

แบบจำลองสำหรับคาดการณ์ความหนาแน่นของฝุ่น PM 2.5 บริเวณภาคเหนือของประเทศไทย
โดยใช้ตัวแปรทางสภาพอากาศ

(Prediction model of PM 2.5 density in Northern region of Thailand using
atmospheric parameters)

นายศิลา ซาลี

รหัส 630510516

ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์

มหาวิทยาลัยเชียงใหม่

ปีการศึกษา 2566

แบบจำลองสำหรับคาดการณ์ความหนาแน่นของฝุ่น PM 2.5 บริเวณภาคเหนือของประเทศไทย
โดยใช้ตัวแปรทางสภาพอากาศ

(Prediction model of PM 2.5 density in Northern region of Thailand using
atmospheric parameters)

นายศิลา ซาลี

630510516

งานค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต

ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์

มหาวิทยาลัยเชียงใหม่

ภาคการศึกษาที่ 2 ปีการศึกษา 2566

แบบจำลองสำหรับคาดการณ์ความหนาแน่นของฝุ่น PM 2.5 บริเวณภาคเหนือของประเทศไทย
โดยใช้ตัวแปรทางสภาพอากาศ

(Prediction model of PM 2.5 density in Northern region of Thailand using
atmospheric parameters)

ได้รับพิจารณาอนุมัติให้เป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต
สาขาวิชาคณิตศาสตร์

คณะกรรมการควบคุมการค้นคว้าอิสระ

Parkpoom Phetpradap

(ผศ. ดร.ภาคภูมิ เพ็ชรประดับ)

ประธานกรรมการ

นอณา ฉัตรสกุลพรหม

(ผศ. ดร.นวินดา ฉัตรสกุลพรหม)

กรรมการ

วันที่ 4 มีนาคม พ.ศ. 2567

กิตติกรรมประกาศ

งานค้นคว้าอิสระฉบับนี้เป็นส่วนหนึ่งของการศึกษาในกระบวนวิชาการศึกษา ค้นคว้าอิสระ (206499) ตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต สาขาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ โดยมี วัตถุประสงค์เพื่อให้ นักศึกษาได้ศึกษา ค้นคว้า ประยุกต์ใช้ ทฤษฎี และแนวคิดทางคณิตศาสตร์ ในการทำการค้นคว้าอิสระ ซึ่งผู้ค้นคว้าได้ศึกษาในเรื่อง แบบจำลองสำหรับคาดการณ์ความหนาแน่นของฝุ่น PM 2.5 บริเวณภาคเหนือของประเทศไทย โดยใช้ตัวแปรทางสภาพอากาศ

ผู้ค้นคว้าขอขอบพระคุณ ประธานกรรมการสอบและในฐานะอาจารย์ที่ปรึกษา งานค้นคว้าอิสระ ที่อนุเคราะห์ทั้งร่างกาย แรงใจ สละเวลาอันมีค่าให้คำปรึกษา คำแนะนำ คำอธิบายในส่วนที่ผู้ค้นคว้าไม่เข้าใจ ให้ความรู้และคำแนะนำที่เกี่ยวข้องกับการทำงานค้นคว้าอิสระ การนำเสนอและรูปเล่มรายงาน ตลอดจนตรวจสอบความสมบูรณ์ถูกต้องการค้นคว้าอิสระนี้จนประสบความสำเร็จตามวัตถุประสงค์ทุกประการ

ขอขอบพระคุณ อาจารย์กรรมการสอบงานค้นคว้าอิสระที่ได้สละเวลาอันมีค่าให้ข้อเสนอแนะ พิจารณาและตรวจสอบความถูกต้องเหมาะสม รวมถึงให้คำแนะนำที่เป็นประโยชน์ต่อการปรับปรุงแก้ไขงานค้นคว้าอิสระฉบับนี้ให้สำเร็จลุล่วงด้วยดี

ทั้งนี้ขอกราบขอบพระคุณ บิดา มารดา ผู้มีพระคุณ ที่ให้กำลังใจ กำลังสติปัญญาอย่างสม่ำเสมอตลอดการ ทำงานตั้งแต่เริ่มต้นจนสำเร็จลุล่วงตามวัตถุประสงค์

ขอขอบคุณคณาจารย์ภาควิชาคณิตศาสตร์ที่ได้กรุณาอบรมสั่งสอนวิชาความรู้ทั้งในและนอกตำราให้แก่ผู้ค้นคว้าอันเป็นประโยชน์และสามารถนำมาประยุกต์ใช้ในการศึกษา ค้นคว้าอิสระในครั้งนี้

ขอขอบคุณเพื่อน พี่ น้อง ภาควิชาคณิตศาสตร์ตลอดจนผู้มีส่วนเกี่ยวข้องทุกท่านที่มีได้กล่าวถึง ที่ให้กำลังใจ ให้ความรู้ให้การช่วยเหลือ เป็นผลให้งานค้นคว้าอิสระนี้ออกมาเป็นไปตามรูปแบบและวัตถุประสงค์ทุกประการ

นายศิลา ชาลี

หัวข้อ (ภาษาไทย)	แบบจำลองสำหรับคาดการณ์ความหนาแน่นของฝุ่น PM 2.5 บริเวณภาคเหนือของประเทศไทย โดยใช้ตัวแปรทางสภาพอากาศ
(ภาษาอังกฤษ)	Prediction model of PM 2.5 density in Northern region of Thailand using atmospheric parameters
ชื่อผู้ทำการค้นคว้าอิสระ	นายศิลา ซาลี รหัสนักศึกษา 630510516
ชื่ออาจารย์ที่ปรึกษา	ผศ. ดร.ภาคภูมิ เพ็ชรประดับ
ชื่อกรรมการสอบ	ผศ. ดร.นวินดา ฉัตรสกุลพรหม

บทคัดย่อ

มลภาวะทางอากาศ โดยเฉพาะอย่างยิ่งฝุ่น PM 2.5 ฝุ่นละอองขนาดเล็ก ได้กลายเป็นปัญหาที่น่ากังวลสำหรับคนส่วนใหญ่ในโลก ณ ปัจจุบัน โดยเฉพาะอย่างยิ่งในบริเวณภาคเหนือของประเทศไทย ซึ่งค่าความหนาแน่นของฝุ่น PM 2.5 ($\mu\text{g}/\text{m}^3$) มักจะเกินระดับที่ปลอดภัย ทำให้เกิดความเสี่ยงต่อสุขภาพของประชาชนอย่างมาก โดยการค้นคว้าอิสระในครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองสำหรับการทำนายค่าความหนาแน่นของฝุ่น PM 2.5 ในบริเวณภาคเหนือของประเทศไทยในอีกหนึ่งชั่วโมงข้างหน้า โดยใช้ตัวแปรทางสภาพอากาศ ด้วยแบบจำลอง Autoregressive Distributed Lag (ARDL) โดยเราจะใช้ตัวแปรในแบบจำลองทั้งหมด 6 ตัวแปร ได้แก่ อุณหภูมิ ความชื้นของอากาศ ปริมาณน้ำฝน ความเร็วลม ความกดอากาศ และ ความหนาแน่นของฝุ่น PM 2.5 และเพื่อหาผลกระทบของตัวแปรทางสภาพอากาศทั้ง 5 ที่ส่งผลกระทบต่อระดับความหนาแน่นของฝุ่น PM 2.5 โดยผลการค้นคว้านี้จะมีความสำคัญต่อระบบสาธารณสุข โดยเฉพาะกับบุคคลที่มีโรคทางเดินหายใจ โดยสามารถระบุช่วงเวลาที่มีความเสี่ยงต่อการได้รับมลพิษทางอากาศนี้

Abstract

Air quality, especially fine particulate matter (PM 2.5), has emerged as a critical environmental and public health concern globally. This is particularly evident in Northern Thailand, where PM 2.5 frequently exceeds safe levels, posing substantial health risks during dry season. This study aims to develop predictive models for PM 2.5 density in Northern Thailand using atmospheric parameters. Employing an Autoregressive Distributed Lag (ARDL) model, we integrated six parameters: temperature, humidity, precipitation, wind speed, air pressure, and PM 2.5 itself, to predict PM 2.5 density in the next hour and to reflect the complex interplay of factors influencing PM 2.5 levels. Additionally, by data visualization from the data, the findings provide significant implications for public health, particularly for individuals with respiratory conditions, by identifying threatening months and general temporal windows of elevated risk.

สารบัญ

เรื่อง	หน้า
กิตติกรรมประกาศ	ก
บทคัดย่อ	ข
บทที่ 1 บทนำ (Introduction)	1
บทที่ 2 ความรู้พื้นฐาน (Preliminaries)	5
บทที่ 3 วิธีการศึกษา (Research methods)	11
บทที่ 4 ผลการศึกษา (Result)	19
บทที่ 5 สรุปผลการศึกษา (Conclusion)	34
บรรณานุกรม	37
ภาคผนวก	39

บทที่ 1

บทนำ (Introduction)

ในช่วงหลายปีที่ผ่านมา ปัญหาคุณภาพอากาศได้กลายเป็นปัญหาสิ่งแวดล้อมที่สำคัญ ซึ่งส่งผลกระทบต่อสุขภาพของประชาชน ภาวะเศรษฐกิจ และความสมดุลของระบบนิเวศ ในบรรดาสารมลพิษต่างๆ ฝุ่นละอองขนาดเล็ก (PM_{2.5}) ได้รับความสนใจเป็นพิเศษ เนื่องจากส่งผลกระทบต่อสุขภาพของมนุษย์ และ สิ่งมีชีวิตอื่น ๆ อย่างร้ายแรงและมักพบปริมาณมากในบริเวณชุมชนเมืองขนาดใหญ่รวมไปถึงพื้นที่บริเวณภาคเหนือของประเทศไทยเช่นเดียวกัน

PM 2.5 หมายถึงอนุภาคขนาดเล็กในบรรยากาศที่มีเส้นผ่านศูนย์กลางน้อยกว่า 2.5 ไมโครเมตร ด้วยขนาดของฝุ่นละอองที่เล็กมาก ๆ ฝุ่น PM 2.5 สามารถแทรกซึมลึกเข้าไปในปอดและเข้าสู่กระแสเลือด สามารถก่อให้เกิดโรคหัวใจ โรคหลอดเลือดสมอง และโรคระบบทางเดินหายใจ ปัญหานี้พบมากในบริเวณภาคเหนือของประเทศไทย ซึ่งมีความหนาแน่นของ PM_{2.5} ที่มักจะเกินระดับที่ปลอดภัย ซึ่งส่งผลกระทบต่อสุขภาพของประชาชนอย่างมาก



รูปที่ 1.1: รูปแสดงลักษณะภูมิประเทศบริเวณภาคเหนือของประเทศไทย [6]

แหล่งกำเนิดของ PM 2.5 ในภาคเหนือของประเทศไทยอาจจะต่างจากบริเวณชุมชนเมืองหรือแถบกรุงเทพมหานครเล็กน้อย เนื่องจากฝุ่นพิษในบริเวณภาคเหนือของประเทศไทยมักจะมีสาเหตุมาจากการเผาพื้นที่การเกษตร โดยเฉพาะการเผาซากข้าวโพด เผาอ้อย และเผาซังต่อข้าวเพื่อใช้สำหรับเตรียมเพาะปลูกในฤดูกาลถัดไป นอกจากนี้ยังมีสาเหตุมาจากไฟป่าในบริเวณภาคเหนือรวมไปถึงประเทศเพื่อนบ้าน ประจวบกับการที่ลักษณะภูมิประเทศของบริเวณภาคเหนือของประเทศไทย

มีลักษณะเป็นหุบเขาและแอ่งแผ่นดินระหว่างภูเขา ทำให้เกิดหมอกควันสะสมในพื้นที่ภาคเหนือของประเทศไทย

โดยงานวิจัยนี้จะมุ่งเน้นที่การพัฒนาแบบจำลอง โดยเราจะดัดแปลงแบบจำลอง Autoregressive Distributed Lag (ARDL) เพื่อใช้สำหรับทำนายความหนาแน่นของฝุ่น PM 2.5 ในอีก 1 ชั่วโมงข้างหน้า และเพื่อใช้สำหรับวิเคราะห์ตัวแปรทางสภาพอากาศต่าง ๆ ที่ส่งผลต่อความหนาแน่นของฝุ่น PM 2.5 ในบริเวณภาคเหนือของประเทศไทย

วัตถุประสงค์ของการค้นคว้าอิสระ (objectives)

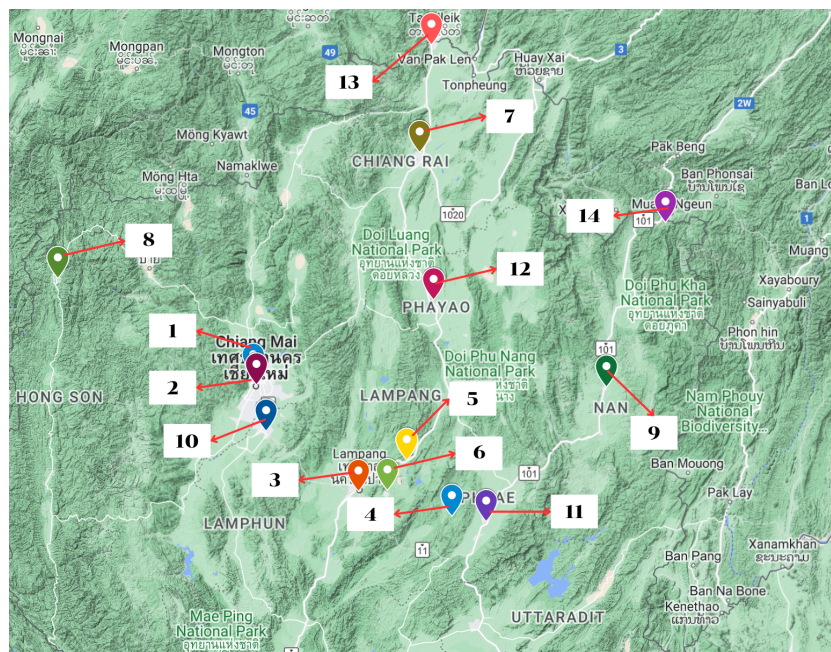
1. เพื่อสร้างแบบจำลองสำหรับทำนายความหนาแน่นของ PM 2.5 โดยใช้ตัวแปรทางสภาพอากาศต่าง ๆ ในบริเวณภาคเหนือของประเทศไทย
2. เพื่อหาผลกระทบของตัวแปรทางสภาพอากาศต่าง ๆ ที่ส่งผลต่อความหนาแน่นของ PM 2.5 ในพื้นที่ภาคเหนือของประเทศไทย

ขอบเขตของการค้นคว้าอิสระ

ในการค้นคว้าอิสระนี้ เราเลือกตัวแปรทางสภาพอากาศที่นำมาใช้ในแบบจำลองจากตัวแปรทางสภาพอากาศต่าง ๆ ที่สถานีตรวจวัดสภาพอากาศได้เก็บรวบรวมข้อมูลไว้ครบถ้วนมากที่สุด โดยตัวแปรทางสภาพอากาศที่เราใช้ในการสร้างแบบจำลองนั้นมีทั้งหมด 6 ตัวแปร ได้แก่

1. ความหนาแน่นของฝุ่น PM 2.5 (มคก./ลบ.ม.)
2. ความเร็วลม (เมตร/วินาที)
3. ความกดอากาศ (มิลลิเมตรปรอท)
4. อุณหภูมิ (องศาเซลเซียส)
5. ปริมาณน้ำฝน (มิลลิเมตร)
6. ความชื้นในอากาศ (% RH)

สาเหตุที่เราเลือกตัวแปรทางสภาพอากาศทั้ง 6 ตัวแปรนี้ เนื่องจากทั้ง 6 ตัวแปรนี้เป็นตัวแปรที่สถานีตรวจวัดสภาพอากาศส่วนใหญ่ได้เก็บข้อมูลย้อนหลังเป็นระยะเวลาที่นานพอสมควร อีกทั้งยังเป็นตัวแปรทางสภาพอากาศพื้นฐานที่สถานีตรวจวัดสภาพอากาศในพื้นที่ภาคเหนือส่วนใหญ่มีเครื่องมือ หรือเครื่องที่สามารถวัดค่าเหล่านี้ได้ จากการใช้ตัวแปรดังที่กล่าวมาข้างต้นน่าจะช่วยให้แบบจำลองที่เราสร้างมีความน่าเชื่อถือและมีความแม่นยำสำหรับพื้นที่ในบริเวณภาคเหนือมากยิ่งขึ้น



รูปที่ 1.2: รูปแสดงตำแหน่งที่ตั้งของ 14 สถานีตรวจวัดสภาพ/คุณภาพอากาศ ในบริเวณภาคเหนือของประเทศไทย

โดยเราได้รวบรวมมาจากสถานีตรวจวัดสภาพอากาศบริเวณพื้นที่ในภาคเหนือของประเทศไทย ได้แก่จังหวัด เชียงใหม่ ลำพูน ลำปาง แพร่ พะเยา น่าน เชียงราย และ แม่ฮ่องสอน ตั้งแต่วันที่ 1 มกราคม 2019 ถึงวันที่ 30 มิถุนายน 2023 โดยแบ่งแยกตามสถานีตรวจวัดสภาพอากาศดังต่อไปนี้

1. ศูนย์ราชการจังหวัดเชียงใหม่
2. โรงเรียนยุพราชวิทยาลัย
3. สถานีอุตุนิยมวิทยาลำปาง
4. โรงพยาบาลส่งเสริมสุขภาพตำบลบ้านสบป่าด
5. โรงพยาบาลส่งเสริมสุขภาพตำบลท่าสี่
6. การประปาส่วนภูมิภาคแม่เมาะ
7. สำนักงานทรัพยากรธรรมชาติและสิ่งแวดล้อมจังหวัดเชียงราย
8. สำนักงานทรัพยากรธรรมชาติและสิ่งแวดล้อมจังหวัดแม่ฮ่องสอน
9. สำนักงานเทศบาลเมืองน่าน
10. อุตุณิยมวิทยาจังหวัดลำพูน
11. อุตุณิยมวิทยาจังหวัดแพร่

12. สนามกีฬาจังหวัดพะเยา
13. สำนักงานสาธารณสุขแม่สาย
14. โรงพยาบาลเฉลิมพระเกียรติ จ.น่าน

โดยเราจะนำแบบจำลอง Auto Regressive Distributed Lag (ARDL) มาช่วยในการวิเคราะห์ผลกระทบของตัวแปรทางสภาพอากาศต่าง ๆ ที่ส่งผลต่อความหนาแน่นของ PM 2.5 และใช้สำหรับทำนายความหนาแน่นของ PM 2.5 โดยใช้ตัวแปรทางสภาพอากาศต่าง ๆ ในบริเวณภาคเหนือของประเทศไทย

วิธีการดำเนินการค้นคว้าอิสระ

1. รวบรวมข้อมูลและวิเคราะห์ข้อมูลเบื้องต้น
2. ศึกษางานวิจัยที่เกี่ยวข้อง
3. วิเคราะห์ผลกระทบของตัวแปรทางสภาพอากาศ
4. สร้างแบบจำลองสำหรับทำนายความหนาแน่นของ PM 2.5 โดยใช้ตัวแปรทางสภาพอากาศ
5. ทดสอบและปรับปรุงประสิทธิภาพของแบบจำลอง
6. สรุปผล
7. เขียนเล่ม และตรวจสอบความถูกต้อง

ประโยชน์ที่คาดว่าจะได้รับจากการค้นคว้าอิสระ (expected benefits and application)

