การเปรียบเทียบวิธีการประมาณค่าสูญหายในการพยากรณ์ความเข้มข้นของ $PM_{2.5}$ ด้วยโครงข่ายประสาทเทียม LSTM

A Comparison of Imputation Methods in Estimating $PM_{2.5}$ Concentrations Using LSTM Neural Networks

ศรีรักษ์ ศรีทองชัย
Sriruk Srithongchai

ภาควิชาวิศวกรรมอุตสาหการและการจัดการ คณะวิศวกรรมศาสตร์และเทคโนโลยีอุตสาหกรรรม
มหาวิทยาลัยศิลปากร วิทยาเขตพระราชวังสนามจันทร์ ต.พระปฐมเจดีย์ อ.เมือง จ.นครปฐม รหัสไปรษณีย์ 73000
srithongchai_s@su.ac.th

บทคัดย่อ

ข้อมูลสูญหายเป็นปัญหาหนึ่งที่พบได้บ่อยในการเก็บรวบรวมข้อมูลเพื่อการวิจัย มักส่งผลกระทบต่อความถูกต้องและความเชื่อมั่นของผลการวิเคราะห์ข้อมูล รวมไปถึงข้อค้นพบและข้อสรุปที่ได้จากการศึกษา วิธีการจัดการความไม่สมบูรณ์ของข้อมูลอาจใช้การกำจัดทิ้งหรือแทนค่าที่สูญหายด้วยการประมาณค่า งานวิจัยนี้นำเสนอวิธีการประมาณค่าข้อมูลสูญหายในการพยากรณ์ความเข้มข้นของ $PM_{2.5}$ ด้วยแบบจำลองโครงข่ายประสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาว (LSTM) จำแนกได้ 4 วิธี ได้แก่ วิธีค่าเฉลี่ย วิธีการประมาณค่าในช่วงแบบเชิงเส้น (LI) วิธีเพื่อนบ้านใกล้สุด (KNN) และวิธีการแทนค่าแบบพหุด้วยสมการลูกโซ่ (MICE) ใช้ค่ารากที่สองความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) และความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) เป็นเครื่องมือวัดความถูกต้องในการทำนายค่าฝุ่น $PM_{2.5}$ ซึ่งเป็นตัวบ่งชี้ที่ใช้เปรียบเทียบประสิทธิภาพของวิธีการจัดการกับข้อมูลสูญหาย โดยข้อมูลจริงที่ทดแทนค่าสูญหายแล้วจะถูกนำไปใช้กับแบบจำลองการพยากรณ์ 4 แบบ ผลการศึกษาพบว่าวิธี KNN (k=11) และ วิธี MICE ให้ค่าความผิดพลาดน้อยที่สุดจึงเป็นวิธีการประมาณค่าสูญหายที่ดีที่สุด ทั้งสองวิธีให้ผลลัพธ์ดีกว่าวิธีค่าเฉลี่ยและวิธี LI และยังแสดงให้เห็นว่าแบบจำลองที่ 2 ให้ความแม่นยำมากที่สุดในทุกวิธีการประมาณค่า

คำสำคัญ: การแทนค่าสูญหาย, ฝุ่น $PM_{2.5}$, โครงข่ายประสาทเทียม, แบบจำลองหน่วยความจำระยะสั้นแบบยาว

Abstract

Missing data is one of the common problems in data collection for research purposes. It often affects the accuracy and reliability of the data analysis results, and the findings and conclusions from the study. Methods for handling incomplete data are deletion or imputation by replacing missing values with estimates. In this study, we propose imputation methods for $PM_{2.5}$ concentration prediction using long short-term memory (LSTM) neural networks. There are four approaches, namely mean, linear interpolation (LI), k-nearest neighbors (KNN) and multiple imputation by chained equations (MICE). The root mean square error (RMSE) and mean absolute error (MAE) were used to evaluate the accuracy of $PM_{2.5}$ predictions, which is an indicator to measure the effectiveness of different imputation methods. Each imputed dataset was applied to four LSTM forecast models. It showed that the KNN (k=11) and MICE methods provided the least errors and therefore were the best imputation techniques. Both gave better results compared to Mean and LI for filling in missing observations. Model 2 had the most accurate forecast among all the data imputation methods.

