Input file
Choose File
No file chosen
Upload File
Upload successfull
Chọn chỉ tiêu quan trắc
PM10
Sử dụng thuật toán hồi qui tuyến tính.
Sử dụng tham số hồi quy mặc định.
Sử dụng tham số hồi quy tùy chọn.
Hàm hồi quy Y =
Hệ số hồi qui :
+
Hệ số hồi qui :
Chỉ tiêu 1 :
MeasPt
+
Add
Remove
Sử dụng bộ dữ liệu trạm khác.
Xử lý

Hình 32. Giao diện điền dữ liệu thiếu sử dụng phép hồi quy tuyến tính tùy biến



Hình 33. Giao diện điền dữ liệu thiếu sử dụng bộ dữ liệu quan trức khác



Hình 34. Giao diện kết quả điền dữ liệu thiếu được hiển thị ra màn hình.

KẾT LUẬN VÀ ĐỊNH HƯỚNG

Luận văn nêu lại những mảng kiến thức tổng quan của bài toán xử lý dữ liệu, các hướng tiếp cận, phương pháp giải quyết, ứng dụng và đánh giá… Với hướng nghiên cứu kết hợp giữa phân tích và tiền xử lý dữ liệu quan trắc môi trường từ đó đề xuất ra quy trình mục đích và quy trình chuẩn hóa dữ liệu quan trắc môi trường tại Việt Nam. Cụ thể là hỗ trợ người dùng xây dựng một tập dữ liệu quan trắc tổng hợp từ các file dữ liệu quan trắc theo ngày riêng lẻ. Bộ dữ liệu này đảm bảo các chuẩn qui ước về cấu trúc, định dạng, đơn vị đo. Khía cạnh thứ 2 đó là chuẩn về chất lượng dữ liệu với những dữ liệu nhiễu được phát hiện và loại bỏ, dữ liệu thiếu thì được điền một cách logic hợp lý dựa trên phương pháp hồi quy tuyến tính.

Cũng từ những đặc điểm của dữ liệu thu được mà luận văn quyết định áp dụng các hướng phân tích tương quan và hồi quy để xử lý dữ liệu nhiễu và thiếu thay vì một số phương pháp mang tính chủ quan dựa trên kinh nghiệm người dùng mà không có cơ sở để xác định với độ tin cậy cao.

Để đánh giá kết quả của quy trình đề xuất. trong phạm vi luận văn này tôi sử dụng hai bộ dữ liệu quan trắc tháng 01/2011 và 01/2012 của trạm Nguyễn Văn Cừ, Hà Nội. và đạt được một số kết quả khả quan làm tiền đề để có thể phát triển hoặc kết hợp thêm các thuật toán khác để cải tiến qui trình trong tương lại.

Mặc dù bài toan được xây dựng dựa trên một qui trình tổng thể nhưng đối với mỗi bước nhỏ trong qui trình, nhà phân tích hoàn toàn có thẻ tùy biến sử dụng qua lại giữa các chức năng tùy thuộc vào yêu cầu bài toán cần xử lý mà không bắt buộc phải chạy lần lượt từng bước nhỏ của quy trình từ đầu đến cuối. Đây cũng là một trong những nguyên nhân đáng kể tới khi mô hình hóa quy trình từ lý thuyết để xây dựng nên hệ thống hỗ trợ xử lý dữ liệu quan trắc môi trường EnvPro. Từ đó hể hiện rõ hơn tính động và khả năng tuy biến với những chỉ tiêu quan trắc môi trường khác nhau hay kết hợp giữa các bước xử lý khác nhau theo yêu cầu người dùng.

Tóm gọn lại với đề tài luận văn được thực hiện từ những bước nghiên cứu tổng quan cho tới đề xuất phương pháp, thực nghiệm và xây dựng hệ thống trực quan mà một quá trình đòi hỏi phải nắm rõ về dữ liệu cũng như mục đích, yêu cầu dữ liệu mong muốn, thêm vào đó đòi hỏi phải có thời gian nghiên cứu dữ liệu lâu dài. Với khả năng nghiên cứu có hạn, trong thời gian thực hiện uận văn này đã tiến hành nghiên cứu và giải quyết được các vấn đề sau:

- Tìm hiểu tổng quan về ô nhiễm không khí. Đánh giả ảnh hưởng, tác động và nguồn phát gây nên hiện tượng ô nhiễm không khí. Từ đó liên hệ tới hiện trạng ô nhiễm không khí tại Việt Nam.
- Tìm hiểu các ký thuật xử lý dữ liệu bao gồm các bước như đánh giá độ phân tán , độ tập trung dữ liệu, xử lý dữ liệu thiếu, xử lý dữ liệu nhiễu, phép phân tích tương quan, hôi quy tuyến tính… Từ đó đề xuất ra qui trình chuẩn hóa dữ liệu quan trắc môi trường tại Việt Nam.
- Sử dụng ngôn ngữ thống kê R để hỗ trợ xử lý dữ liệu quan trắc môi trường
- Tìm hiểu và đề xuất những tiêu chuẩn về cấu trúc dữ liệu, đơn vị đo, định dạng dữ liệu quan trắc.
- Đề xuất khoảng dữ liệu tin cậy cho chỉ tiêu quan trắc PM10 thông qua các báo cáo môi trường và các công trình nghiê cứu đã thực hiện.
- Nắm được các phương pháp đánh giá, thử nghiệm dữ liệu nhằm xác định được tính chính xác, độ tin cậy của phương pháp đề xuất
- Nghiên cứu các ngôn ngữ lập trình như PHP Yii 2.0 Framework, Jquery và PostgreSQL. Từ đó làm nền tảng xây dựng hệ thống hỗ trợ xử lý dữ liệu quan trắc môi trường EnvPro tại Việt Nam

Bên cạnh những kiến thức hữu ích đạt được từ luận văn tôi cũng có những đánh giá, nhận xét về những nhược điểm mà trong luận văn này vẫn chưa giải quyết được và những yêu cầu thực tê khách quan đối với quy trình cũng như hệ thống. Những vấn đề này sẽ là những hướng phát triển trong tương lai đảm bảo cho kết quả như người dùng mong muốn.

Về quy trình chuẩn hóa dữ liệu quan trắc môi trường:

- Về bước xử lý dữ liệu nhiễu: ngoài cách phát hiện bất thường sử dụng phép phân tích tương quan thì có thể sử dụng các quy chuẩn về chất lượng không khí đê phát hiện bất thường như quy chuẩn QCVN05 2013 của Việt Nam. Cũng giống như việc phát hiện bất thường từ việc so sánh tương quan ngày và tháng. Người dùng có thể đánh giá chi tiết những ngày vượt chuẩn để đưa ra quyết định xử lý.
- Với bước xử lý dữ liệu thiếu: Nếu sử dụng mô hình hồi quy truyến tính thì dữ liệu thiếu chỉ điền được khi các giá trị dùng để xây dựng mô hình hồi quy có dữ liệu. Vậy với những trường hợp mà thời điểm quan trắc thiếu toàn bộ giá trị quan trắc của tất cả các chỉ tiêu thì không thể xử lý được. Có một số chuyên gia vê môi trường đề xuất là sử dụng các trạm quan trắc trong cùng khu vực khí tượng, thời tiết để chao đổi dữ nhiệu cho những

bản ghi thiếu. Xét trên điều kiện các trạm quan trắc thực tế ở Việt Nam phương pháp pháp này có thể áp dụng được để nâng cao chất lượng bộ dữ liệu.

Về hệ thống hỗ trợ xử lý dữ liệu quan trắc môi trường:

- Hoàn thiện các chức năng khác của hệ thống trong thời gian tới. Đảm bảo hệ thống hoàn chỉnh đầy đủ các chức năng phù hợp với những yêu cầu đã được phân tích.
- Với khả năng tùy biến trong quá trình xử lý dữ liệu, có thể nói mỗi bước xử lý lại cho kết quả là những bộ dữ liệu khác nhau đòi hỏi người dùng tốn nhiều công sức để lưu trữ và ghi nhớ thông tin chi tiết từng file cụ thể. Thêm vào đo slaf khả năng truy xuất dữ liệu kém nếu không được lưu trữ trực tuyên. Vì vậy tôi thiết nghĩ sẽ tích hợp một đám mây với hệ thống. Cho phép người dùng xử lý và lưu trữ trực tuyến trên đám mây. Ứng với mỗi người dùng sẽ có một tài khoản riêng đẻ lưu trữ file. Từ đó người dùng có thể dễ dàng truy xuất và xử lý dữ liệu ở bất kì đâu.
- Với lượng dữ liệu được thu thập rất lớn và được chuẩn hóa thì sẽ dùng để làm gi? Ở thời điểm hiện tại và tương lai, các công cụ liên quan tới bản đồ số sẽ rất phát triển vì vậy ngài những nhiệm vụ như chỉ đường, xá định vị trí nhà hàng, công ty… thì những thông tin quan trắc, khí tương hay đánh giá chát lượng ô nhiễm không khí có thể được tích hợp và hiển thị đồng thời lên bản đồ tại vị trí mà người dùng đang tìm kiếm.
- Một hướng nghiên cứu khác đó là tích hợp với dữ liệu thời gian thực được quan trắc từ các trạm cảm biến không dây. Hệ thống sẽ tự động phân tích và gửi báo cáo phát hiện dữ liệu bất thường và xử lý dữ liệu thiếu định kì theo yêu cầu của người dùng. Tạo ra nguồn dữ liệu tin cậy và đảm bảo khi đến tay nhà quản lý.

Hy vọng những vấn đề được đề cập trong luận văn từ lớn đến nhỏ, từ tổng quan đến chi tiết, từ cách tiếp cận cho bài toán chuẩn hóa dữ liệu quan trắc môi trường đến đề xuất, đánh giá quy trình hay những vướng mắc khi giải quyết các bước thực hiện và đưa ra cách giải quyết, cách tư duy trong những bài toán thực tế nói chung và bài toán xử lý dữ liệu nói riêng, sẽ góp phần nào đó chứng minh được tính khả thi của đề xuất nghiên cứu này tại Việt Nam

TÀI LIỆU THAM KHẢO

Tiếng Việt.

[1] Luật bảo vệ môi trường Việt Nam, năm 1993.

[2] Báo cáo môi trường quốc gia 2013, Môi trường không khí, Bộ Tài nguyên Môi trường

[3] QCVN 05:2013/BTNMT – Quy chuẩn kỹ thuật quốc gia về chất lượng không khí xung quanh

[4] Bài giảng Môi trường và phát triển, Khoa Môi trường, Trường đại học khao học Huế 2010

[5] Thực trạng ô nhiễm không khí đô thị ở Việt Nam 2011, GS.TSKH. Phạm Ngọc Đăng, Chủ tịch Hội Môi trường Xây dựng Việt Nam, Phó Chủ tịch Hội Bảo vệ Thiên nhiên và Môi trường Việt Nam

[6] Bài giảng khai phá dữ liệu 2011, Trường đại học Hàng Hải Việt Nam, Khoa công nghệ thông tin, Bộ môn Hệ thống thông tin

[7] Chương trình huấn luyện y khoa – Ykhoa.net Training – Nguyễn Văn Tuấn

[8] Đại học công nghệ, ĐHQGHN, Luận văn thạc sĩ, Nghiên cứu xây dựng hệ thống webGIS phục vụ chia sẻ thông tin cảnh báo ô nhiễm không khí, Lê Xuân Thành, 2015

[9] Nguyễn Văn Tuấn, Phân tích số liệu và tạo biểu đồ bằng R- Hướng dãn thực hành

[10] Hiện trạng và quy luật diễn biến của chất lượng không khí ở Hà Nội, Phạm Duy Hiển, 03-2006

[11] Thực trạng ô nhiễm môi trường không khí Hà Nội và kiến nghị nhằm giảm thiểu ô nhiễm, Đặng Mạnh Đoàn,Trần ThịDiệu Hằng, Phan Ban Mai - Viện Khoa học Khí Tượng - ThuỷVăn và Môi Trường

[12] Báo cáo môi trường quốc gia 2010, Bộ tài nguyên môi trường.

Tiếng Anh.

[13] Air Pollution in China: Mapping of Concentrations and Sources Robert, A. Rohde, Richard A. Muller

[14] Measurement of high order Kerr refractive index of major air components: erratum V. Loriot, E. Hertz, O. Faucher, and B. Lavorel, 2010 Optical Society of America, Vol. 18, No. 3 / OPTICS EXPRESS 3011

[15] Ambient Air Quality Monitoring System for a City Using Wireless Gas Sensors Dr. K Karuppasamy, S. Shanthini, S. Shobana, J. Jeevin

Chandrakumar, 6 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099

[16] Urban Air Quality Modelling and Management in Hanoi, Vietnam, PhD Thesis, 2010, Ngo Tho Hung, AARHUS University.