1. ทราบแบบจำลองสำหรับทำนายความหนาแน่นของ PM 2.5 โดยใช้ตัวแปรทางสภาพอากาศต่าง ๆ ในบริเวณภาคเหนือของประเทศไทย
2. ทราบผลกระทบของตัวแปรทางสภาพอากาศต่าง ๆ ที่ส่งผลต่อความหนาแน่นของ PM 2.5 ในพื้นที่ภาคเหนือของประเทศไทย

บทที่ 2

ความรู้พื้นฐาน (Preliminaries)

ในบทนี้เราได้รวบรวมเนื้อหาที่เป็นความรู้พื้นฐานทั้งตัวบทนิยามและทฤษฎีบทที่ใช้ในงานค้นคว้าอิสระนี้เพื่อเพิ่มความเข้าใจให้แก่ผู้อ่าน โดยเริ่มจาก ความแปรปรวนร่วม (Covariance), ค่าสัมประสิทธิ์สหสัมพันธ์ (Pearson correlation coefficient), เมทริกซ์สหสัมพันธ์ (Correlation matrix), ค่าเฉลี่ย (Mean), ข้อมูลอนุกรมเวลา (Time series data), ข้อมูลที่มีความนิ่ง (Stationary data), แบบจำลองการถดถอยอัตโนมัติ (Autoregressive model), แบบจำลอง ARDL (Autoregressive distributed lag model), การตรวจสอบความเหมาะสมของแบบจำลอง (Diagnostic plots) และการประเมินประสิทธิภาพของแบบจำลอง (Model evaluation metrics) ซึ่งนำมาสู่ผลลัพธ์ของงานค้นคว้าอิสระนี้

บทนิยาม 2.1 [7] ค่าเฉลี่ย (Mean) คือ ผลบวกของค่าข้อมูลทั้งหมดหารด้วยจำนวนข้อมูล โดยสามารถคำนวณค่าเฉลี่ยของตัวแปร X ได้จากสูตร

$$\mu_X = \frac{1}{n} \sum_{i=1}^n X_i$$

โดย

μ_X คือ ค่าเฉลี่ยของตัวแปร X

X_i คือ ค่าของตัวแปร X ตัวที่ i

n คือ จำนวนข้อมูลทั้งหมด

บทนิยาม 2.2 [7] ความแปรปรวนร่วม (Covariance) คือ ค่าชี้วัดการเปลี่ยนแปลงของค่าตัวแปรใดตัวแปรหนึ่งเมื่อมีการเปลี่ยนแปลงในค่าตัวแปรอีกตัวหนึ่ง ซึ่งแสดงระดับความสัมพันธ์เชิงเส้นตรงของตัวแปร 2 ตัวแปร โดยค่าความแปรปรวนร่วมของ X และ Y ของประชากรสามารถคำนวณได้จาก

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N (X_i - \mu_X)(Y_i - \mu_Y)$$

โดย

σ_{XY} คือ ความแปรปรวนร่วมของ X และ Y

X_i คือ ค่าของตัวแปร X ตัวที่ i

Y_i คือ ค่าของตัวแปร Y ตัวที่ i

μ_X คือ ค่าเฉลี่ยของตัวแปร X

μ_Y คือ ค่าเฉลี่ยของตัวแปร Y

N คือ จำนวนข้อมูลรวมทั้งหมด

และค่าความแปรปรวนร่วมของ X และ Y ของกลุ่มตัวอย่างสามารถคำนวณได้จาก

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

โดย

s_{XY} คือ ความแปรปรวนร่วมของ X และ Y

X_i คือ ค่าของตัวแปร X ตัวที่ i

Y_i คือ ค่าของตัวแปร Y ตัวที่ i

$\bar{X}$ คือ ค่าเฉลี่ยของตัวแปร X

$\bar{Y}$ คือ ค่าเฉลี่ยของตัวแปร Y

n คือ จำนวนข้อมูลรวมทั้งหมด

บทนิยาม 2.3 [7] ค่าสัมประสิทธิ์สหสัมพันธ์ (Pearson correlation coefficient) คือ ค่าความแปรปรวนร่วมระหว่างตัวแปร 2 ตัวแปรที่ถูกทำให้เป็นบรรทัดฐาน แทนด้วยสัญลักษณ์ $\rho_{X,Y}$ โดยค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร X และ Y สามารถคำนวณได้จาก

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

โดย

$\rho_{X,Y}$ คือ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร X และ Y

$\text{Cov}(X,Y)$ คือ ค่าความแปรปรวนร่วมของ X และ Y

σ_X คือ ค่าเบี่ยงเบนมาตรฐานของ X

σ_Y คือ ค่าเบี่ยงเบนมาตรฐานของ Y

ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์ มีค่าที่เป็นไปได้อยู่ระหว่าง -1 และ 1

บทนิยาม 2.4 [7] เมทริกซ์สหสัมพันธ์ (Correlation matrix) คือ เมทริกซ์ที่ใช้สำหรับอธิบายความสัมพันธ์ระหว่างแต่ละคู่ของตัวแปร โดยเมทริกซ์สหสัมพันธ์ของชุดตัวแปร $X_1, X_2, \dots, X_n$ แทนด้วยสัญลักษณ์ $corr(X)$ สามารถสร้างได้ ดังนี้

$$corr(X) = \begin{bmatrix} 1 & \rho_{X_1, X_2} & \dots & \rho_{X_1, X_n} \\ \rho_{X_2, X_1} & 1 & \dots & \rho_{X_2, X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{X_n, X_1} & \rho_{X_n, X_2} & \dots & 1 \end{bmatrix}$$

บทนิยาม 2.5 [2] อนุกรมเวลา (Time series) คือ ค่าสังเกตหนึ่งชุดซึ่งถูกกำหนดขึ้น ณ เวลาต่าง ๆ โดยแบ่งออกเป็นอนุกรมเวลาต่อเนื่อง คือค่าสังเกตที่กระทำในเวลาต่อเนื่องกัน และอนุกรมเวลาไม่ต่อเนื่อง คือค่าสังเกตที่กระทำ ณ จุดเวลาที่ไม่ต่อเนื่องกัน

บทนิยาม 2.6 [2] ข้อมูลอนุกรมเวลา (Time series data) คือ ชุดของข้อมูลอนุกรมเวลาที่มีการเก็บรวบรวม อย่างเป็นระบบ

บทนิยาม 2.7 [2] ข้อมูลที่มีความนิ่ง (Stationary data) คือ ข้อมูลอนุกรมเวลาที่มีค่าเฉลี่ยและความแปรปรวนคงที่เมื่อเวลาเปลี่ยนแปลงไป

บทนิยาม 2.8 [9] การทดสอบความนิ่งของข้อมูล (Unit root test) เป็น การทดสอบข้อมูลอนุกรมเวลาว่ามี ลักษณะหนึ่งที่ระดับ level กล่าวคือ Integrated of order 0 = $I(0)$ หรือหนึ่งหลังจากผ่านการ difference ข้อมูล นั่นคือ Integrated of order $d = I(d), d > 0$

บทนิยาม 2.9 [9] การทดสอบ Augmented-Dickey fuller (ADF test) เป็นวิธีหนึ่งในการทดสอบ unit root โดยมีความสัมพันธ์ดังนี้

$$\Delta X_t = \alpha + \rho X_{t-1} + \sum_{i=1}^n \phi_i \Delta X_{t-1} + \varepsilon_t$$

โดย

ΔX_t คือ ผลต่างข้อมูลอนุกรมเวลาของตัวแปร X ณ เวลา t และ $t - 1$

ρ คือ ค่าสัมประสิทธิ์ (Coefficient) ของตัวแปร X_{t-1}

ϕ_i คือ สัมประสิทธิ์ของตัวแปรล่าช้า (Lagged variables)

ε_t คือ ความคลาดเคลื่อน ณ เวลา t

และสมมติฐานการทดสอบ ADF ได้แก่

$H_0 : \rho = 0$ (Non-stationary)

$H_1 : \rho < 0$ (Stationary)

บทนิยาม 2.10 การทำนายอนุกรมเวลา (Time series forecasting) คือ การพยากรณ์ค่าของตัวแปรตามในอนาคต โดยใช้ข้อมูลอนุกรมเวลาของตัวแปรนั้นมาศึกษาหาความสัมพันธ์ต่างๆ

บทนิยาม 2.11 [7] แบบจำลองการถดถอยอัตโนมัติ (Autoregressive model) คือ แบบจำลองสำหรับทำนายข้อมูลอนุกรมเวลาชนิดหนึ่ง โดยแบบจำลองนี้เป็นรูปแบบที่แสดงว่าค่าที่เราต้องการทำนาย X_t ถูกกำหนดจากค่าของ $X_{t-1}, \dots, X_{t-p}$ หรือค่าก่อนหน้าจำนวน p ค่า โดยแบบจำลอง AR(p) สามารถเขียนให้อยู่ในรูปสมการได้ดังนี้

$$X_t = \mu + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \dots + \varphi_p X_{t-p} + \varepsilon_t$$

โดยที่

X_t คือ ค่าของ X ณ เวลา t

μ คือ ค่าคงที่ (Constant term)

φ_i คือ ค่าสัมประสิทธิ์ (Coefficient) หน้า X_{t-i}

ε_t คือ ความคลาดเคลื่อน ณ เวลา t

บทนิยาม 2.12 [9] แบบจำลอง ARDL (Autoregressive distributed lag model) คือ แบบจำลองสำหรับทำนายข้อมูลอนุกรมเวลาชนิดหนึ่ง โดยแบบจำลอง ARDL($p, q_1, q_2, \dots, q_k$) สามารถเขียนให้อยู่ในรูปสมการได้ดังนี้

$$Y_t = \alpha + \sum_{i=1}^p \beta_i Y_{t-i} + \sum_{j=0}^{q_1} \gamma_{1j} X_{1,t-j} + \dots + \sum_{j=0}^{q_k} \gamma_{kj} X_{k,t-j} + \varepsilon_t$$

โดยที่

Y_t : คือค่าของ Y ณ เวลา t

α : คือ ค่าคงที่ (Constant term)

β_i : ค่าสัมประสิทธิ์ (Coefficient) ของค่าย้อนหลังของตัวแปรตาม (Lagged value) Y_{t-i}

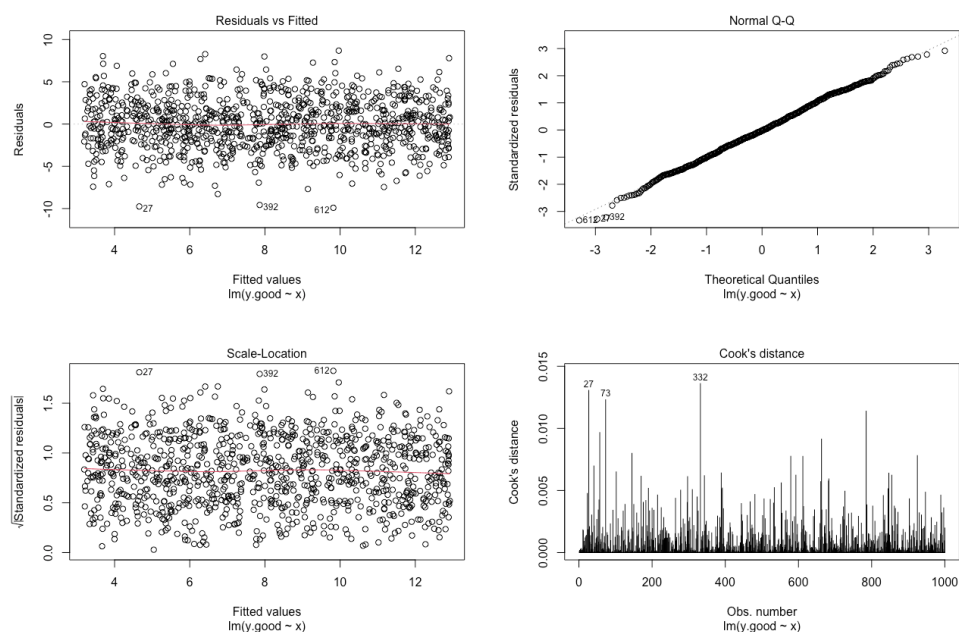
Y_{t-i} : ค่าย้อนหลังของตัวแปรตาม

$\gamma_{1j}, \dots, \gamma_{kj}$: คือ ค่าสัมประสิทธิ์ (Coefficient) ของค่าย้อนหลังของแต่ละตัวแปรต้น

$X_{1,t-j}, \dots, X_{k,t-j}$ คือ ค่าย้อนหลังของแต่ละตัวแปรต้น

ε_t คือ ความคลาดเคลื่อน ณ เวลา t

บทนิยาม 2.13 [3] การตรวจสอบความเหมาะสมของแบบจำลอง (Diagnostic plots) คือ เครื่องมือสำหรับแสดงผลที่ใช้สำหรับการวิเคราะห์ว่าตัวแบบจำลองที่เราใช้กับชุดข้อมูลนั้นมีความน่าเชื่อถือหรือสอดคล้องกับสมมติฐานของโมเดลนั้นมากน้อยเพียงใด โดยจะแบ่งเป็น 4 พล็อต ได้แก่



รูปที่ 2.1: รูปภาพแสดง diagnostic plots แบบต่าง ๆ

1. Residuals vs Fitted values คือ เครื่องมือที่ใช้สำหรับเช็คความสัมพันธ์เชิงเส้นตรงของแบบจำลอง
2. Normal q-q plot คือ เครื่องมือที่ใช้สำหรับการพิจารณาว่าค่าความคลาดเคลื่อนของแบบจำลองมีการกระจายตัวแบบปกติหรือไม่
3. Scale-location plot คือ เครื่องมือที่ใช้สำหรับในการตรวจสอบสมมติฐานการเป็น homoscedasticity
4. Residuals vs leverage plot คือ เครื่องมือที่ใช้สำหรับตรวจสอบผลกระทบของข้อมูลในชุดข้อมูล ว่าส่งผลทำให้แบบจำลองคลาดเคลื่อนหรือไม่

บทนิยาม 2.14 [7] การประเมินประสิทธิภาพของแบบจำลอง (Model evaluation metrics) คือ การประเมินเพื่อตรวจสอบว่าแบบจำลองที่เราได้ตีพิมพ์ที่จะเอาไปใช้งานได้หรือไม่ โดยพิจารณาจากความแม่นยำ หรือ ความถูกต้อง โดยการประเมินแบบจำลองสามารถประเมินได้หลายแบบเช่น

1. ค่า R^2 (R-squared) คือ ตัวสถิติที่ใช้วัดว่าตัวแบบคณิตศาสตร์ที่ได้นี้มีความสมรूपกับข้อมูลมากน้อยอย่างไร จากชุดข้อมูล n ข้อมูล โดยให้ y_i คือค่าจริงที่ได้จากการเก็บข้อมูล, $\hat{y}_i$ คือค่าที่ได้จากการทำนายของแบบจำลอง และ $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ คือค่าเฉลี่ยของค่าจริงที่ได้จากการเก็บข้อมูล และให้ $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ เป็นความแปรผันที่ไม่สามารถอธิบายได้ และ $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$ เป็นความแปรผันทั้งหมด หรือความแปรผันของตัวแปรตาม จะได้ว่าค่า

R^2 สามารถคำนวณได้จาก

$$R^2 = 1 - \frac{SSE}{TSS} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

2. MAE (Mean absolute error) เป็นค่าเฉลี่ยของผลต่างสัมบูรณ์ ระหว่างค่าจริง (y) กับค่าที่เกิดจากการทำนายของแบบจำลอง ($\hat{y}$) สามารถเขียนเป็นสมการได้ดังนี้

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

3. MAPE (Mean absolute percentage error) เป็นค่าเฉลี่ยของร้อยละผลต่างสัมบูรณ์ ระหว่างค่าจริง (y) กับค่าที่เกิดจากการทำนายของแบบจำลอง ($\hat{y}$) สามารถเขียนเป็นสมการได้ดังนี้

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

4. MSE (Mean square error) เป็นค่าเฉลี่ยผลต่างยกกำลังสองระหว่างค่าจริง (y) กับค่าที่เกิดจากการทำนายของแบบจำลอง ($\hat{y}$) สามารถเขียนเป็นสมการได้ดังนี้

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

5. RMSE (Root mean square error) เป็นรากที่สองของค่าเฉลี่ยผลต่างยกกำลังสองระหว่างค่าจริง (y) กับค่าที่เกิดจากการทำนายของแบบจำลอง ($\hat{y}$) สามารถเขียนเป็นสมการได้ดังนี้

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

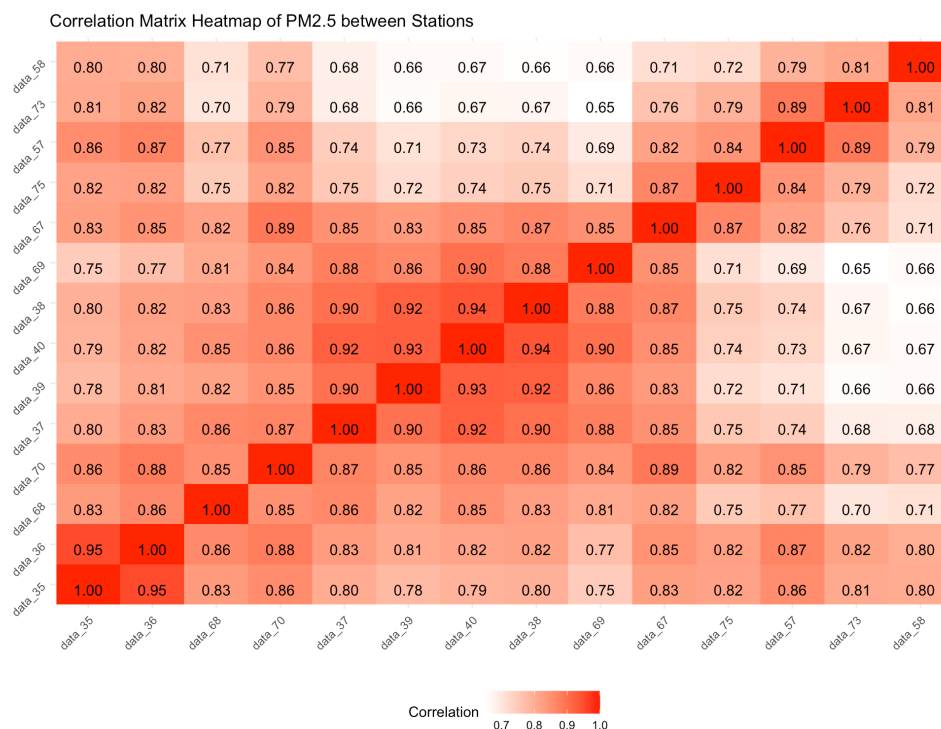
บทที่ 3

วิธีการศึกษา (Research methods)

งานค้นคว้าอิสระในหัวข้อ แบบจำลองสำหรับคาดการณ์ความหนาแน่นของฝุ่น PM 2.5 บริเวณภาคเหนือของประเทศไทย โดยใช้ตัวแปรทางสภาพอากาศ มีวิธีการศึกษาดังนี้