Keywords: Imputation, $PM_{2.5}$, Neural Network, LSTM

1. บทนำ

มลพิษทางอากาศ [1] คือภาวะอากาศที่ปนเปื้อนด้วยสารพิษต่าง ๆ โดยทางเคมี กายภาพและชีววัตถุ ส่วนใหญ่มีแหล่งกำเนิดมาจากการเผาผลาญของเครื่องยนต์ การใช้ยานพาหนะ กระบวนการผลิตของโรงงาน การเผาพื้นที่เกษตรและไฟป่า สารมลพิษทางอากาศที่อันตรายส่งผลเสียต่อสุขภาพอนามัยของมนุษย์ ได้แก่ ฝุ่นละอองขนาดเล็กและก๊าซพิษ เช่น คาร์บอนมอนอกไซด์ ซัลเฟอร์ไดออกไซด์ ไนโตรเจนไดออกไซด์ และโอโซน เป็นต้นเหตุก่อให้เกิดโรคเกี่ยวกับปอดและระบบทางเดินหายใจซึ่งอาจมีอาการรุนแรงถึงขั้นเสียชีวิต ฝุ่น $PM_{2.5}$ นั้นมีอนุภาคขนาดเล็กไม่เกิน 2.5 ไมครอน ถือเป็นภัยคุกคามด้านสิ่งแวดล้อมระดับโลกที่หลายประเทศกำลังเผชิญอยู่ โดยเฉพาะในเมืองใหญ่ที่แออัดมีการจราจรคับคั่งและเขตประกอบการอุตสาหกรรม ประเทศไทยก็ประสบกับปัญหานี้เช่นเดียวกันเห็นได้ชัดในตอนต้นปี พ.ศ. 2562 เกิดปรากฏการณ์ฝุ่นละอองหมอกควันปกคลุมอย่างหนาแน่น ไม่ว่าจะเป็นบริเวณภาคเหนือ ภาคตะวันออกเฉียงเหนือ กรุงเทพมหานครและปริมณฑล พบว่าค่าฝุ่น $PM_{2.5}$ เกินค่ามาตรฐาน [2] (ค่าเฉลี่ยรายปีและค่าเฉลี่ยใน 24 ชั่วโมงของประเทศไทยกำหนดไว้ที่ 25 และ 50 ไมโครกรัมต่อลูกบาศก์เมตรตามลำดับ [3]) ส่งผลให้สถานศึกษาต้องปิดการเรียนการสอน ประชาชนต่างป้องกันตนเองด้วยการสวมใส่หน้ากากป้องกันฝุ่น และตลอดระยะเวลาหลายปีที่ผ่านมาเว็บไซต์รายงานคุณภาพอากาศ IQAir.com ได้จัดให้เชียงใหม่และกรุงเทพมหานครติดอันดับเมืองที่มีปริมาณมลพิษสูงลำดับต้น ๆ ของโลก [4]

การตรวจวัดฝุ่น $PM_{2.5}$ ในอากาศจึงมีความสำคัญมากเพื่อประโยชน์ในการเฝ้าระวังและแจ้งเตือนภัยสถานการณ์ฝุ่นละอองเตรียมการรับมือ และจัดการแก้ไขปัญหา แต่การตรวจวัดของสถานีตรวจวัดคุณภาพอากาศมีข้อจำกัดด้านงบประมาณและเครื่องมือวัดที่ไม่เพียงพอ ในประเทศไทยมีเครื่องมือตรวจวัดคุณภาพอากาศที่เป็นของหน่วยงานภาครัฐ 138 เครื่อง จำนวนมากกว่า 50 เครื่องอยู่ในพื้นที่กรุงเทพมหานคร [5] การคาดการณ์ความเข้มข้นของฝุ่น $PM_{2.5}$ ผ่านแบบจำลองทางคณิตศาสตร์เป็นอีกทางเลือกหนึ่งซึ่งนิยมใช้กันเนื่องจากเสียเวลาและค่าใช้จ่ายน้อยกว่าการตรวจวัดจริง จากงานวิจัยก่อนหน้านี้ของผู้วิจัย [6] ที่พัฒนาแบบจำลองโครงข่ายใยประสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาว (Long-Short Term Memory: LSTM) เพื่อพยากรณ์ค่า $PM_{2.5}$ ในพื้นที่กรุงเทพมหานคร ได้อาศัยข้อมูลคุณภาพอากาศ (Air Quality) และข้อมูลอุตุนิยมวิทยา (Metrological Data) ใช้การคัดเลือกตัวแปรพยากรณ์เข้าสู่แบบจำลองและแทนค่าสูญหาย (Missing Value) ด้วยค่าเฉลี่ย ซึ่งการสูญหายของข้อมูลมีหลายสาเหตุ เช่น การสอบเทียบเครื่องมือวัด (Calibration Instrument) อุปกรณ์ชำรุด และไฟฟ้าดับ ถึงกระนั้นก็ตามการใช้วิธีการจัดการข้อมูลสูญหายที่ไม่เหมาะสมย่อมมีผลทำให้ผลวิจัยผิดไปจากความเป็นจริง