[17] VOL. 9, No. 5, May 2014. ARPN Journal of Agricultural and Biological Science, Impact of rice straw burning methods on soil temperature and microorganism distribution in thepaddy soil ecosystems, Nguyen Song Tung, Nguyen Xuan Cu, Nguyen Xuan Hai

[18] Gadde, B., S. Bonnet, C. Menke, S. Garivait, Air Pollutant Emissions from Rice Straw Open Field Bur ning in India, Thailand and Philippines, 2009,Journal of Environmental Pollution.

[19] Current Situation and Possibilities of Rice Straw Management in Vietnam, Pham Thuy Duong, Higano Yoshiro, University of Tsukuba

[20] Viet Nam: Air Quality Profile 2010 Edition - Clean Air Initiative for Asian Cities (CAI-Asia) Center

[21] Roadside BTEX and other gaseous air pollutants in relation to emission sources - Vo Thi Quynh Truc, Nguyen Thi Kim Oanh

[22] Effects of local, regional meteorology and emission sources on mass and compositions of particulate matter in Hanoi Cao Dung Hai, Nguyen Thi Kim Oanh

[23] Roadside levels and traffic emission rates of PM2.5 and BTEX in Ho Chi Minh City, Vietnam - Nguyen Tran Huong Giang, Nguyen Thi Kim Oanh

[24] New indices for wet scavenging of air pollutants (O3, CO, NO2, SO2, and PM10) by summertime rain Jung-Moon Yoo a, Yu-Ri Lee b, Dongchul Kim c,g,*, Myeong-Jae Jeong d, William R. Stockwell e, Prasun K. Kundu f,g, Soo-Min Oh a, Dong-Bin Shin b, Suk-Jo Lee

[25] Impact of Meteorological Parameters and Gaseous Pollutants on PM2.5 and PM10 Mass Concentrations during 2010 in Xi’an, China

[26] Determination of O3, NO2, SO2, CO and PM10 measured in Belgrade urban area, Dragan M. Marković, Dragan A. Marković, Anka Jovanović, Lazar Lazić, and Zoran Mijić

Website.

[27] http://vnexpress.net/tin-tuc/khoa-hoc/moi-truong/bac-kinh-phat-bao-dong-do-lan-hai-vi-o-nhiem-khong-khi-3329876.html

[28] http://www.theguardian.com/world/2016/feb/22/beijing-raises-red-alert-

threshold-for-air-pollution-warning

[29] http://edition.cnn.com/2015/12/07/asia/china-beijing-pollution-red-alert/

[30] http://www.bbc.com/news/world-asia-china-35173709

[31] http://www.bbc.com/news/world-asia-china-35026363

[32] http://www.who.int/mediacentre/factsheets/fs313/en/

[33] http://www.nytimes.com/2015/08/14/world/asia/study-links-polluted-air-in-china-to-1-6-million-deaths-a-year.html?_r=0

[34] http://vnexpress.net/tin-tuc/khoa-hoc/trong-nuoc/bao-dong-o-nhiem-khong-khi-o-ha-noi-3364621.html

[35] http://thoitiet.net/index.asp?newsid=6407&PageNum=1

[36] http://genk.vn/kham-pha/mua-axit-tac-hai-va-cach-phong-ngua-20120724043655836.chn

[37] http://moitruongvietco.vn/nguyen-nhan-va-anh-huong-cua-mua-axit.html

[38] http://dantri.com.vn/xa-hoi/moi-ngay-tphcm-tang-them-300-o-to-1367241300.htm

[39] http://www.impe-qn.org.vn/impe-qn/vn/portal/InfoDetail.jsp?area=58&cat=1104&ID=2764

[40] http://moitruong.quangtri.gov.vn/index.php?language=vi&nv=news&op=Anh-huong-cua-o-nhiem-khong-khi/Anh-huong-cua-o-nhiem-khong-khi-86

[41] http://bis.net.vn/forums/t/489.aspx

[42] http://www.php.com.vn/nhung-dieu-khai-quat-co-ban-ve-php-framework.html

[43] http://hanoi.gov.vn/