3.1 เกณฑ์การแบ่งพื้นที่ในการวิเคราะห์และทำนายความหนาแน่นของฝุ่น PM 2.5

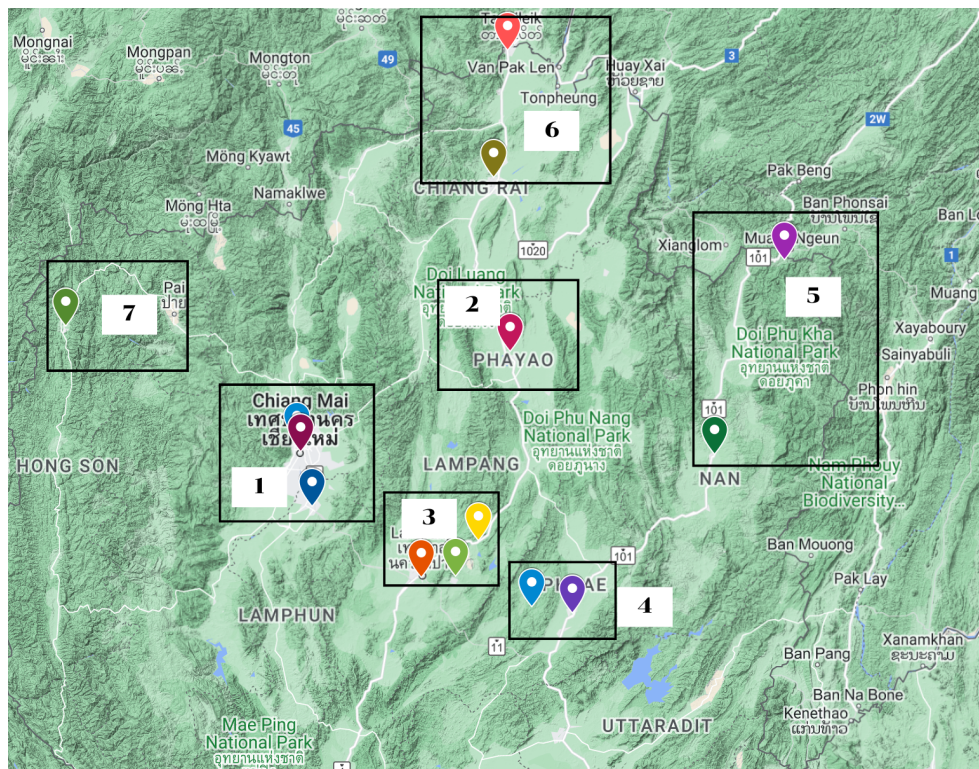
เนื่องจากค่าของตัวแปรทางสภาพอากาศต่าง ๆ ที่เราได้ขอข้อมูลมานั้นมาจาก 14 สถานีตรวจวัดสภาพอากาศที่แตกต่างกัน เราจึงพยายามยุบรวมบางสถานีเข้าด้วยกัน เพื่อให้ง่ายต่อการวิเคราะห์และสร้างแบบจำลองสำหรับทำนายความหนาแน่นของฝุ่น PM 2.5 รายพื้นที่ โดยเรานำค่า PM 2.5 ของแต่ละสถานีตรวจวัดสภาพอากาศในช่วงเวลาเดียวกัน มาหาค่าสัมประสิทธิ์สหสัมพันธ์ ได้ดังนี้



รูปที่ 3.1: รูปแสดงเมทริกซ์ของค่าสัมประสิทธิ์สหสัมพันธ์ของ PM 2.5 ระหว่างสถานีตรวจวัดสภาพอากาศต่าง ๆ

จากตารางแสดงเมทริกซ์ของค่าสัมประสิทธิ์สหสัมพันธ์ของ PM 2.5 ระหว่างสถานีตรวจวัด

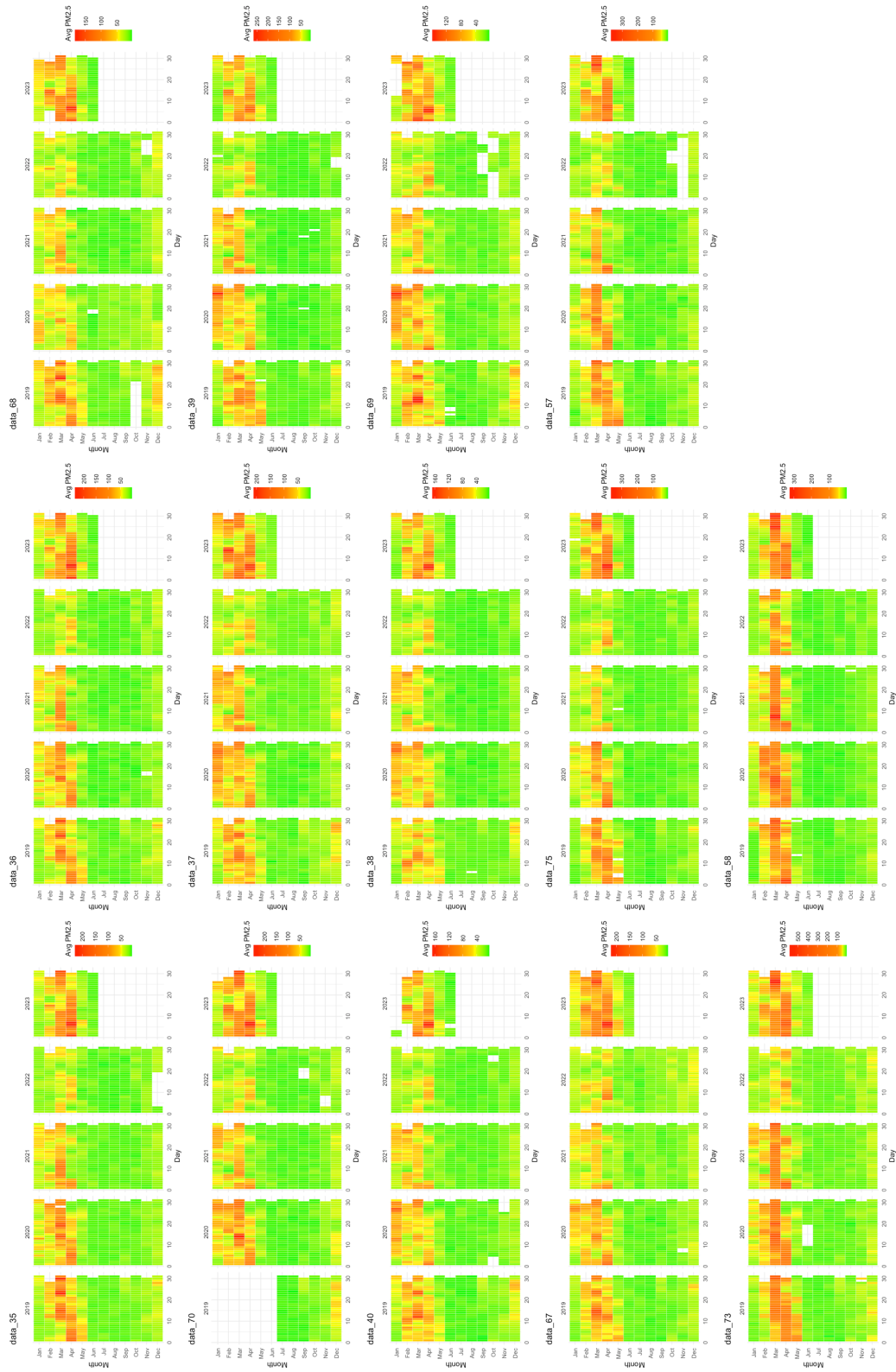
สภาพอากาศต่าง ๆ ทำให้เราสามารถรวมสถานีบางสถานีเข้าเป็นพื้นที่เดียวกันเพื่อให้ง่ายต่อการวิเคราะห์และง่ายต่อการสร้างแบบจำลองรายพื้นที่ โดยเราจะรวมสถานีที่ตั้งอยู่ในพื้นที่ใกล้เคียงกัน และมีค่าสัมประสิทธิ์สหสัมพันธ์ของ PM 2.5 มากกว่าหรือเท่ากับ 0.80 โดยจะแบ่งเป็นพื้นที่ได้ทั้งหมด 7 แอ่ง ได้แก่



รูปที่ 3.2: รูปแสดงการแบ่งพื้นที่ตามค่าสัมประสิทธิ์สหสัมพันธ์ของ PM 2.5 ของแต่ละสถานีตรวจวัดสภาพอากาศ

1. แอ่งเชียงใหม่ และลำพูน
2. แอ่งพะเยา
3. แอ่งลำปาง
4. แอ่งแพร่
5. แอ่งน่าน
6. แอ่งเชียงราย
7. แอ่งแม่ฮ่องสอน

ต่อไปเราจะเลือกช่วงระยะเวลาที่จะนำมาสร้างแบบจำลองสำหรับทำนายค่าความหนาแน่นของฝุ่น PM 2.5 โดยเรานำค่าความหนาแน่นของฝุ่น PM 2.5 เฉลี่ยต่อวัน มาสร้าง heatmap ดูแต่ละวันเพื่อดูค่าว่ามีความหนาแน่นมากในวัน หรือเดือนใด ได้ผลดังรูปที่ 3.3

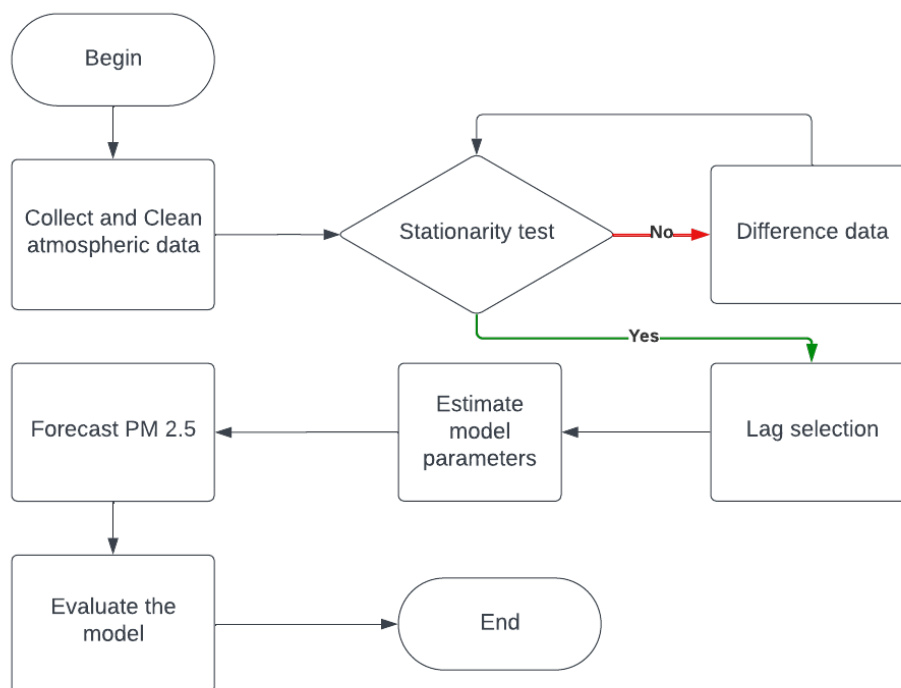


รูปที่ 3.3: Heat map แสดงค่าความหนาแน่นของ PM 2.5 เฉลี่ยต่อวันตั้งแต่ปี 2019 ถึง กลางปี 2023

จากรูป 3.3 จะเห็นได้ว่ามีแนวโน้มที่ค่อนข้างชัดเจนของระดับ PM 2.5 ที่มีการเพิ่มขึ้นอย่างมีนัยสำคัญตั้งแต่ต้นเดือนธันวาคม ถึงปลายเดือนพฤษภาคม ของทุกปี (2019-2023) ดังนั้นในแบบจำลองที่เราจะสร้างในครั้งนี้ จะมีขอบเขตของข้อมูลที่ใช้สร้างแบบจำลองอยู่ในช่วงเดือน ธันวาคม ถึงเดือนพฤษภาคมของทุกปี ซึ่งการเลือกช่วงระยะเวลาของข้อมูลที่จะนำมาสร้างแบบจำลองนั้น สอดคล้องกับวัตถุประสงค์ในการสร้างเครื่องมือการทำนายที่สามารถทำนายช่วงที่ฝุ่นมีค่าสูงได้แม่นยำยิ่งขึ้น

3.2 ขั้นตอนวิธีการสร้างแบบจำลอง ARDL สำหรับทำนายค่าฝุ่น Pm 2.5 และวิเคราะห์ตัวแปรทางสภาพอากาศต่าง ๆ ที่ส่งผลกระทบต่อความหนาแน่นของฝุ่น PM 2.5

หลังจากศึกษาตัวแบบจำลอง ARDL พบว่าการสร้างแบบจำลอง ARDL มีขั้นตอนวิธีดังนี้



รูปที่ 3.4: รูปแสดงขั้นตอนวิธีการสร้างแบบจำลอง ARDL

โดยแบบจำลอง ARDL สำหรับงานครั้งนี้สามารถเขียนให้อยู่ในรูปสมการได้ดังนี้

$$Y_t = \alpha + \sum_{i=1}^p \beta_i Y_{t-i} + \sum_{j=0}^{q_1} \gamma_j X_{1,t-j} + \sum_{j=0}^{q_2} \delta_j X_{2,t-j} + \sum_{j=0}^{q_3} \phi_j X_{3,t-j} + \sum_{j=0}^{q_4} \sigma_j X_{4,t-j} + \sum_{j=0}^{q_5} \lambda_j X_{5,t-j} + \varepsilon_t$$

โดยที่

α คือ ค่าคงที่ (Constant term)

Y_t คือ ค่าของ PM 2.5 ณ เวลา t

$X_{1,t}$ คือ ค่าของความเร็วลม ณ เวลา t

$X_{2,t}$ คือ ค่าของอุณหภูมิ ณ เวลา t

$X_{3,t}$ คือ ค่าของความชื้นในอากาศ ณ เวลา t

$X_{4,t}$ คือ ค่าความกดอากาศ ณ เวลา t

$X_{5,t}$ คือ ค่าของปริมาณน้ำฝน ณ เวลา t

β_i คือ ค่าสัมประสิทธิ์ (coefficient) ของ Y_{t-i}

γ_j คือ ค่าสัมประสิทธิ์ (coefficient) ของ $X_{1,t-j}$

δ_j คือ ค่าสัมประสิทธิ์ (coefficient) ของ $X_{2,t-j}$

ϕ_j คือ ค่าสัมประสิทธิ์ (coefficient) ของ $X_{3,t-j}$

σ_j คือ ค่าสัมประสิทธิ์ (coefficient) ของ $X_{4,t-j}$

λ_j คือ ค่าสัมประสิทธิ์ (coefficient) ของ $X_{5,t-j}$

ε_t คือ ค่าความคลาดเคลื่อน ณ เวลา t

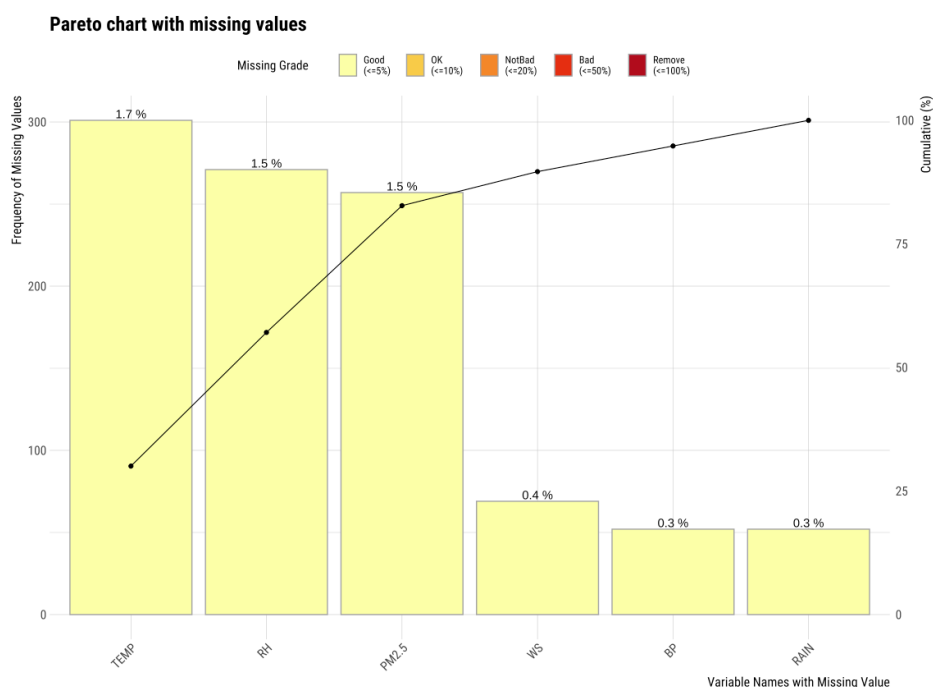
ซึ่งจะเห็นได้ว่าการทำนายค่า PM 2.5 ณ เวลา t ด้วยแบบจำลอง ARDL นั้นจะต้องอาศัยค่าของ PM 2.5 ณ เวลา $t-1, t-2, \dots, t-i$ และอาศัยค่าของตัวแปรทางสภาพอากาศอื่น ๆ ณ เวลา $t, t-1, t-2, \dots, t-j$ ซึ่งอาจจะไม่เหมาะสมสำหรับการนำไปประยุกต์ใช้ในการทำนายค่าฝุ่น PM 2.5 ในอนาคต ดังนั้นในงานนี้เราจึงปรับเปลี่ยนตัวแบบจำลองเล็กน้อยเพื่อให้เหมาะกับการนำไปใช้ในสถานการณ์จริง โดยเราจะใช้ค่าของตัวแปรต่าง ๆ ในอดีตทั้งหมด เพื่อใช้สำหรับทำนายค่าฝุ่น PM 2.5 ในอนาคต โดยจะได้แบบจำลอง ARDL ที่ปรับเปลี่ยนเล็กน้อยดังนี้

$$Y_t = \alpha + \sum_{i=1}^p \beta_i Y_{t-i} + \sum_{j=1}^{q_1} \gamma_j X_{1,t-j} + \sum_{j=1}^{q_2} \delta_j X_{2,t-j} + \sum_{j=1}^{q_3} \phi_j X_{3,t-j} + \sum_{j=1}^{q_4} \sigma_j X_{4,t-j} + \sum_{j=1}^{q_5} \lambda_j X_{5,t-j} + \varepsilon_t \quad (3.1)$$

ตัวอย่างการสร้างแบบจำลอง ARDL

เพื่อให้เห็นภาพมากยิ่งขึ้นเราจะทำการยกตัวอย่างขั้นตอนวิธีอย่างละเอียดในการสร้างแบบจำลองสำหรับทำนายค่าฝุ่น PM 2.5 ใน 1 ชั่วโมงข้างหน้าและสำหรับวิเคราะห์ตัวแปรทางสภาพอากาศต่าง ๆ ที่ส่งผลต่อความหนาแน่นของฝุ่น PM 2.5 ด้วยโปรแกรม R studio โดยเราจะยกตัวอย่างโดยใช้ข้อมูลของอ่างพะเยา ซึ่งขั้นตอนวิธีมีดังนี้

1. เก็บรวบรวมข้อมูลและจัดการกับข้อมูล



รูปที่ 3.5: รูปแสดงการวิเคราะห์ Missing values ของตัวแปรทางสภาพอากาศต่าง ๆ

ในขั้นตอนแรกเราได้ขอข้อมูลจากสถานีตรวจวัดสภาพอากาศในจังหวัดพะเยา หลังจากได้ข้อมูลมา เราจะจัดการกับข้อมูลบางค่าที่ขาดหายไป (Missing values) ในบางตัวแปรสภาพอากาศ ซึ่งจากการวิเคราะห์ตามรูปที่ 3.5 พบว่า ในแต่ละตัวแปรทางสภาพอากาศที่เราเลือกใช้ทั้ง 6 ตัวแปร มีค่าที่หายไปไม่เกิน 5% ดังนั้นเราสามารถไม่สนใจข้อมูลบางค่าที่ขาดหายไปเหล่านั้นได้ (Shafer, 1999)