การวิจัยครั้งนี้จึงได้ทำการศึกษาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหาย (Imputation Method) 4 วิธี ได้แก่ วิธีที่ใช้ค่าเฉลี่ย (Mean) วิธีการประมาณค่าในช่วงแบบเชิงเส้น (Linear Interpolation: LI) วิธีเพื่อนบ้านใกล้สุด (K-Nearest Neighbor: KNN) และวิธีการทดแทนแบบพหุด้วยสมการลูกโซ่ (Multiple Imputation by Chained Equations: MICE) เพื่อใช้เป็นแนวทางในการตัดสินใจเลือกวิธีการจัดการที่เหมาะสม โดยนำข้อมูลที่ใส่ค่าสูญหายแล้วไปใช้กับแบบจำลองที่กล่าวข้างต้น จากนั้นประเมินความแม่นยำในการพยากรณ์ด้วยรากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error: RMSE) และความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error, MAE) วิธีการแทนค่าสูญหายใดที่มีค่า RMSE และ MAE ต่ำสุดจะเป็นวิธีการที่ดีที่สุด [7] ซึ่งพบว่าวิธี KNN และ MICE ให้ผลการประมาณค่าสูญหายได้ดีกว่าวิธีอื่น

2. ข้อมูลสูญหาย

ข้อมูลสูญหายนับว่าเป็นปัญหาที่พบในการทำงานวิจัยที่ยากจะหลีกเลี่ยง ข้อมูลไม่ครบถ้วนขาดหายไปอาจเกิดขึ้นจากปัจจัยหลายประการ เช่น ไม่ได้รับความร่วมมือจากผู้ให้ข้อมูล อุปกรณ์ทำงานผิดปกติ และข้อผิดพลาดในการป้อนข้อมูล เป็นต้น ซึ่งมีโอกาสทำให้ผลการวิเคราะห์เกิดความผิดพลาดขาดความน่าเชื่อถือได้ โดยทั่วไปข้อมูลสูญหายแบ่งออกได้ 3 ประเภทด้วยกัน [8] ประเภทแรกคือการสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing Completely at Random: MCAR) เกิดเมื่อค่าที่ขาดหายไปไม่มีความสัมพันธ์กับข้อมูลใด ๆ เลย ประเภทสองคือการสูญหายแบบสุ่ม (Missing at Random: MAR) หมายถึงส่วนสูญหายขึ้นอยู่กับค่าของข้อมูลอื่นที่ไม่เกี่ยวกับข้อมูลของตนเอง และประเภทสุดท้ายเป็นการสูญหายแบบไม่สุ่ม (Missing Not at Random: MNAR) โดยค่าสูญหายเกี่ยวข้องกับข้อมูลของตนเองหรือข้อมูลตัวอื่น จากการศึกษาของ Pollice และ Lasinio [9] และ Marwala [10] กล่าวว่าคุณภาพอากาศมีกลไกการสูญหายของข้อมูลแบบ MAR สำหรับการดำเนินการกับข้อมูลขาดหายนิยมใช้วิธีการประมาณค่าเพื่อนำไปแทนให้กับค่าที่สูญหาย งานวิจัยด้านมลพิษทางอากาศ เช่น ฝุ่น $PM_{2.5}$ ฝุ่น PM_{10} โอโซน (O_3) และไนโตรเจนไดออกไซด์ (NO_2) มักจะใช้การจำลองข้อมูลให้มีบางส่วนขาดหายและเติมเต็มโดยใช้วิธีต่าง ๆ เพื่อค้นหาวิธีแทนค่าข้อมูลสูญหายที่เหมาะสม [11-14] วิธีการใส่ค่าเฉลี่ยเป็นวิธีการดำเนินการกับข้อมูลขาดหายแบบดั้งเดิมที่ใช้งานง่ายในการเติม