2. ทดสอบความนิ่งของข้อมูลทางสภาพอากาศต่าง ๆ

เนื่องจากข้อมูลตัวแปรทางสภาพอากาศที่เราจะนำมาใช้ในตัวแบบจำลองต้องมีลักษณะที่นิ่ง (Stationary) จากบทนิยาม 2.9 เราจะได้ผลการทดสอบความนิ่งของแต่ละตัวแปรดังนี้ จากตารางที่ 3.1 เห็นได้ว่าตัวแปรทางสภาพอากาศทุกตัวมีลักษณะนิ่งที่ $I(0)$ ดังนั้นเราจึงสามารถใช้ข้อมูลอนุกรมเวลาของตัวแปรทางสภาพอากาศทั้ง 6 ตัวแปรนี้เพื่อสร้างแบบจำลองได้

3. การหาค่าสัมประสิทธิ์ในแบบจำลอง

หลังจากที่เราตรวจสอบว่าข้อมูลอนุกรมเวลาของตัวแปรทางสภาพอากาศต่าง ๆ มีลักษณะที่นิ่งในระดับ Level หรือที่ 1^{st} difference ($I(1)$) เนื่องจากการหา optimal lag ของแต่ละตัวแปรไม่สามารถการันตีได้ว่าเมื่อนำมาสร้างแบบจำลองแล้วจะให้แบบจำลองที่มี

ตัวแปรทางสภาพอากาศ	ระดับของ Unit root	ADF t- statistic	p-value	result
ฝุ่น PM 2.5	$I(0)$	-7.00	0.01	stationary
ความเร็วลม	$I(0)$	-14.38	0.01	stationary
อุณหภูมิ	$I(0)$	-6.17	0.01	stationary
ความชื้นในอากาศ	$I(0)$	-6.82	0.01	stationary
ความกดอากาศ	$I(0)$	-6.10	0.01	stationary
ปริมาณน้ำฝน	$I(0)$	-7.00	0.01	stationary

ตารางที่ 3.1: การทดสอบ Unit Root ที่ระดับ level ($I(0)$) ด้วยวิธี Augmented Dickey-Fuller Test (ADF)

ประสิทธิภาพหรือแม่นยำมากที่สุด เราจึงใช้ฟังก์ชันในโปรแกรม Rstudio ที่มีชื่อว่า `auto.ardl()` เพื่อหา optimal lag ของแต่ละตัวแปรในแบบจำลองที่ทำให้แบบจำลองอธิบายตัวแปรตามที่เราต้องการหรือ ค่าความหนาแน่นของฝุ่น PM 2.5 ได้ดีที่สุด โดยผลลัพธ์ของการสร้างแบบจำลองจะได้ดังภาพที่ 3.6 และได้ Diagnostic plots ดังภาพที่ 3.7

```
Time series regression with "ts" data:
Start = 3, End = 13409

Call:
dynlm::dynlm(formula = full_formula, data = data, start = start,
  end = end)

Residuals:
    Min       1Q   Median       3Q      Max
-78.263  -3.410   -0.182    3.073  147.779

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  93.172243   24.021389   3.879  0.000106 ***
L(PM2.5, 1)   0.972905    0.008570  113.528 < 2e-16 ***
L(PM2.5, 2)  -0.028462    0.008518   -3.342  0.000835 ***
WS           -0.875864    0.128417   -6.820  9.46e-12 ***
TEMP         -0.172625    0.020559   -8.397 < 2e-16 ***
RH           -0.022573    0.017829   -1.266  0.205507
L(RH, 1)      0.179384    0.030530    5.876  4.31e-09 ***
L(RH, 2)     -0.225336    0.017479  -12.892 < 2e-16 ***
BP           -0.112084    0.032695   -3.428  0.000610 ***
RAIN         -0.174989    0.066908   -2.615  0.008924 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.796 on 13397 degrees of freedom
Multiple R-squared:  0.9325,    Adjusted R-squared:  0.9324
F-statistic: 2.056e+04 on 9 and 13397 DF,  p-value: < 2.2e-16
```

รูปที่ 3.6: รูปแสดงผลลัพธ์จากการสร้างแบบจำลองทำนายค่าฝุ่น PM 2.5 โดยอาศัยตัวแปรทางสภาพอากาศต่าง ๆ

จากภาพที่ 3.6 จะได้ว่าแบบจำลองนี้มีค่า R^2 อยู่ที่ 0.9325 ซึ่งถือว่ามีค่าสูงมาก โดยเราจะนำแบบจำลองนี้ไปทดสอบกับ test data เพื่อหาว่าแบบจำลองนี้มีประสิทธิภาพมากน้อยเพียงใด จากสมการที่ 3.1 เราจะได้สมการของการทำนายค่าฝุ่น PM 2.5 ในอีก 1 ชั่วโมงข้างหน้า

จากรูปที่ 3.6 คือ

$$Y_t = 93.17 + 0.97(Y_{t-1}) - 0.03(Y_{t-2}) - 0.88(X_{1,t-1}) \\ - 0.17(X_{2,t-1}) - 0.02(X_{3,t-1}) + 0.18(X_{3,t-2}) \\ - 0.23(X_{3,t-3}) - 0.11(X_{4,t-1}) - 0.17(X_{5,t-1})$$

โดยที่

Y_t คือ ค่าของ PM 2.5 ณ เวลา t

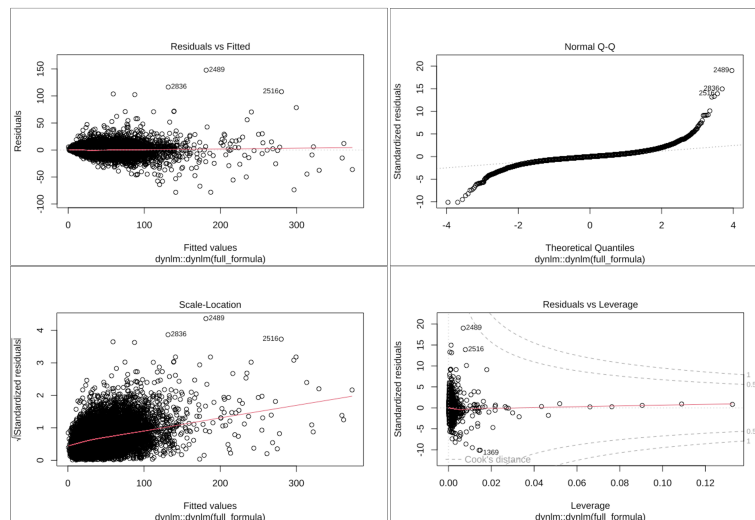
$X_{1,t}$ คือ ค่าของความเร็วลม ณ เวลา t

$X_{2,t}$ คือ ค่าของอุณหภูมิ ณ เวลา t

$X_{3,t}$ คือ ค่าของความชื้นในอากาศ ณ เวลา t

$X_{4,t}$ คือ ค่าความกดอากาศ ณ เวลา t

$X_{5,t}$ คือ ค่าของปริมาณน้ำฝน ณ เวลา t



รูปที่ 3.7: รูปแสดงdiagnostic plots ของแบบจำลอง

จากบทนิยามที่ 2.13 ได้ว่าแบบจำลองนั้นค่อนข้างเป็นไปตามสมมติฐาน โดยเราจะวัดว่าแบบจำลองนั้นน่าเชื่อถือหรือไม่จากการนำไปทดสอบกับตัว test data ว่ามีความแม่นยำหรือไม่อีกทีหนึ่ง

สำหรับการสร้างแบบจำลองในแง่อื่น ๆ สามารถทำได้โดยอาศัยวิธีเดียวกันกับที่ได้กล่าวไปข้างต้น โดยผลการทดสอบประสิทธิภาพและความแม่นยำของแบบจำลองโดยอาศัยตัวแปรทางสถิติ rmse จะแสดงในบทที่ 4 ต่อไป

บทที่ 4

ผลการศึกษา (Result)

งานค้นคว้าอิสระในหัวข้อ แบบจำลองสำหรับคาดการณ์ความหนาแน่นของฝุ่น PM 2.5 บริเวณภาคเหนือของประเทศไทย โดยใช้ตัวแปรทางสภาพอากาศ จะแบ่งผลการศึกษาออกเป็น 2 หัวข้อย่อย โดยจะแบ่งตามความครบถ้วนของข้อมูลของตัวแปรทางสภาพอากาศต่าง ๆ โดยจะแบ่งเป็นแบบจำลองสำหรับพื้นที่ที่มีข้อมูลทางสภาพอากาศครบถ้วนทุกตัวแปรได้แก่แอ่งพะเยา แอ่งแพร่ แอ่งน่าน และแอ่งแม่ฮ่องสอน และ แบบจำลองสำหรับพื้นที่ที่มีข้อมูลบางตัวแปรขาดหายไป พื้นที่เหล่านี้ได้แก่แอ่งเชียงใหม่ แอ่งลำพูน และแอ่งเชียงราย เนื่องจากพื้นที่เหล่านี้มีบางสถานีตรวจวัดสภาพอากาศในพื้นที่นั้น ๆ ขาดข้อมูลของตัวแปรความกดอากาศ (Air pressure) ซึ่งไม่ได้มีการบันทึกข้อมูลไว้ เราจึงแบ่งแบบจำลองสำหรับพื้นที่ที่มีข้อมูลบางตัวแปรขาดหายไปออกเป็นอีก 2 กรณีย่อยได้แก่ แบบจำลองในกรณีที่ไม่นำตัวแปรความกดอากาศมาพิจารณาในแบบจำลอง และกรณีที่อนุমানข้อมูลตัวแปรความกดอากาศจากสถานีที่อยู่ในพื้นที่เดียวกันที่ใกล้ที่สุด ซึ่งจะได้ผลการศึกษาดังนี้

4.1 แบบจำลองสำหรับทำนายความหนาแน่นของฝุ่น PM 2.5 ของแอ่งพะเยา แอ่งแพร่ แอ่งน่าน และแอ่งแม่ฮ่องสอน

หลังจากการสร้างแบบจำลองตามขั้นตอนวิธีดังรูป 3.4 จะได้ตารางค่าสัมประสิทธิ์ในแบบจำลองสำหรับคาดการณ์ความหนาแน่นของฝุ่น PM 2.5 บริเวณภาคเหนือของประเทศไทยในแอ่งพะเยา (Zone 2) แอ่งแพร่ (Zone 4) แอ่งน่าน (Zone 5) และแอ่งแม่ฮ่องสอน (Zone 7) ของแต่ละตัวแปรจากสมการ 3.1 ดังตารางที่ 4.1 และ 4.2

Parameters	intercept	PM2.5 ($\mu g/m^3$)					Windspeed (m/s)						Temperature (C)					
		t-1	t-2	t-3	t-4	t-5	t-1	t-2	t-3	t-4	t-5	t-6	t-1	t-2	t-3	t-4	t-5	t-6
Zone 2	93.17	0.97	-0.03				-0.88						-0.17					
Zone 4	4.21	1.08	-0.19	-0.04	0.08	0.03	-1.95	0.92	0.38	-0.14	0.15	0.69	-0.07	-0.58	0.53			
Zone 5	21.19	1.09	-0.19	0.00	0.02	0.04	-2.94	0.58	0.85	0.04	-0.03	-0.25	0.02	-0.67	0.6			
Zone 7	289.3	1.14	-0.37	0.15	-0.03	0.08	-2.79						-2.6	-0.75	2.21	-0.03	-0.15	0.91

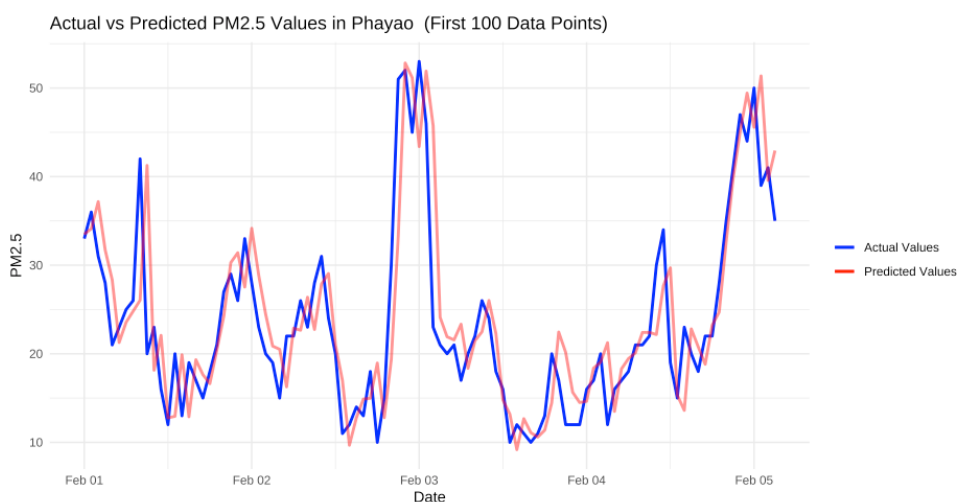
ตารางที่ 4.1: ตารางแสดงค่าสัมประสิทธิ์ของ ค่าคงที่ในแบบจำลอง Pm 2.5 ความเร็วลม และอุณหภูมิ

Parameters	Humidity (% RH)						Rain (mm)						Pressure (mmHg)					
	t-1	t-2	t-3	t-4	t-5	t-6	t-1	t-2	t-3	t-4	t-5	t-6	t-1	t-2	t-3	t-4	t-5	t-6
Zone 2	-0.02	0.18	-0.23				-0.11						-0.17					
Zone 4	-0.03												0.6	-0.11	-0.31	-0.57	-0.36	0.64
Zone 5	0.02	-0.17	0.11										0.4	-0.19	-0.25	-0.47	-0.16	0.67
Zone 7	-0.64	-0.07	0.50										0.01	0.20	-0.03	-0.19	-0.60	

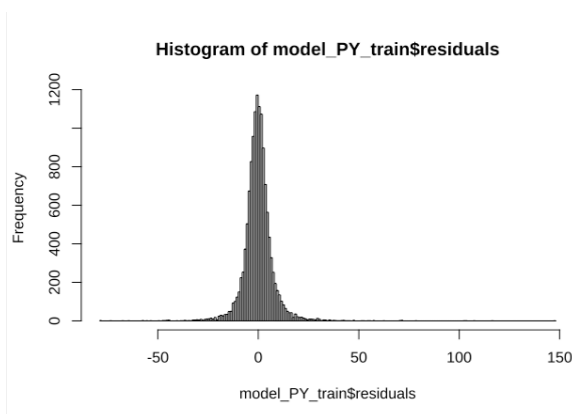
ตารางที่ 4.2: ตารางแสดงค่าสัมประสิทธิ์ของความชื้นในอากาศ ปริมาณน้ำฝน และความกดอากาศ

หลังจากที่ได้แบบจำลองสำหรับคาดการณ์ความหนาแน่นของฝุ่น PM 2.5 บริเวณภาคเหนือของประเทศไทยในพื้นที่ดังกล่าวแล้ว เราได้สร้างฮิสโตแกรม (Histogram) ของค่าความคลาดเคลื่อนของแบบจำลองในแต่ละพื้นที่ โดยแกน X แสดงถึงค่าความคลาดเคลื่อน และแกน Y แสดงถึงค่าความถี่ของค่าความคลาดเคลื่อนแต่ละค่า จากนั้นเราจึงนำแบบจำลองไปทำนายกับชุดข้อมูลที่แบบจำลองไม่เคยพบเจอ (Test data set) และพล็อตระหว่างค่า PM 2.5 ที่เราเก็บค่าได้จริง (Actual values) กับ ค่าที่แบบจำลองทำนายได้ (Predicted value) เนื่องจากข้อมูลของเรามีจำนวนข้อมูลที่ค่อนข้างเยอะ (Dense data) เราจึงเลือกช่วงบางช่วงมาพล็อตเพื่อให้เห็นถึงความแตกต่างของค่าจริงที่ได้จากเครื่องวัดของสถานีตรวจวัดสภาพอากาศต่าง ๆ กับค่าความหนาแน่นของฝุ่น PM 2.5 ที่แบบจำลองทำนายไว้ดังต่อไปนี้

ประสิทธิภาพของแบบจำลองในแอ่งพะเยา

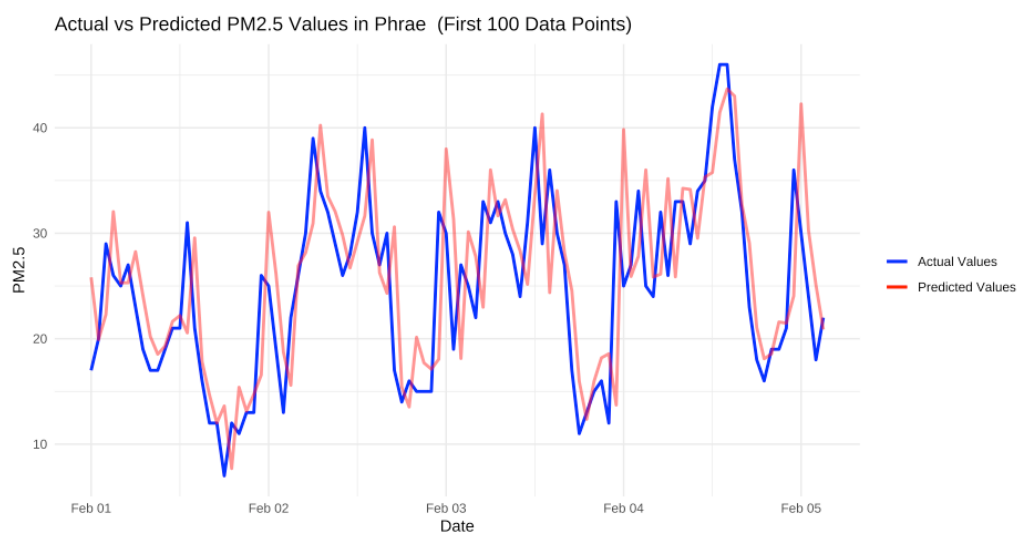


รูปที่ 4.1: กราฟระหว่างค่าจริงและค่าที่แบบจำลองทำนายค่าของ PM 2.5 ระหว่างวันที่ 1 กุมภาพันธ์ 2566 ถึง 5 กุมภาพันธ์ 2566