ค่าที่หายไปให้ครบถ้วน [15, 16] Saeipourdizaj, Sarbakhsh และ Gholampour [12] ศึกษาวิธีประมาณค่าสูญหายของข้อมูล PM_{10} และ O_3 โดยนำเสนอ 4 เทคนิค คือ วิธี LI, ค่าเฉลี่ยเคลื่อนที่ (Moving Average: MA), KNN และการจับคู่ค่าเฉลี่ยแบบคาดการณ์ (Predictive Mean Matching: PMM) และเมื่อพิจารณาจากค่า RMSE, MAE และค่าสัมประสิทธิการกำหนด (Coefficient of Determination: R^2) พบว่าวิธี LI, MA และ KNN สามารถนำไปใช้ในการเติมข้อมูลได้อย่างมีประสิทธิภาพ Zainuri, Jemain และ Muda [17] เปรียบเทียบวิธีการจัดการข้อมูลสูญหายกับข้อมูลจำลองคุณภาพอากาศในพื้นที่ตอนกลาง ตอนเหนือและตอนใต้ของคาบสมุทรมลายู พบว่าวิธีค่าคาดหมายสูงสุด (Expectation-Maximization: EM) วิธี KNN และวิธีที่พัฒนามาจากเพื่อนบ้านใกล้สุด (Sequential K-Nearest Neighbor: SKNN) ดีกว่าวิธีค่าเฉลี่ย วิธีมัธยฐาน (Median) และวิธีการแยกค่าเอกฐาน (Singular Value Decomposition: SVD) Ahn, Sun และ Kim [7] ทดสอบประสิทธิภาพวิธีแทนค่าสูญหายโดยอาศัยความแม่นยำของการพยากรณ์ที่ได้จากแบบจำลองโครงข่าย LSTM ใช้ข้อมูลจริงจากหลายแหล่งหลายช่วงเวลา ได้แก่ ข้อมูลคุณภาพอากาศ อุณหภูมิของระบบให้ความร้อน และราคาหุ้น พบว่าวิธี KNN ดีที่สุดเมื่อเทียบกับวิธีอื่น ๆ ทั้งวิธีค่าเฉลี่ย วิธีแทนค่าด้วยข้อมูลสุดท้ายก่อนเกิดการสูญหาย (Last Observation Carried Forward: LOCF) วิธีแทนค่าด้วยข้อมูลล่าสุดหลังจากเกิดค่าสูญหาย (Next Observation Carried Backward: NOCB) วิธี EM และวิธี MICE สำหรับการวิจัยนี้ใช้วิธีการประมาณค่าข้อมูลที่หายไป 4 วิธี ดังต่อไปนี้

2.1 วิธีค่าเฉลี่ย

การหาค่าเฉลี่ย [8] เป็นวิธีแบบดั้งเดิม แทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของตัวแปรที่เกิดค่าสูญหาย

2.2 วิธีการประมาณค่าในช่วงแบบเชิงเส้น (LI)

การประมาณค่าในช่วงแบบเชิงเส้น [18] จะประมาณค่าข้อมูลสูญหายโดยอาศัยความสัมพันธ์เชิงเส้นระหว่างตัวแปรและคำนวณจากข้อมูลที่อยู่ติดกับค่าสูญหายนั้น ตัวอย่างเช่น กำหนดให้ y เป็นข้อมูลสูญหาย x, (x_1,y_1) และ (x_2,y_2) เป็นข้อมูลที่ทราบค่า ถ้า (x,y) อยู่ในช่วงของค่า (x_1,y_1) และ (x_2,y_2) จะได้ว่า

$$y = y_1 + (x - x_1)\frac{(y_2 - y_1)}{(x_2 - x_1)} \qquad (1)$$

2.3 วิธีเพื่อนบ้านใกล้สุด (KNN)