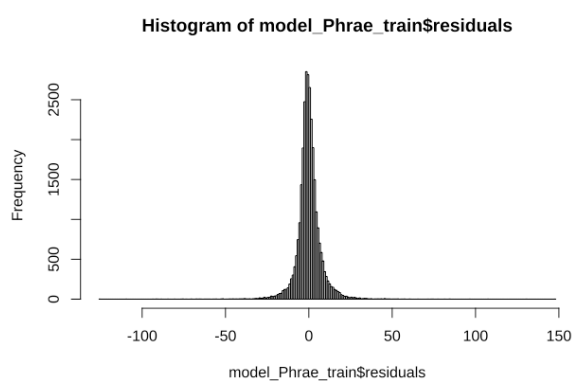


รูปที่ 4.2: กราฟแสดงฮิสโตแกรมของความคลาดเคลื่อนของแบบจำลอง

ประสิทธิภาพของแบบจำลองในแอ่งแพร่

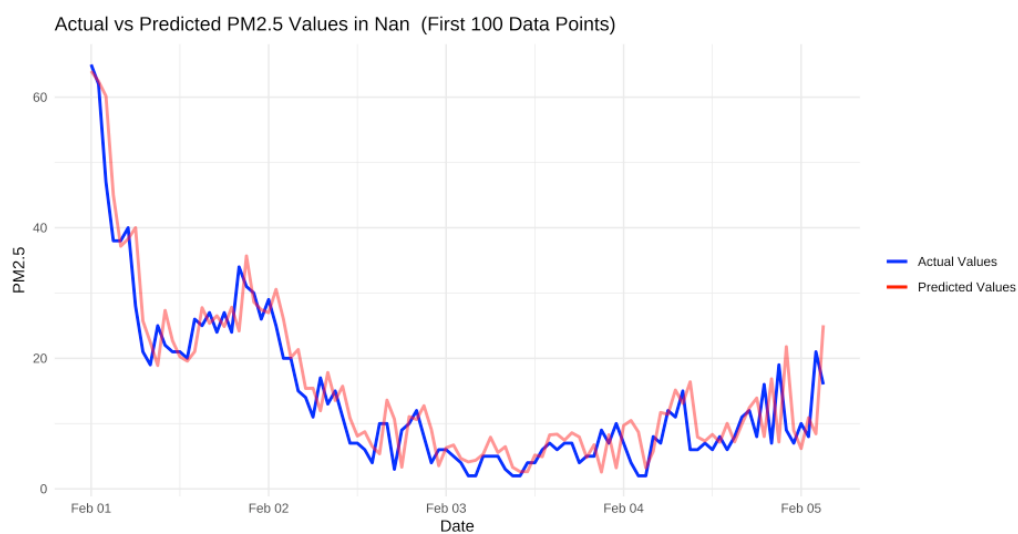


รูปที่ 4.3: กราฟระหว่างค่าจริงและค่าที่แบบจำลองทำนายค่าของ PM 2.5 ระหว่างวันที่ 1 กุมภาพันธ์ 2566 ถึง 5 กุมภาพันธ์ 2566

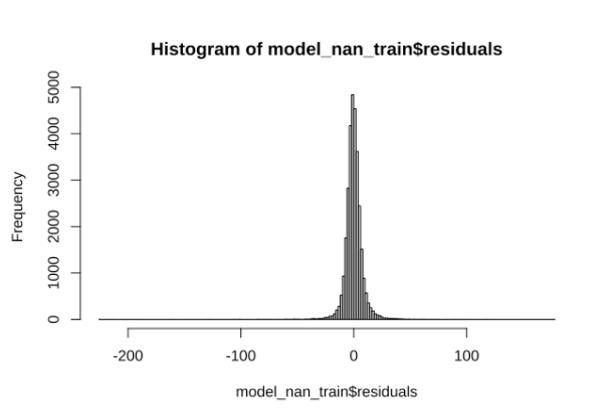


รูปที่ 4.4: กราฟแสดงฮิสโตแกรมของความคลาดเคลื่อนของแบบจำลอง

ประสิทธิภาพของแบบจำลองในแอ่งน่าน

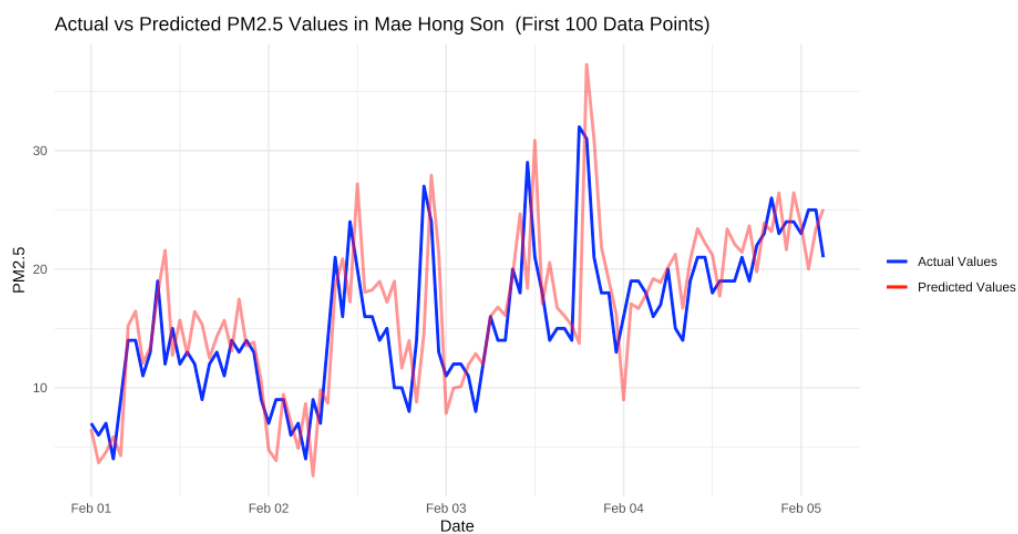


รูปที่ 4.5: กราฟระหว่างค่าจริงและค่าที่แบบจำลองทำนายค่าของ PM 2.5 ระหว่างวันที่ 1 กุมภาพันธ์ 2566 ถึง 5 กุมภาพันธ์ 2566

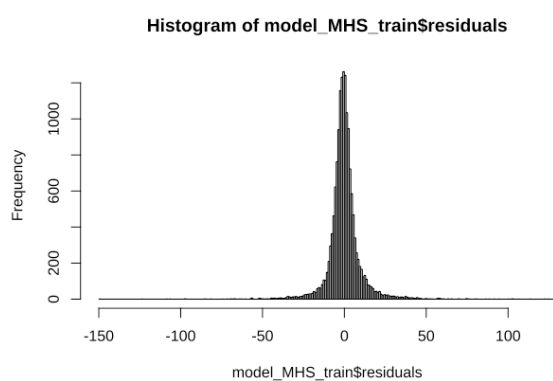


รูปที่ 4.6: กราฟแสดงฮิสโตแกรมของความคลาดเคลื่อนของแบบจำลอง

ประสิทธิภาพของแบบจำลองในแอ่งแม่ฮ่องสอน



รูปที่ 4.7: กราฟระหว่างค่าจริงและค่าที่แบบจำลองทำนายค่าของ PM 2.5 ระหว่างวันที่ 1 กุมภาพันธ์ 2566 ถึง 5 กุมภาพันธ์ 2566



รูปที่ 4.8: กราฟแสดงฮิสโตแกรมของความคลาดเคลื่อนของแบบจำลอง

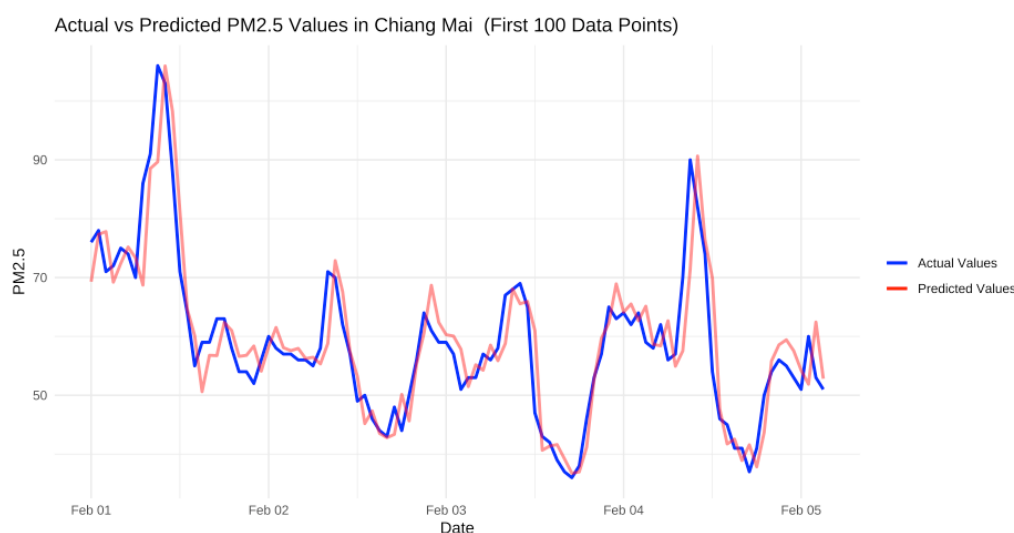
4.2 แบบจำลองสำหรับทำนายความหนาแน่นของฝุ่น PM 2.5 ของแอ่งเชียงใหม่และลำพูน แอ่งลำปาง และแอ่งเชียงราย

สำหรับแอ่งเชียงใหม่และลำพูน (Zone 1) แอ่งลำปาง (Zone 3) และแอ่งเชียงราย (Zone 6) จะมีบางสถานีตรวจวัดสภาพอากาศในพื้นที่นั้น ๆ ไม่ได้เก็บข้อมูลของตัวแปรทางสภาพอากาศบางตัวแปรไว้ เราจึงแบ่งแบบจำลองสำหรับ 3 พื้นที่นี้ออกเป็น 2 แบบจำลองได้แก่ 1.แบบจำลองที่ไม่นำตัวแปรที่ไม่ได้เก็บข้อมูลมาวิเคราะห์เลย และ 2. แบบจำลองที่เราจะอนุมานค่าตัวแปรที่บางสถานีตรวจวัดสภาพอากาศในพื้นที่นั้น ๆ ไม่ได้เก็บข้อมูลไว้ด้วยสถานีตรวจวัดสภาพอากาศในพื้นที่เดียวกันที่ใกล้ที่สุดและ ณ เวลาเดียวกัน หลังจากนั้นจะสร้างแบบจำลองตามขั้นตอนวิธีดังรูป 3.4 และจะได้ตารางค่าสัมประสิทธิ์ในแบบจำลองสำหรับคาดการณ์ความหนาแน่นของฝุ่น PM 2.5 บริเวณภาคเหนือของประเทศไทยในแอ่งเชียงใหม่และลำพูน แอ่งลำปาง และแอ่งเชียงราย ของแต่ละตัวแปร จากสมการ 3.1 ดังต่อไปนี้

4.2.1 แบบจำลองในกรณีที่ไม่นำตัวแปรที่ไม่ได้เก็บข้อมูลมาวิเคราะห์ในแบบจำลอง

ตารางค่าสัมประสิทธิ์ในแบบจำลองของกรณีนี้แสดงดังตารางที่ 4.3 และ 4.4 และเช่นเดียวกับ 4.1 เราได้ทำการวัดประสิทธิภาพของแบบจำลอง ได้ผลดังต่อไปนี้

ประสิทธิภาพของแบบจำลองในแอ่งเชียงใหม่และลำพูน



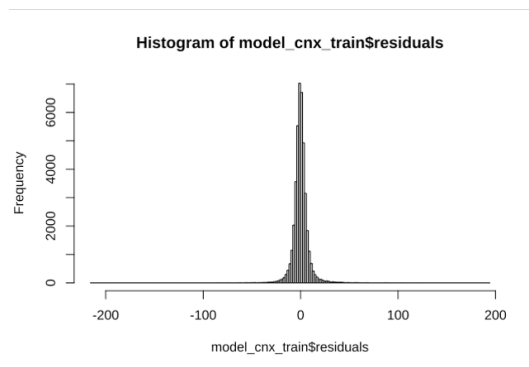
รูปที่ 4.9: กราฟระหว่างค่าจริงและค่าที่แบบจำลองทำนายค่าของ PM 2.5 ระหว่างวันที่ 1 กุมภาพันธ์ 2566 ถึง 5 กุมภาพันธ์ 2566

Parameters	intercept	PM2.5 ($\mu\text{g}/\text{m}^3$)						Windspeed (m/s)					
		t-1	t-2	t-3	t-4	t-5		t-1	t-2	t-3	t-4	t-5	t-6
Zone 1	7.31	1.14	-0.21	0.03				-2.09	-0.33	0.45			
Zone 3	1.68	1.25	-0.34	-0.01	-0.02	0.09		-0.97	0.01	0.36			
Zone 6	5.86	1.33	-0.44	0.11	-0.08	0.05		-2.26	0.06	0.82			

ตารางที่ 4.3: ตารางแสดงค่าสัมประสิทธิ์ของ ค่าคงที่ในแบบจำลอง Pm 2.5 และ ความเร็วลม

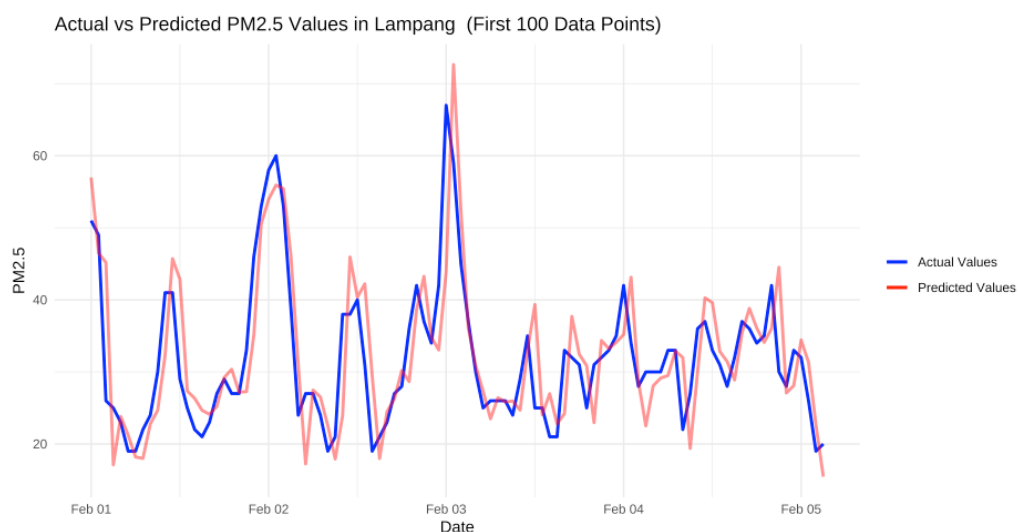
Parameters	Temperature (C)						Humidity (% RH)						Rain (mm)					
	t-1	t-2	t-3	t-4	t-5	t-6	t-1	t-2	t-3	t-4	t-5	t-6	t-1	t-2	t-3	t-4	t-5	t-6
Zone 1	-0.04	-0.85	0.03	0.30	0.25	0.22	-0.06	-0.04	-0.05	0.05	-0.01	0.07						
Zone 3	-0.01	-0.32	0.01	0.12	0.06	0.15												
Zone 6	0.05	-1.2	0.4	0.13	0.36	0.19	-0.01	-0.27	0.09	0.02	0.05	0.06	-0.22	0.13				

ตารางที่ 4.4: ตารางแสดงค่าสัมประสิทธิ์ของ อุณหภูมิ ความชื้นในอากาศและปริมาณน้ำฝน

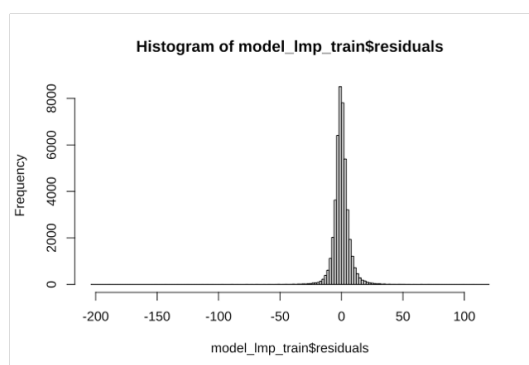


รูปที่ 4.10: กราฟแสดงฮิสโตแกรมของความคลาดเคลื่อนของแบบจำลอง

ประสิทธิภาพของแบบจำลองในแอ่งลำปาง

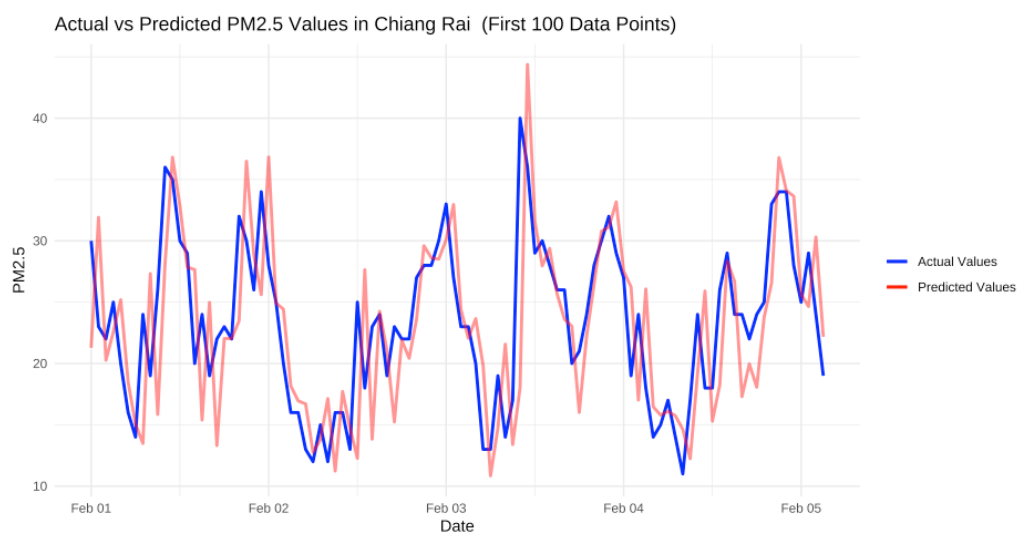


รูปที่ 4.11: กราฟระหว่างค่าจริงและค่าที่แบบจำลองทำนายค่าของ PM 2.5 ระหว่างวันที่ 1 กุมภาพันธ์ 2566 ถึง 5 กุมภาพันธ์ 2566

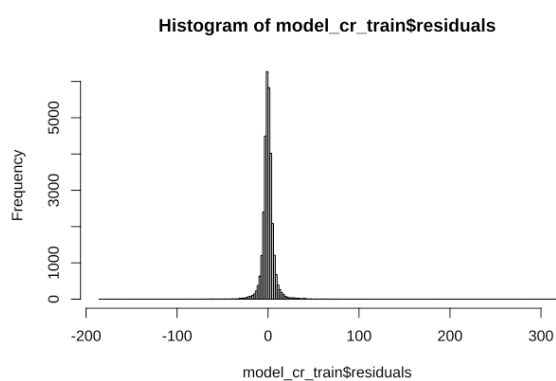


รูปที่ 4.12: กราฟแสดงฮิสโตแกรมของความคลาดเคลื่อนของแบบจำลอง

ประสิทธิภาพของแบบจำลองในแอ่งเชียงใหม่



รูปที่ 4.13: กราฟระหว่างค่าจริงและค่าที่แบบจำลองทำนายค่าของ PM 2.5 ระหว่างวันที่ 1 กุมภาพันธ์ 2566 ถึง 5 กุมภาพันธ์ 2566