วิธีเพื่อนบ้านใกล้สุด [19] เป็นอัลกอลิทึมหนึ่งที่สามารถใช้งานได้ดีเพื่อเติมเต็มข้อมูลที่ขาดหายไป ใช้หลักการวัดระยะห่างระหว่างข้อมูลขาดหายกับข้อมูลแต่ละตัว แล้วประมาณค่าที่ขาดหายไปด้วยค่าเฉลี่ยของข้อมูลที่ใกล้ที่สุดจำนวน k ตัว นิยมใช้การคำนวณระยะทางแบบยูคลิเดียน (Euclidean Distance) นั่นคือ ถ้า d เป็นระยะห่างระหว่างจุด (x_1,y_1) และ (x_2,y_2) จะได้

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (2)$$

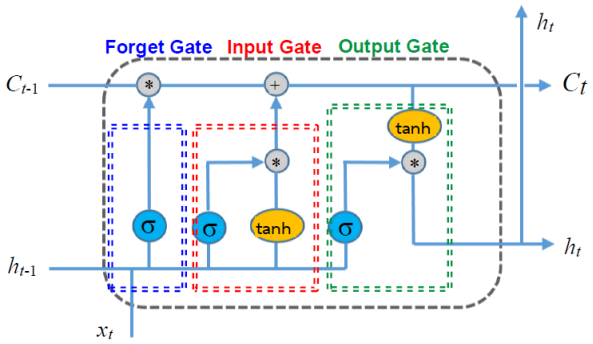
2.4 วิธีการทดแทนแบบพหุด้วยสมการลูกโซ่ (MICE)

หลักการทดแทนแบบพหุด้วยสมการลูกโซ่ [20] จัดการกับข้อมูลสูญหายโดยเติมค่าสูญหายหลายครั้งด้วยการทำงานซ้ำหลาย ๆ รอบ จะหยุดการทำงานตามเงื่อนไขที่ระบุไว้อาจเป็นจำนวนรอบที่แน่นอนหรือค่าความผิดพลาดที่ยอมรับได้ ในแต่ละรอบค่าสูญหายจะถูกประมาณโดยพิจารณาจากค่าของตัวแปรอื่นในชุดข้อมูล อีกทั้งได้นำเอาอัลกอริทึมแบบป่าสุ่ม (Random Forest: RF) เข้ามาร่วมในการสร้างตัวแบบจำลองสำหรับทำนายค่าที่ขาดหาย ซึ่งเป็นเทคนิคการสร้างต้นไม้ตัดสินใจ (Decision Tree) หลาย ๆ ต้น จากการสุ่ม แต่ละต้นใช้ข้อมูลที่แตกต่างกัน แล้วนำผลลัพธ์ทั้งหมดจากหลายต้นมาสรุปรวมกันออกมาเป็นคำตอบที่ต้องการ

3. โครงข่ายประสาทเทียม

แนวคิดเรื่องโครงข่ายประสาทเทียมก่อกำเนิดขึ้นมาครั้งแรกในปี พ.ศ. 2486 (ค.ศ. 1943) โดย McCulloch และ Pitts [21] เสนอโครงข่ายประสาทเทียมอย่างง่ายมีพื้นฐานมาจากศาสตร์ด้านสรีรวิทยาและการทำงานของเซลล์ประสาทในสมองของ

มนุษย์ ซึ่งได้รับความสนใจและเกิดการพัฒนาอย่างต่อเนื่องให้ครอบคลุมปัญหาที่มีความยุ่งยากและซับซ้อน ต่อมาในปี พ.ศ. 2540 (ค.ศ. 1997) โครงข่ายประสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาว (LSTM) ได้ถูกคิดค้นขึ้นโดย Hochreiter และ Schmidhuber [22] เพื่อแก้ปัญหาการสูญหายของเกรเดียน (Vanishing Gradient Problem) ของโครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Network) ด้วยการออกแบบให้โครงข่ายมีความสามารถในการเรียนรู้และจดจำได้นานขึ้น โดยใช้การทำงานของฟังก์ชันต่าง ๆ เพื่ออ่าน เขียนและลบข้อมูล ซึ่งฟังก์ชันเหล่านี้เปรียบเสมือนประตู (Gate) คอยควบคุมการไหลของข้อมูล ประกอบด้วย 3 ประตูได้แก่ ประตูลืม (Forget Gate) ประตูทางเข้า (Input Gate) และประตูทางออก (Output Gate) ลักษณะการทำงานภายในเซลล์ของ LSTM แสดงดังรูปที่ 1 โดยที่ x_t และ h_t คือข้อมูลนำเข้าและข้อมูลส่งออกที่เวลา t ตามลำดับ และ C_t คือสถานะซ่อน *ประตูลืม* มีหน้าที่กำหนดข้อมูลที่เข้ามาว่าควรเก็บไว้หรือทิ้งไป โดยรับข้อมูลปัจจุบันและข้อมูลสถานะซ่อน (Hidden State) ก่อนหน้า ซึ่งจะใช้เป็นแหล่งข้อมูลเข้าสำหรับประตูอื่น ๆ ด้วยเช่นกัน แล้วคำนวณผ่านฟังก์ชันซิกมอยด์ *ประตูทางเข้า* คือประตูที่ใช้เปิดรับข้อมูลที่เข้ามาใหม่แล้วจึงทำการบันทึกข้อมูลลงในเซลล์ของมันเอง การคำนวณจะใช้ฟังก์ชันซิกมอยด์ ขณะเดียวกันจะปรับค่าสถานะเซลล์ด้วยฟังก์ชัน tanh *ประตูทางออก* ใช้บอกค่าสถานะซ่อนอยู่ถัดไปรวมถึงส่งผลลัพธ์ของเซลล์ไปเป็นข้อมูลออก ข้อมูลเข้าถูกส่งผ่านฟังก์ชันซิกมอยด์ และสถานะเซลล์ผ่านฟังก์ชัน tanh



รูปที่ 1 การทำงานของ LSTM (ดัดแปลงจาก [23])

4. การดำเนินงานวิจัย

โครงข่ายประสาทเทียมเลียนแบบระบบการทำงานของเซลล์ประสาทในร่างกาย สามารถเรียนรู้ทำความเข้าใจได้ด้วยตนเองจากข้อมูลตัวอย่างที่ได้ฝึกสอน และทำการประมวลผลข้อมูลอื่นโดยใช้ประสบการณ์ความรู้ที่ผ่านมาในการไขคำตอบของปัญหา ผู้วิจัยได้เลือกใช้โครงข่ายประสาทเทียม LSTM สร้างแบบจำลองเพื่อหาระดับความเข้มข้นของฝุ่นละออง $PM_{2.5}$ โดยทำการคัดเลือกตัวแปร (Variable Selection) ที่มีอิทธิพลมากต่อตัวแปรตามด้วยค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson Correlation Coefficient) ซึ่งใช้หาความสัมพันธ์ระหว่างตัวแปรว่ามีมากน้อยเพียงใด และได้นำเอาความคลาดเคลื่อน RMSE และ MAE ที่ได้จากการพยากรณ์มาเป็นเครื่องมือสำหรับศึกษาเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าสูญหายแต่ละวิธี

4.1 ข้อมูลการวิจัย

ข้อมูลที่ใช้ศึกษาเป็นข้อมูลอนุกรมเวลาเก็บรวบรวมในพื้นที่เขตบางนา กรุงเทพมหานคร โดยกรมควบคุมมลพิษ กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม และกรมอุตุนิยมวิทยา กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม ตั้งแต่เดือนมกราคม พ.ศ. 2560 ถึงเดือนธันวาคม พ.ศ. 2563 ประกอบไปด้วย 2 ส่วน คือข้อมูลคุณภาพอากาศของสถานีตรวจวัดคุณภาพอากาศบางนา ได้แก่ ฝุ่น $PM_{2.5}$ ฝุ่น PM_{10} ก๊าซคาร์บอนมอนอกไซด์ ก๊าซซัลเฟอร์ไดออกไซด์ ก๊าซไนโตรเจนไดออกไซด์และก๊าซโอโซน และข้อมูลอุตุนิยมวิทยาของสถานีอุตุนิยมวิทยาเขตบางนา ได้แก่ อุณหภูมิ ความชื้นสัมพันธ์ ปริมาณน้ำฝน ความกดอากาศ ความเร็วลมและทิศทางลม โดยที่ข้อมูลคุณภาพอากาศเก็บบันทึกเป็นรายชั่วโมง อุณหภูมิ ความชื้นสัมพัทธ์และปริมาณน้ำฝนมีลักษณะรายวัน ในขณะที่ความกดอากาศ ความเร็วลมและทิศทางลมเป็นแบบราย 3 ชั่วโมง ตารางที่ 1 สรุปผลข้อมูลคุณภาพอากาศและอุตุนิยมวิทยา พบว่าค่า $PM_{2.5}$ สูงสุดอยู่ในระดับที่เกินกว่ามาตรฐานซึ่งอาจส่งผลกระทบต่อสุขภาพโดยเฉพาะผู้อยู่ในกลุ่มเสี่ยง ทั้งในผู้มีโรคระบบทางเดินหายใจ เด็ก และผู้สูงอายุ