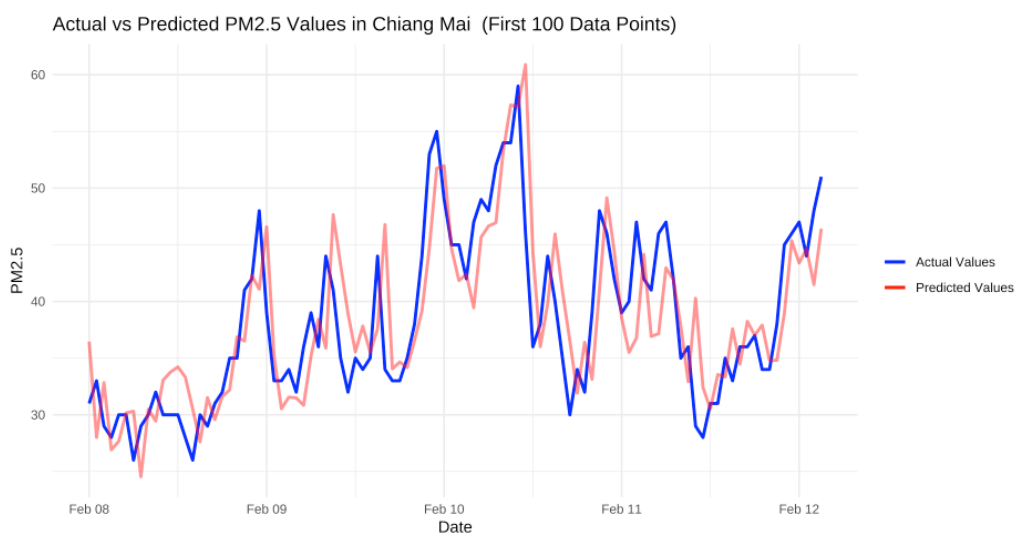


รูปที่ 4.14: กราฟแสดงฮิสโตแกรมของความคลาดเคลื่อนของแบบจำลอง

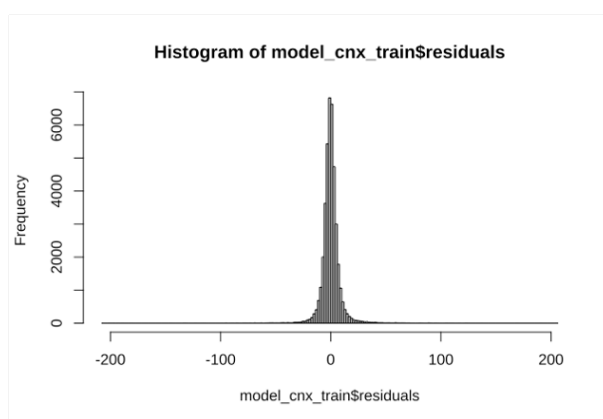
4.2.2 แบบจำลองในกรณีที่ใช้การอนุมานตัวแปรที่บางสถานีในพื้นที่นั้นไม่ได้เก็บข้อมูลมา

ตารางค่าสัมประสิทธิ์ในแบบจำลองของกรณีนี้แสดงดังตารางที่ 4.5 และ 4.6 และเช่นเดียวกับ 4.1 เราได้ทำการวัดประสิทธิภาพของแบบจำลอง ได้ผลดังต่อไปนี้

ประสิทธิภาพของแบบจำลองในแอ่งเชียงใหม่และลำพูน



รูปที่ 4.15: กราฟระหว่างค่าจริงและค่าที่แบบจำลองทำนายค่าของ PM 2.5 ระหว่างวันที่ 8 กุมภาพันธ์ 2566 ถึง 12 กุมภาพันธ์ 2566



รูปที่ 4.16: กราฟแสดงฮิสโตแกรมของความคลาดเคลื่อนของแบบจำลอง

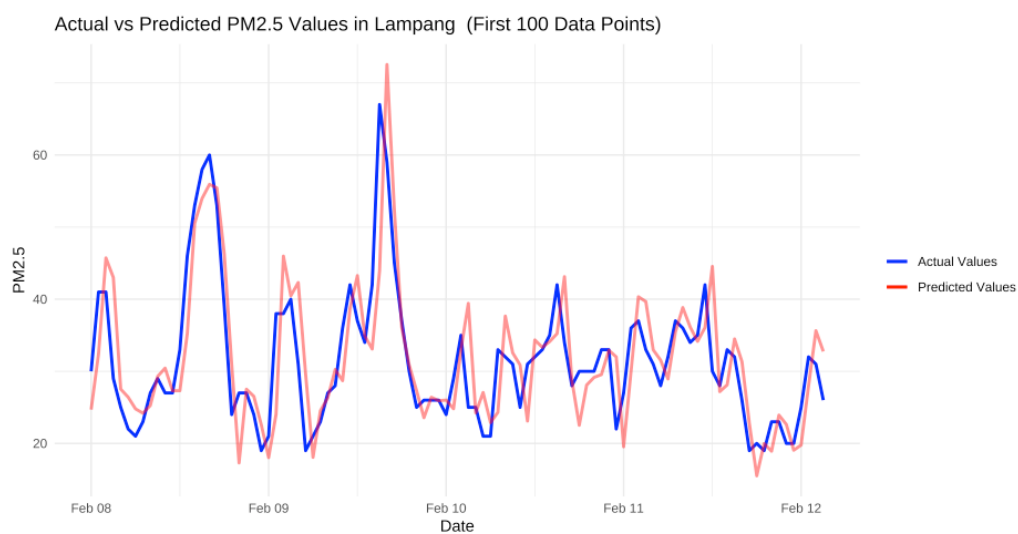
Parameters	intercept	PM2.5 (ug/m^3)					Windspeed (m/s)						Temperature (C)					
		t-1	t-2	t-3	t-4	t-5	t-1	t-2	t-3	t-4	t-5	t-6	t-1	t-2	t-3	t-4	t-5	t-6
Zone 1	21.26	1.08	-0.17	0.04			-2.29						-0.31	-0.51	0.04	0.24	0.22	0.21
Zone 3	1.78	1.25	-0.34	-0.01	-0.01	-0.09	-0.96	0.02	0.38				-0.01	-0.33	0.02	0.08	0.10	0.13
Zone 6	207.76	1.33	-0.45	0.12	-0.07	0.04	-1.64						-0.2					

ตารางที่ 4.5: ตารางแสดงค่าสัมประสิทธิ์ของ ค่าคงที่ในแบบจำลอง PM 2.5 ความเร็วลม และ อุณหภูมิ

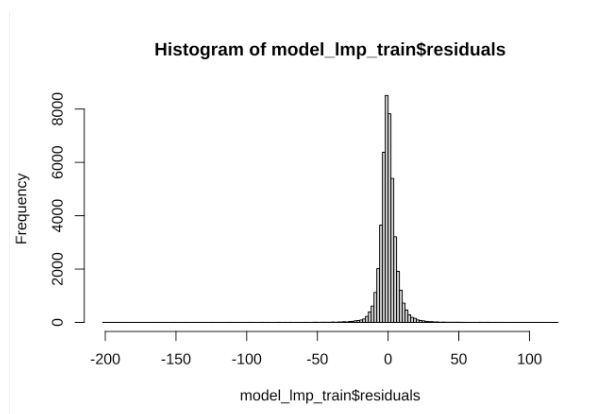
Parameters	Humidity (% RH)						Rain (mm)						Pressure (mmHg)					
	t-1	t-2	t-3	t-4	t-5	t-6	t-1	t-2	t-3	t-4	t-5	t-6	t-1	t-2	t-3	t-4	t-5	t-6
Zone 1	-0.04	-0.09	-0.01	0.01	-0.08	0.09							0.84	-0.1				
Zone 3							-0.12	-0.23	-0.12	-0.17								
Zone 6	-0.02	-0.09	0.04				-0.02	0.04					0.52	0.08	-0.61	-0.62	-0.01	0.36

ตารางที่ 4.6: ตารางแสดงค่าสัมประสิทธิ์ของความชื้นในอากาศ ปริมาณน้ำฝน และความกดอากาศ

ประสิทธิภาพของแบบจำลองในแอ่งลำปาง

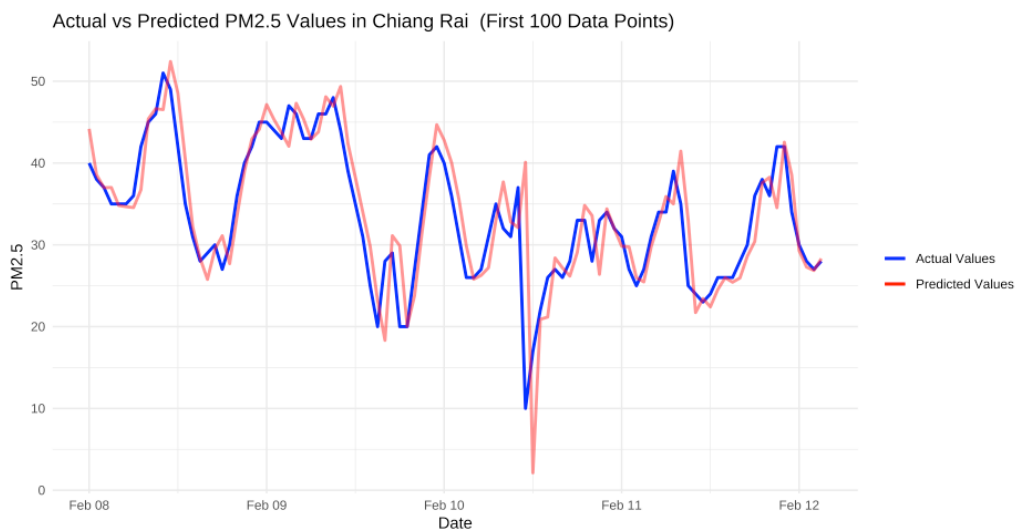


รูปที่ 4.17: กราฟระหว่างค่าจริงและค่าที่แบบจำลองทำนายค่าของ PM 2.5 ระหว่างวันที่ 8 กุมภาพันธ์ 2566 ถึง 12 กุมภาพันธ์ 2566

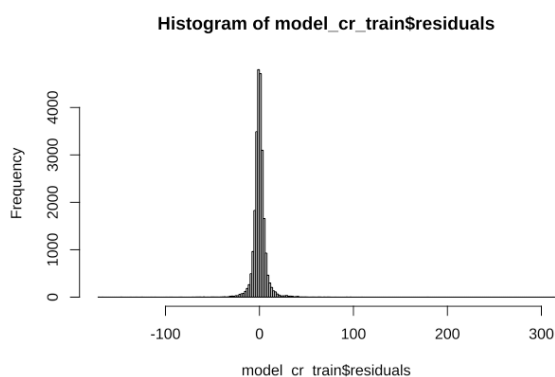


รูปที่ 4.18: กราฟแสดงฮิสโตแกรมของความคลาดเคลื่อนของแบบจำลอง

ประสิทธิภาพของแบบจำลองในแอ่งเชียงใหม่



รูปที่ 4.19: กราฟระหว่างค่าจริงและค่าที่แบบจำลองทำนายค่าของ PM 2.5 ระหว่างวันที่ 8 กุมภาพันธ์ 2566 ถึง 12 กุมภาพันธ์ 2566



รูปที่ 4.20: กราฟแสดงฮิสโตแกรมของความคลาดเคลื่อนของแบบจำลอง

4.3 ความแม่นยำของแบบจำลอง

หลังจากที่เราได้แบบจำลองสำหรับทำนายค่าฝุ่น PM 2.5 ในแต่ละพื้นที่แล้ว เราได้วัดประสิทธิภาพและความแม่นยำของแบบจำลองด้วยค่า RMSE ในแต่ละช่วงของการทำนายค่าฝุ่น PM 2.5 โดยค่าแต่ละช่วงอ้างอิงมาจากกรมควบคุมมลพิษที่ระบุความรุนแรงของฝุ่น PM 2.5 เรียงตามระดับความรุนแรงจากน้อยไปมาก ดังช่วงต่อไปนี้ $0-37 \mu\text{g}/\text{m}^3$, $37-50 \mu\text{g}/\text{m}^3$, $50-90 \mu\text{g}/\text{m}^3$ และมากกว่า $90 \mu\text{g}/\text{m}^3$ โดยเราจะนำแบบจำลองที่เราได้มาเปรียบเทียบกับแบบจำลอง Autoregressive (AR) เพื่อให้เห็นประสิทธิภาพของตัวโมเดลชัดเจนยิ่งขึ้น ดังตารางที่ 4.7

พื้นที่	Model	0 – 37 $\mu g/m^3$	37 – 50 $\mu g/m^3$	50 – 90 $\mu g/m^3$	> 90 $\mu g/m^3$
เชียงใหม่และลำพูน	AR	26.35	5.27	22.44	79.72
	ARDL	4.60	6.16	8.59	14.47
พะเยา	AR	21.22	5.27	25.45	101.06
	ARDL	4.59	6.01	8.90	21.21
ลำปาง	AR	23.49	4.12	24.79	79.17
	ARDL	4.12	6.22	8.59	15.24
แพร่	AR	19.38	4.56	28.43	89.71
	ARDL	4.60	7.03	10.90	23.07
น่าน	AR	27.23	4.64	26.02	164.78
	ARDL	4.56	5.83	8.40	34.51
เชียงราย	AR	30.20	11.24	14.75	213.99
	ARDL	4.87	6.55	7.88	18.94
แม่ฮ่องสอน	AR	30.81	10.16	19.09	157.55
	ARDL	4.69	6.08	9.12	34.52

ตารางที่ 4.7: ตารางแสดงค่า rmse ของแบบจำลอง ARDL กับแบบจำลอง AR ในแต่ละช่วงของค่า PM 2.5

จากตารางที่ 4.7 เห็นได้ชัดว่าประสิทธิภาพความแม่นยำที่วัดจาก RMSE นั้น แบบจำลอง ARDL ที่เรานำมาประยุกต์ใช้ในงานนี้ซึ่งใช้ทั้งข้อมูลของ PM 2.5 และข้อมูลตัวแปรทางสภาพอากาศอื่น ๆ นั้นมีความแม่นยำในการทำนายค่าฝุ่น PM 2.5 ในอีกหนึ่งชั่วโมงข้างหน้า มากกว่าแบบจำลอง AR ที่ใช้เพียงแค่ข้อมูล PM 2.5 ในทุก ๆ พื้นที่ในบริเวณภาคเหนือที่เรานำมาวิเคราะห์ในงานนี้

บทที่ 5

สรุปผลการศึกษา (Conclusion)

5.1 สรุปผลการศึกษา

ในการจัดทำงานค้นคว้าอิสระในหัวข้อ แบบจำลองสำหรับคาดการณ์ความหนาแน่นของฝุ่น PM 2.5 บริเวณภาคเหนือของประเทศไทย โดยใช้ตัวแปรทางสภาพอากาศ จากการศึกษาหลักการทำงานของแบบจำลองที่จะใช้ในการทำนายค่า PM 2.5 ในชั่วโมงถัดไป รวมถึงการวิเคราะห์ผลกระทบของตัวแปรทางสภาพอากาศที่ส่งผลต่อความหนาแน่นของฝุ่น PM 2.5 จากค่าสัมประสิทธิ์ของตัวแปรทางสภาพอากาศต่าง ๆ ในแบบจำลอง พบว่า ค่าความหนาแน่นของฝุ่น PM 2.5 ย้อนหลัง นั้นมีนัยสำคัญกับการทำนายค่า PM 2.5 ของแบบจำลอง และส่งผลกระทบต่อค่า PM 2.5 ในอีก 1 ชั่วโมงข้างหน้ามากที่สุด ส่วนปริมาณน้ำฝนนั้นแทบจะไม่มีผลต่อการทำนายค่า PM 2.5 ของแบบจำลอง และส่งผลกระทบต่อค่า PM 2.5 ในอีก 1 ชั่วโมงข้างหน้าน้อยที่สุด

โดยสามารถสร้างตารางที่อธิบายถึงผลกระทบทางบวกหรือทางลบของตัวแปรทางสภาพอากาศต่าง ๆ ที่ส่งผลต่อค่าความหนาแน่นของฝุ่น PM 2.5 ว่าส่งผลมากน้อยเพียงใด ดังตารางที่ 5.3 และ 5.4 โดยแบ่งระดับสีตามระดับนัยสำคัญดังแสดงในตารางที่ 5.1 และ 5.2 ซึ่งระดับนัยสำคัญนั้นคือค่า p-value จากการสร้างแบบจำลองโดยที่

* แทน ระดับนัยสำคัญ 0.05

** แทน ระดับนัยสำคัญ 0.01

*** แทน ระดับนัยสำคัญ 0.001

ตารางที่ 5.1: ส่งผลกระทบทางเดียวกัน

	=	ไม่มีนัยสำคัญ	
	=	*	ระดับนัยสำคัญระดับ 0.05
	=	**	ระดับนัยสำคัญระดับ 0.01
	=	***	ระดับนัยสำคัญระดับ 0.001

ตารางที่ 5.2: ส่งผลกระทบทางตรงข้าม

	=	ไม่มีนัยสำคัญ	
	=	*	ระดับนัยสำคัญระดับ 0.05
	=	**	ระดับนัยสำคัญระดับ 0.01
	=	***	ระดับนัยสำคัญระดับ 0.001

Parameters	PM2.5 (ug/m^3)					Windspeed (m/s)					Temperature (C)					Humidity (% RH)			Rain (mm)	Pressure (mmHg)				
Zone/time	t-1	t-2	t-3	t-4	t-5	t-1	t-2	t-3	t-6	t-1	t-2	t-3	t-6	t-1	t-2	t-3	t-1	t-1	t-3	t-4	t-5	t-6		
Zone 2																								
Zone 4																								
Zone 5																								
Zone 7																								

ตารางที่ 5.3: ตารางแสดงผลกระทบทางบวกหรือทางลบของตัวแปรทางสภาพอากาศต่าง ๆ ที่ส่งผลต่อค่าความหนาแน่นของฝุ่น PM 2.5 ของพื้นที่ 2 4 5 และ 7

Parameters	PM2.5 (ug/m^3)					Windspeed (m/s)					Temperature (C)					Humidity (% RH)					Rain (mm)					Pressure (mmHg)							
Zone/time	t-1	t-2	t-3	t-4	t-5	t-1	t-2	t-3	t-6	t-1	t-2	t-3	t-4	t-5	t-6	t-1	t-2	t-3	t-5	t-6	t-1	t-2	t-3	t-4	t-1	t-2	t-3	t-4	t-1	t-2	t-3	t-4	t-6
Zone 1																																	
Zone 3																																	
Zone 6																																	

ตารางที่ 5.4: ตารางแสดงผลกระทบทางบวกหรือทางลบของตัวแปรทางสภาพอากาศต่าง ๆ ที่ส่งผลต่อค่าความหนาแน่นของฝุ่น PM 2.5 ของพื้นที่ 1 3 และ 6

5.2 อภิปรายผลการศึกษา

การค้นคว้าอิสระในหัวข้อ แบบจำลองสำหรับคาดการณ์ความหนาแน่นของฝุ่น PM 2.5 บริเวณภาคเหนือของประเทศไทย โดยใช้ตัวแปรทางสภาพอากาศ เราได้ประยุกต์ใช้แบบจำลอง Autoregressive Distributed Lag (ARDL) เพื่อใช้ทำนายความหนาแน่นของฝุ่น PM 2.5 ในอีกหนึ่งชั่วโมงข้างหน้าและหาผลกระทบของตัวแปรทางสภาพอากาศต่อความหนาแน่นของฝุ่น PM 2.5 ซึ่งเรานำแบบจำลองที่ได้มาเปรียบเทียบกับแบบจำลอง Autoregressive (AR) เพื่อให้เห็นประสิทธิภาพของแบบจำลองได้ชัดเจนยิ่งขึ้น ซึ่งจากการเปรียบเทียบความแม่นยำด้วยค่า RMSE พบว่าการประยุกต์ใช้แบบจำลอง ARDL มีประสิทธิภาพและความแม่นยำกว่าแบบจำลอง AR ในทุก ๆ พื้นที่ในภาคเหนือที่เรานำมาวิเคราะห์ในงานในครั้งนี้ โดยจุดเด่นของแบบจำลอง ARDL ที่เรานำมาประยุกต์ใช้นั้นคือเราสามารถนำข้อมูลอนุกรมเวลาของตัวแปรทางสภาพอากาศมาวิเคราะห์ร่วมกับข้อมูลอนุกรมเวลาของความหนาแน่นของฝุ่น PM 2.5 ซึ่งจะทำให้เราเข้าใจถึงผลกระทบของตัวแปรทางสภาพอากาศที่ส่งผลต่อความหนาแน่นของฝุ่น PM 2.5 ได้ครอบคลุมมากขึ้น ในทางกลับกันแบบจำลอง AR อาศัยเพียงข้อมูลอนุกรมเวลาของฝุ่น PM 2.5 เพียงตัวแปรเดียวซึ่งมองข้ามปัจจัยสภาพอากาศต่าง ๆ ที่อาจจะส่งผลต่อความหนาแน่นของฝุ่น PM 2.5 โดยแบบจำลองจากงานในครั้งนี้สามารถนำไปประยุกต์ใช้กับการจัดการคุณภาพอากาศและนำไปเตือนประชาชนทั่วไปเมื่อค่าฝุ่นมีแนวโน้มที่จะเพิ่มขึ้นเกินระดับความปลอดภัยที่กรมควบคุมมลพิษได้กำหนดไว้ได้อย่างทันท่วงที

5.3 ข้อเสนอแนะ

ในแบบจำลองเพื่อทำนายค่าฝุ่น PM 2.5 นี้มีวิธีการจัดการข้อมูลที่ขาดหายไปบางส่วนด้วยการตัดข้อมูลในชั่วโมงนั้นออก ไม่ได้มีการคาดคะเนข้อมูลที่หายไปแต่อย่างใดเนื่องจากมีปริมาณข้อมูลที่ขาดหายไปน้อยกว่า 5% แต่ข้อมูลที่ขาดไปอาจจะมีความสำคัญหรือส่งผลกับตัวแบบจำลองได้ รวมไปถึงอาจจะสร้างแบบจำลองเพื่อทำนายค่าฝุ่น PM 2.5 ด้วยแบบจำลองอื่น ๆ และเปรียบเทียบความแม่นยำกับแบบจำลองนี้ โดยใช้ข้อมูลเดียวกัน

บรรณานุกรม

- [1] เกศรินทร์ บุญเรือง. (2554). การวิเคราะห์ปัจจัยที่มีผลกระทบต่ออัตราผลตอบแทนของหลักทรัพย์ในกลุ่มการท่องเที่ยวและสถานการณ์ของตลาดหลักทรัพย์แห่งประเทศไทย โดยใช้วิธีเออาร์ดีแอล [Analysis of factors affecting the rate of return of securities in the Tourism Business Sector, the Stock Exchange of Thailand Using ARDL Method]. มหาวิทยาลัยเชียงใหม่.
- [2] ศิริลักษณ์ เล็กสมบูรณ์. (2531). การวิเคราะห์อนุกรมเวลา: ตำราประกอบการเรียนวิชาอนุกรมเวลาเลขดัชนี. คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒมหาสารคาม.
- [3] Choonghyun Ryu (2021). dlookr: Tools for Data Diagnosis, Exploration, Transformation. R package version 0.4.0. <https://CRAN.R-project.org/package=dlookr>
- [4] Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, 74(366a).
- [5] Hassler, U., & Wolters, J. (2005). Autoregressive distributed lag models and cointegration. *Diskussionsbeiträge*, No. 2005/22. Freie Universität Berlin, Fachbereich Wirtschaftswissenschaft.
- [6] Hdammm. (2012, January 16). Topography of northern Thailand with corrected position and corrected name of Khun Tan Range. https://commons.wikimedia.org/wiki/File:Topography_of_northern_Thailand.png
- [7] Inkeaw, P. (2023). Introduction to Data Science [วิทยาการข้อมูลเบื้องต้น]. Chiang Mai University Press.
- [8] Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, 9(1), 207-218.
- [9] Nkoro, E., & Uko, A. K. (2016). Autoregressive Distributed Lag (ARDL) cointegration technique: application and interpretation. *Journal of Statistical and Econometric Methods*, 5(4), 63-91.
- [10] Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, 8(1), 3-15.
- [11] Siegel, A. F., & Wagner, M. R. (2022). *Practical Business Statistics* (8th ed.).

- [12] Watakajaturaphon, S., & Phetpradap, P. (2020, November). PM 2.5 Problem in Chiang Mai, Thailand: The Application of Maximizing Expected Utility with Imbalanced Loss Functions. In International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (pp. 72-83). Springer International Publishing.

ภาคผนวก

หมายเหตุ สำหรับแบบจำลองในพื้นที่อื่น ๆ จะใช้ code เดียวกันต่างกันแค่ตอน import ข้อมูล จึงขอละไว้ในฐานที่เข้าใจ

โปรแกรมสำหรับการสร้างแบบจำลองสำหรับคาดการณ์ความหนาแน่นของฝุ่น PM 2.5 บริเวณภาคเหนือของประเทศไทย โดยใช้ตัวแปรทางสภาพอากาศ ในพื้นที่ที่ 1

```
1 library(vars)
2 library(ARDL)
3 library(tseries)
4 library(stats)
5 library(ggpubr)
6 library(readxl)
7 library(lmtest)
8 library(dynlm)
9 library(ggplot2)
10 library(dlookr)
11 library(visdat)
12 library(plotly)
13 library(missRanger)
14 library(tidyverse)
15 library(naniar)
16 library(mice)
17 library(imputeTS)
18
19 # import data
20 # Area 1 (CNX)
21 data_35 <- read_excel("~/Desktop/206499/data_update2023/(35t)
                .xlsx")
22 data_36 <- read_excel("~/Desktop/206499/data_update2023/(36t)
                (1).xlsx")
```

```

23 data_68 <- read_excel("~/Desktop/206499/data_update2023/(68t)
    .xlsx")
24
25 # clean data
26 # delete first row of each dataframe
27 data_35 <- data_35[-1,]
28 data_36 <- data_36[-1,]
29 data_68 <- data_68[-1,]
30
31 # Modify the name of the first column for each dataframe
32 colnames(data_35)[1] <- "Date"
33 colnames(data_36)[1] <- "Date"
34 colnames(data_68)[1] <- "Date"
35
36 # Filter function to select rows based on month Jan-May
37 filter_months_string <- function(df) {
38   # Extract the month part of the date (assuming format is Y/M/
    D)
39   month <- substr(df$Date, 3, 4)
40   # Filter rows where month is between "12" (December) and "05"
    (May)
41   return(df[month %in% c("12", "01", "02", "03", "04", "05"), ])
42 }
43
44 data_35 <- filter_months_string(data_35)
45 data_36 <- filter_months_string(data_36)
46 data_68 <- filter_months_string(data_68)
47
48
49 #implies data from full data column
50 #rain
51 data_36$RAIN <- data_68$Rain
52 #BP
53 data_35$BP <- data_68$Pressure
54 data_36$BP <- data_68$Pressure

```

```

55
56 # select only some column from each df
57 data_35 <- data_35[, c(13,10,11,12,15,14)]
58 data_36 <- data_36[, c(12,10,11,14,15,13)]
59 data_68 <- data_68[, c(10,12,13,15,14,16)]
60
61 #Change name
62 colnames(data_35) <- colnames(data_68)
63 colnames(data_36) <- colnames(data_68)
64
65 #Merge 2 df in verticle
66 data_cnx <- rbind(data_35,data_36)
67 data_cnx <- rbind(data_cnx,data_68)
68
69
70 colnames(data_cnx)[1] <- "WS"
71 colnames(data_cnx)[2] <- "TEMP"
72 colnames(data_cnx)[3] <- "RH"
73 colnames(data_cnx)[4] <- "RAIN"
74 colnames(data_cnx)[5] <- "BP"
75 colnames(data_cnx)[6] <- "PM2.5"
76
77 #change all variables to numeric
78 data_cnx$WS <- as.numeric(data_cnx$WS)
79 data_cnx$TEMP <- as.numeric(data_cnx$TEMP)
80 data_cnx$RH <- as.numeric(data_cnx$RH)
81 data_cnx$RAIN <- as.numeric(data_cnx$RAIN)
82 data_cnx$BP <- as.numeric(data_cnx$BP)
83 data_cnx$PM2.5 <- as.numeric(data_cnx$PM2.5)
84
85 #Handle Missing Values in Time Series For Beginners
86 n_miss(data_cnx)
87 data_cnx %>% is.na() %>% colSums()
88
89 # Display the number of missing values in a dataset
90 plot_na_pareto(data_cnx, only_na = TRUE)

```

```

91
92 data_cnx <- na.omit(data_cnx)
93
94 # Shift the time (we want to forecast pm2.5 by using all past
    values)
95 data_cnx$PM2.5 <- c(data_cnx$PM2.5[-1], NA)
96 data_cnx <- na.omit(data_cnx)
97
98 #####
99
100 # 1. Check for stationary
101 adf.test(data_cnx$WS)
102 adf.test(data_cnx$TEMP)
103 adf.test(data_cnx$RH)
104 adf.test(data_cnx$BP)
105 adf.test(data_cnx$RAIN)
106 adf.test(data_cnx$PM2.5)
107
108 # 2. Select optimal lag
109 varselect_ws <- VARselect(data_cnx[, "WS"], lag.max = 5, type =
    "both")
110 varselect_temp <- VARselect(data_cnx[, "TEMP"], lag.max = 5,
    type = "both")
111 varselect_rh <- VARselect(data_cnx[, "RH"], lag.max = 5, type =
    "both")
112 varselect_bp <- VARselect(data_cnx[, "BP"], lag.max = 5, type =
    "both")
113 varselect_rain <- VARselect(data_cnx[, "RAIN"], lag.max = 5,
    type = "both")
114 varselect_pm25 <- VARselect(data_cnx[, "PM2.5"], lag.max = 5,
    type = "both")
115
116 # See each lag of each variable (all 5 lags)
117 print(varselect_ws$selection)
118 print(varselect_temp$selection)
119 print(varselect_rh$selection)

```

```

120 print(vartest_bp$selection)
121 print(vartest_rain$selection)
122 print(vartest_pm25$selection)
123
124
125 # Using train/test data
126 # Assuming you choose an 80/20 split
127 split_index <- round(nrow(data_cnx) * 0.8)
128
129 # Splitting the data
130 train_data <- data_cnx[1:split_index, ]
131 test_data <- data_cnx[(split_index + 1):nrow(data_cnx), ]
132
133 start.time <- Sys.time()
134 model_cnx_train <- ardl(PM2.5 ~ WS + TEMP + RH + BP , data=
    train_data, order = c(3,0,5,5,1))
135 summary(model_cnx_train)
136 end.time <- Sys.time()
137 time.taken <- round(end.time - start.time,2)
138 time.taken
139
140 plot(model_cnx_train)
141 hist(model_cnx_train$residuals,breaks = 200)
142 mean(model_cnx_train$residuals)
143 summary(model_cnx_train$residuals)
144
145 # Calculate the number of over-predictions
146 num_over_predictions <- sum(model_cnx_train$residuals > 0)
147
148 # Calculate the number of under-predictions
149 num_under_predictions <- sum(model_cnx_train$residuals < 0)
150
151 # Total number of predictions
152 total_predictions <- num_under_predictions + num_over_
    predictions
153

```

```

154 # Proportion of over-predictions
155 prop_over_predictions <- num_over_predictions / total_
    predictions
156
157 # Proportion of under-predictions
158 prop_under_predictions <- num_under_predictions / total_
    predictions
159
160 # Output the proportions
161 print(paste("Proportion of Over-Predictions:", prop_over_
    predictions))
162 print(paste("Proportion of Under-Predictions:", prop_under_
    predictions))
163
164
165 coefficients <- c(
166 `(Intercept)` = 21.257673,
167 `L(PM2.5, 1)` = 1.082133,
168 `L(PM2.5, 2)` = -0.172711,
169 `L(PM2.5, 3)` = 0.049316,
170 `WS` = -2.293018,
171 `TEMP` = -0.317585,
172 `L(TEMP, 1)` = -0.511763,
173 `L(TEMP, 2)` = 0.041584,
174 `L(TEMP, 3)` = 0.246477,
175 `L(TEMP, 4)` = 0.223715,
176 `L(TEMP, 5)` = 0.219272,
177 `RH` = -0.040326,
178 `L(RH, 1)` = -0.088466,
179 `L(RH, 2)` = -0.011717,
180 `L(RH, 3)` = 0.015163,
181 `L(RH, 4)` = -0.008761,
182 `L(RH, 5)` = 0.095013,
183 `BP` = 0.084342,
184 `L(BP, 1)` = -0.102619
185 )

```

```

186
187 # Ensure that the lagged variables are the same length as the
    test_data
188 for (lag in 1:5) {
189   # Check if the variable has the lag according to the model
    output
190   if(paste0("L(PM2.5, ", lag, ")") %in% names(coefficients)) {
191     test_data[paste0("PM2.5_lag", lag)] <- c(rep(NA, lag), head(
      (test_data$PM2.5, -lag))
192   }
193   if(paste0("L(WS, ", lag, ")") %in% names(coefficients)) {
194     test_data[paste0("WS_lag", lag)] <- c(rep(NA, lag), head(
      test_data$WS, -lag))
195   }
196   if(paste0("L(TEMP, ", lag, ")") %in% names(coefficients)) {
197     test_data[paste0("TEMP_lag", lag)] <- c(rep(NA, lag), head(
      test_data$TEMP, -lag))
198   }
199   if(paste0("L(RH, ", lag, ")") %in% names(coefficients)) {
200     test_data[paste0("RH_lag", lag)] <- c(rep(NA, lag), head(
      test_data$RH, -lag))
201   }
202 }
203 # Remove NA values that were introduced by shifting for lags
204 test_data <- na.omit(test_data)
205
206 # Use the actual coefficient names as strings within the `with`
    function
207 predicted_PM2.5 <- with(test_data,
208   coefficients["(Intercept)"] +
209   coefficients["L(PM2.5, 1)"] * lag(PM2.5, 1) +
210   coefficients["L(PM2.5, 2)"] * lag(PM2.5, 2) +
211   coefficients["L(PM2.5, 3)"] * lag(PM2.5, 3) +
212   coefficients["WS"] * WS +
213   coefficients["TEMP"] * TEMP +
214   coefficients["L(TEMP, 1)"] * lag(TEMP, 1) +

```

```

215 coefficients["L(TEMP, 2)"] * lag(TEMP, 2) +
216 coefficients["L(TEMP, 3)"] * lag(TEMP, 3) +
217 coefficients["L(TEMP, 4)"] * lag(TEMP, 4) +
218 coefficients["L(TEMP, 5)"] * lag(TEMP, 5) +
219 coefficients["RH"] * RH +
220 coefficients["L(RH, 1)"] * lag(RH, 1) +
221 coefficients["L(RH, 2)"] * lag(RH, 2) +
222 coefficients["L(RH, 3)"] * lag(RH, 3) +
223 coefficients["L(RH, 4)"] * lag(RH, 4) +
224 coefficients["L(RH, 5)"] * lag(RH, 5) +
225 coefficients["BP"] * BP +
226 coefficients["L(BP, 1)"] * lag(BP, 1)
227 )
228
229
230 # Add the predicted values to your test_data dataframe
231 test_data$predicted_PM2.5 <- predicted_PM2.5
232
233 test_data <- na.omit(test_data)
234
235 # Calculate performance metrics on the test data
236 mae_test <- mean(abs(test_data$predicted_PM2.5[6:length(test_
    data$predicted_PM2.5)] - test_data$PM2.5[6:length(test_data$
    predicted_PM2.5)]), na.rm = TRUE)
237 # Exclude the first five observations as they do not have
    lagged values
238 actual_values <- test_data$PM2.5
239 predicted_values <- test_data$predicted_PM2.5
240
241 # Calculate MAPE
242 mape_test <- mean(abs((actual_values - predicted_values) /
    actual_values), na.rm = TRUE) * 100
243 rmse_test <- sqrt(mean((predicted_values - actual_values)^2, na
    .rm = TRUE))
244
245 print(paste("Test MAE:", mae_test))

```

```

246 print(paste("MAPE on the test data:", mape_test))
247 print(paste("Test RMSE:", rmse_test))
248
249
250 # Evaluate by range of PM 2.5
251 test_data$accuracy_range <- cut(test_data$predicted_PM2.5,
252 breaks = c(-Inf, 37, 50, 90, Inf),
253 labels = c("0-37", "37-50", "50-90", "more than 90"),
254 include.lowest = TRUE)
255 # Create subsets for each accuracy range
256 range_0_37 <- subset(test_data, accuracy_range == "0-37")
257 range_37_50 <- subset(test_data, accuracy_range == "37-50")
258 range_50_90 <- subset(test_data, accuracy_range == "50-90")
259 range_more_than_90 <- subset(test_data, accuracy_range == "more
    than 90")
260
261 # Function to calculate MAE
262 calculate_mae <- function(predicted, actual) {
263   return(mean(abs(predicted - actual), na.rm = TRUE))
264 }
265
266 # Function to calculate MAPE
267 calculate_mape <- function(predicted, actual) {
268   return(mean(abs((actual - predicted) / actual) * 100, na.rm =
    TRUE))
269 }
270
271 # Calculate MAE, MAPE, and RMSE for each range
272 mae_0_37 <- calculate_mae(range_0_37$predicted_PM2.5, range_0_
    37$PM2.5)
273 mape_0_37 <- calculate_mape(range_0_37$predicted_PM2.5, range_0
    _37$PM2.5)
274 rmse_0_37 <- sqrt(mean((range_0_37$predicted_PM2.5 - range_0_37
    $PM2.5)^2, na.rm = TRUE))
275

```

```

276 mae_37_50 <- calculate_mae(range_37_50$predicted_PM2.5, range_
    37_50$PM2.5)
277 mape_37_50 <- calculate_mape(range_37_50$predicted_PM2.5, range
    _37_50$PM2.5)
278 rmse_37_50 <- sqrt(mean((range_37_50$predicted_PM2.5 - range_37
    _50$PM2.5)^2, na.rm = TRUE))
279
280 mae_50_90 <- calculate_mae(range_50_90$predicted_PM2.5, range_
    50_90$PM2.5)
281 mape_50_90 <- calculate_mape(range_50_90$predicted_PM2.5, range
    _50_90$PM2.5)
282 rmse_50_90 <- sqrt(mean((range_50_90$predicted_PM2.5 - range_50
    _90$PM2.5)^2, na.rm = TRUE))
283
284 mae_more_than_90 <- calculate_mae(range_more_than_90$predicted_
    PM2.5, range_more_than_90$PM2.5)
285 mape_more_than_90 <- calculate_mape(range_more_than_90$
    predicted_PM2.5, range_more_than_90$PM2.5)
286 rmse_more_than_90 <- sqrt(mean((range_more_than_90$predicted_
    PM2.5 - range_more_than_90$PM2.5)^2, na.rm = TRUE))
287
288 # Print or use the MAE, MAPE, and RMSE values as needed
289 print(paste("MAE for 0-37 accurate range:", mae_0_37))
290 print(paste("MAPE for 0-37 accurate range:", mape_0_37))
291 print(paste("RMSE for 0-37 accurate range:", rmse_0_37))
292
293 print(paste("MAE for 37-50 accurate range:", mae_37_50))
294 print(paste("MAPE for 37-50 accurate range:", mape_37_50))
295 print(paste("RMSE for 37-50 accurate range:", rmse_37_50))
296
297 print(paste("MAE for 50-90 accurate range:", mae_50_90))
298 print(paste("MAPE for 50-90 accurate range:", mape_50_90))
299 print(paste("RMSE for 50-90 accurate range:", rmse_50_90))
300
301 print(paste("MAE for more than 90 accurate range:", mae_more_
    than_90))

```

```

302 print(paste("MAPE for more than 90 accurate range:", mape_more_
      than_90))
303 print(paste("RMSE for more than 90 accurate range:", rmse_more_
      than_90))
304
305
306 # ggplot2 scatter plot of predicted vs actual PM2.5 values
307 ggplot(test_data, aes(x = predicted_PM2.5, y = PM2.5)) +
308 geom_point(color = "black") +
309 geom_abline(intercept = 0, slope = 1, color = "red", linetype =
      "dashed") +
310 labs(x = "Predicted PM2.5", y = "Actual PM2.5", title = "
      Predicted vs. Actual PM2.5") +
311 theme_minimal()
312
313 #plot in same graph
314 # Add an index column to your test_data for plotting
315 test_data$Index <- 1:nrow(test_data)
316 ggplot(test_data, aes(x = Index)) +
317 geom_line(aes(y = PM2.5, color = "Actual Values"), linewidth =
      1) +
318 geom_line(aes(y = predicted_PM2.5, color = "Predicted Values"),
      linewidth = 1,alpha=0.5) +
319 scale_color_manual(values = c("Actual Values" = "blue", "
      Predicted Values" = "red")) +
320 labs(x = "Index", y = "PM2.5", title = "Actual vs Predicted PM2
      .5 Values") +
321 theme_minimal() +
322 theme(legend.title = element_blank())

```

Listing 5.1: R code for PM 2.5 forecasting model



Prediction model of PM 2.5 density in Northern region of Thailand using atmospheric parameters

Sila Chalee, and Asst. Prof. Dr. Parkpoom Phetpradap
Department of Mathematics, Faculty of science, Chiang Mai university



Abstract

Air quality, especially fine particulate matter (PM 2.5), has emerged as a critical environmental and public health concern globally. This is particularly evident in Northern Thailand, where PM 2.5 frequently exceeds safe levels, posing substantial health risks during dry season. This study aims to develop predictive models for PM 2.5 density in Northern Thailand using atmospheric parameters. Employing an Autoregressive Distributed Lag (ARDL) model, we integrated six parameters: temperature, humidity, precipitation, wind speed, air pressure, and PM 2.5 itself, to predict PM 2.5 density in the next hour and to reflect the complex interplay of factors influencing PM 2.5 levels. Additionally, by data visualization from the data, the findings provide significant implications for public health, particularly for individuals with respiratory conditions, by identifying threatening months and general temporal windows of elevated risk.

4. Results

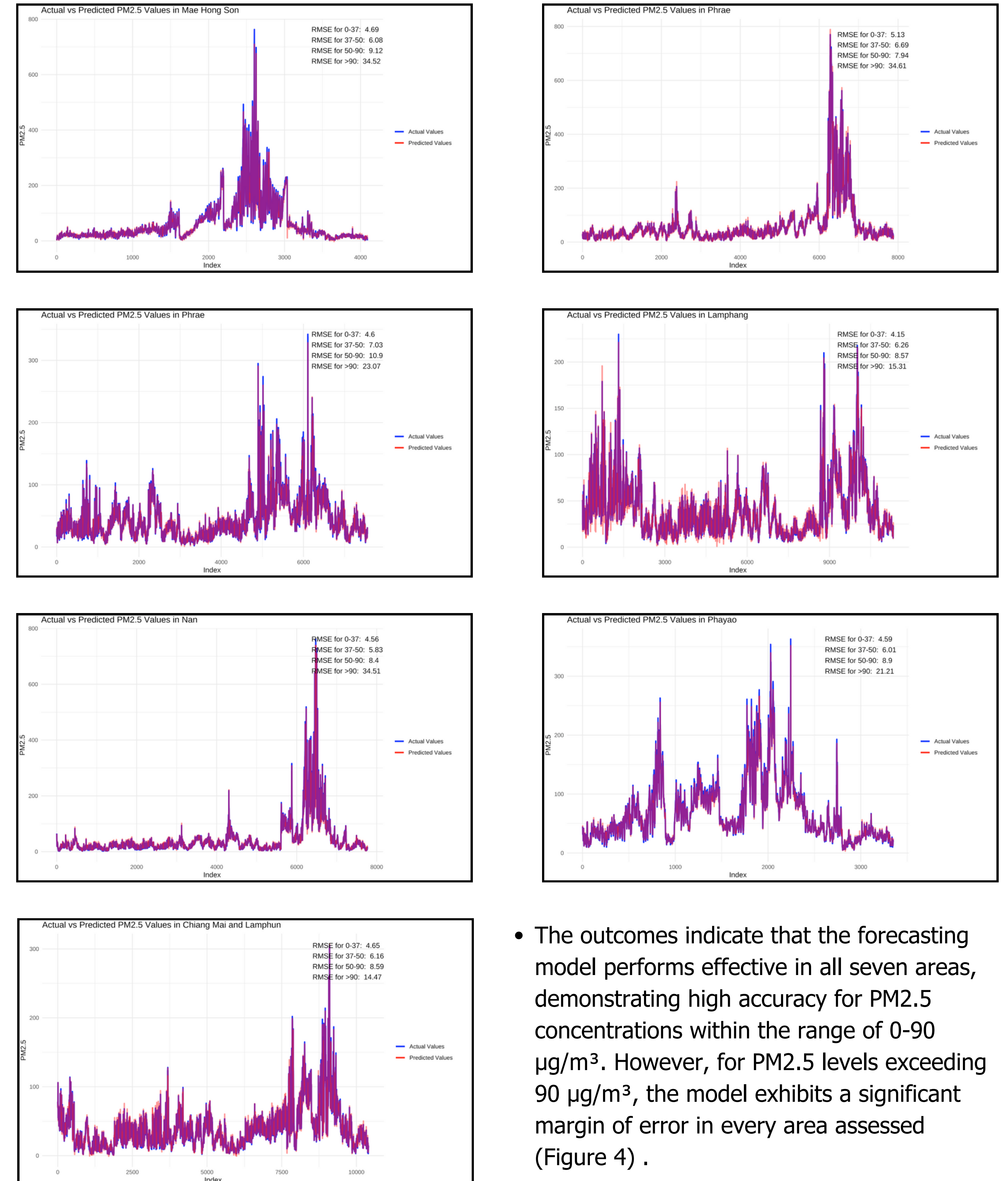


Figure 4: Comparison of Actual vs. Predicted PM2.5 Values (µg/m³) with Root Mean Square Error (RMSE) in each range in each Area

- The outcomes indicate that the forecasting model performs effective in all seven areas, demonstrating high accuracy for PM2.5 concentrations within the range of 0-90 µg/m³. However, for PM2.5 levels exceeding 90 µg/m³, the model exhibits a significant margin of error in every area assessed (Figure 4).

- The table illustrates the coefficients assigned to different parameters within our PM2.5 predictive model, segmented by area and time lag. Positive coefficients (green) suggest a direct relationship, while negative values (red) indicate an inverse relationship with future PM2.5 concentrations.

Parameters	PM2.5 (ug/m³)					Wind speed (m/s)					Temperature (C)				
Zone/Time	t-1	t-2	t-3	t-4	t-5	t-1	t-2	t-3	t-4	t-5	t-1	t-2	t-3	t-4	t-5
Zone 1	0.11	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Zone 2	0.11	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Zone 3	0.11	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Zone 4	0.11	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Zone 5	0.11	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Zone 6	0.11	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Zone 7	0.11	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01

Figure 5: Coefficient Values by Parameter and Time Lag for PM2.5 Prediction Model in each Area

5. Conclusions

In this project on the topic of "Prediction model of PM 2.5 density in Northern region of Thailand using atmospheric parameters" by studying the operational principles of the models used to predict PM 2.5 values in the next hour, including the analysis of the impact of weather variables on PM 2.5 density, it was found that historical PM 2.5 values and wind speed are the most influential factors, with past PM 2.5 levels having a positive impact and wind speed showing a clear negative effect on future PM 2.5 density. These insights suggest that while PM 2.5 tends to accumulate based on previous concentrations, increased wind speed can significantly disperse particulate matter. Temperature, air pressure, and humidity present both positive and negative effects across various time lags, indicating that their impact on PM 2.5 levels is not uniform and may depend on meteorological conditions. Meanwhile, rainfall has a relatively minor influence, underscoring the complexity of predicting PM 2.5 fluctuations and the need for tailored approaches considering local environmental dynamics.

References

- [1] Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. Journal of the American Statistical Association, 74(366a).
- [2] Hassler, U., & Wolters, J. (2005). Autoregressive distributed lag models and coin- tegration. Diskussionsbeiträge, No. 2005/22. Freie Universität Berlin, Fachbereich Wirtschaftswissenschaft.
- [3] Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time Series Missing Value Imputation in R. The R Journal, 9(1), 207-218.
- [4] Nkoro, E., & Uko, A. K. (2016). Autoregressive Distributed Lag (ARDL) cointegration technique: application and interpretation. Journal of Statistical and Econometric Methods, 5(4), 63-91.
- [5] Watakajaturaphon, S., & Phetpradap, P. (2020, November). PM 2.5 Problem in Chiang Mai, Thailand: The Application of Maximizing Expected Utility with Imbalanced Loss Functions. In International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (pp. 72-83).

1. Introduction

In recent years, air quality especially particulate matter less than 2.5 micrometers in diameter (PM2.5) has emerged as a critical environmental issue, significantly impacting human health, economic conditions, and ecological balance. This issue is especially prevalent in the northern regions of Thailand, where PM2.5 levels often exceed safe thresholds, posing significant health risks to the population. This research focuses on developing a model by adapting the Autoregressive Distributed Lag (ARDL) approach to predict the density of PM2.5 particles one hour ahead and to analyze various weather-related variables that influence PM2.5 density in Northern Thailand.

2. Study Area & Data

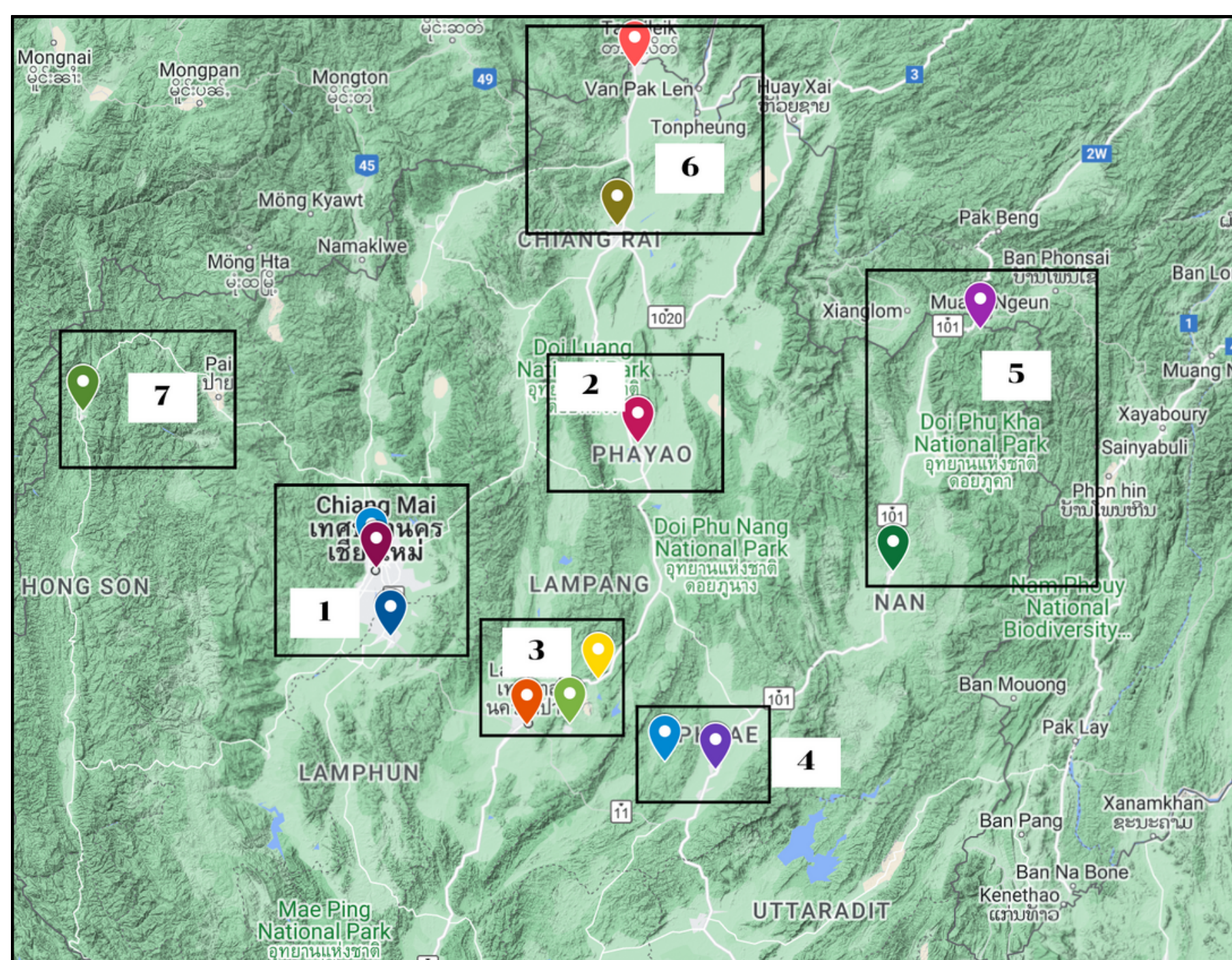


Figure 1: Location of study areas

- We collect Atmospheric data from 14 Weather stations. By data analysis, we can separate the area of study into 7 areas consisting of

 1. Chiang Mai and Lamphun
 2. Phayao
 3. Lampang
 4. Phrae
 5. Nan
 6. Chiang Rai
 7. Mae Hong Son

- We use atmospheric time series data from 1 Jan 2019 to 30 June 2023.
- The data is collected from Ministry of Pollution Control, Ministry of Natural Resources and Environment.
- We focus the period of study to be from December to May from each year, where PM 2.5 level is above the safety level standard.

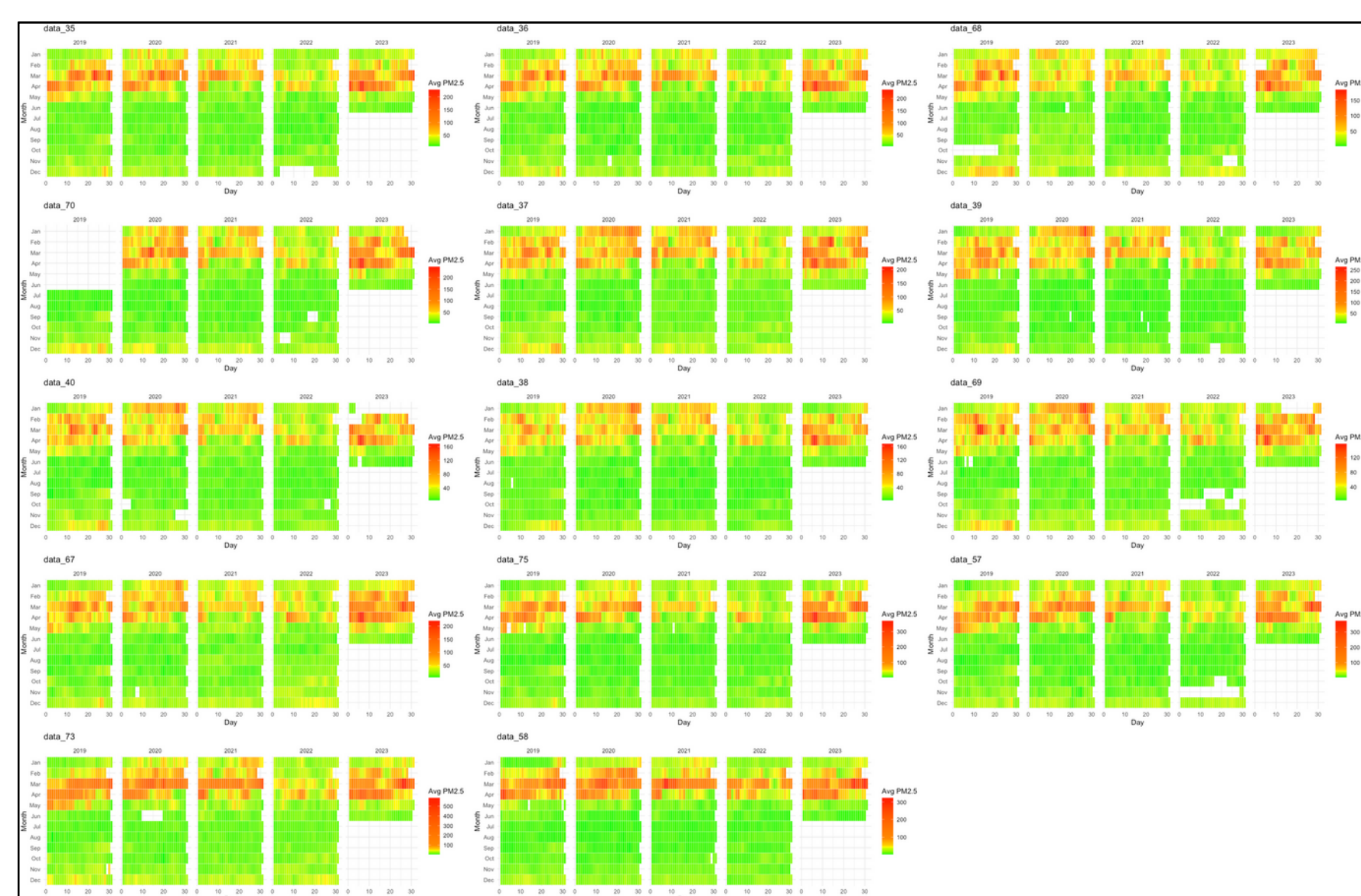


Figure 2: Heatmap representation of daily average PM2.5 density across different months and years in each stations

3. Methodology

We apply the adapted Autoregressive Distributed Lag (ARDL) model to forecast PM 2.5 density in the next hour, and investigate the dynamic relationship between PM 2.5 density, and 5 environmental factors, over time. The adapted ARDL model captures lagged effects, making it possible to assess the influence of past values of both the dependent variable and the independent variables on current PM 2.5 levels.

$$Y_t = \alpha + \sum_{i=1}^p \beta_i Y_{t-i} + \sum_{j=1}^{q_1} \gamma_j X_{1,t-j} + \sum_{j=1}^{q_2} \delta_j X_{2,t-j} + \sum_{j=1}^{q_3} \phi_j X_{3,t-j} + \sum_{j=1}^{q_4} \sigma_j X_{4,t-j} + \sum_{j=1}^{q_5} \lambda_j X_{5,t-j} + \varepsilon_t$$

where α is the constant term
 Y_t is the value of PM 2.5 at time t
 $X_{1,t}$ is the value of wind speed at time t
 $X_{2,t}$ is the value of air pressure at time t
 $X_{3,t}$ is the value of temperature at time t
 $X_{4,t}$ is the value of precipitation at time t
 $X_{5,t}$ is the value of air humidity at time t

β_i is the coefficient of Y_{t-i}
 γ_j is the coefficient of $X_{1,t-j}$
 δ_j is the coefficient of $X_{2,t-j}$
 ϕ_j is the coefficient of $X_{3,t-j}$
 σ_j is the coefficient of $X_{4,t-j}$
 λ_j is the coefficient of $X_{5,t-j}$
 ε_t is the error term at time t