ตารางที่ 1 ผลสรุปข้อมูลรายวันในเขตบางนา กรุงเทพมหานคร ปี 2560-2563

ข้อมูล	ตัวแปร (หน่วยวัด)	ค่าเฉลี่ย	ค่าเบี่ยงเบน	ค่าต่ำสุด/สูงสุด	จำนวนข้อมูล (%)
ฝุ่น $PM_{2.5}$	PM2.5 (µg/m^3)	21.83	16.47	1-173	33,407 (95.27)
ฝุ่น PM_{10}	PM10 (µg/m^3)	38.67	23.55	3-273	34,097 (97.24)
คาร์บอนมอนอกไซด์	CO (ppm)	0.47	0.28	0-3	26,586 (75.82)
ซัลเฟอร์ไดออกไซด์	SO2 (ppb)	1.48	1.41	0-23	26,533 (75.67)
ไนโตรเจนไดออกไซด์	NO2 (ppb)	16.76	14.74	0-120	26,738 (76.25)
โอโซน	O3 (ppb)	20.10	17.81	0-52.52	26,833 (76.53)
อุณหภูมิ	Temp (°C)	29.19	1.65	20.70-33.40	1,461 (100)
ความชื้นสัมพัทธ์	Hum (%)	74.26	7.96	46-96	1,461 (100)
ปริมาณน้ำฝน	Rainfall (mm)	5.89	13.48	0-119	1,101 (75.36)
ความกดอากาศ	Press (hPa)	1009.38	2.93	999.80-1022.10	11,688 (100)
ความเร็วลม	WindSp (knot)	2.94	2.44	0-12	11,688 (100)
ทิศทางลม	WindDir (degree)	118.96	102.87	0-360	11,688 (100)

4.2 การเตรียมข้อมูล

ข้อมูลงานวิจัยจัดเก็บในรูปของตาราง (Data Frame) แต่ละคอลัมน์เป็นข้อมูลของตัวแปรแต่ละตัว แต่ละแถวในแนวนอนแสดงข้อมูลตามลำดับเวลา จากการสำรวจข้อมูลพบว่าข้อมูลไม่สมบูรณ์ขาดหายบางส่วนแสดงดังตารางที่ 1 ข้อมูลที่สูญหายต่อเนื่องยาวนานจะถูกตัดทิ้ง (Listwise Deletion) และค่าสูญหายอื่นถูกแทนที่ด้วยค่าประมาณ ยกตัวอย่างเช่น ในช่วง 9 เดือนหลังของปี 2563 ไม่มีข้อมูล “คาร์บอนมอนอกไซด์” “ซัลเฟอร์ไดออกไซด์” “ไนโตรเจนไดออกไซด์” ส่วน “โอโซน” หายไปมากถึง 92% จึงลบทิ้งแถวข้อมูลในช่วงเวลาดังกล่าว และข้อมูลปริมาณน้ำฝนในช่วงระยะ 4 ปีขาดหายไป 25% ทำให้ตัวแปร Rainfall ถูกยกเลิกออกไป ข้อมูล (รายชั่วโมง) จึงลดลงเหลือ 32,880 แถวและ 11 คอลัมน์ (จากเดิมมีแถวทั้งหมด 35,064 แถวและ 12 คอลัมน์) ในรูปที่ 2 เป็นแผนภาพเมทริกซ์แสดงปริมาณและการกระจายตัวข้อมูลสูญหายของตัวแปร PM2.5, PM10, CO, SO2, NO2 และ O3 ซึ่งมีอัตราการสูญหายเท่ากับ 5.08%, 2.99%, 6.34%, 6.51%, 6.10% และ 7.18% ตามลำดับ สีขาวหมายถึงข้อมูลที่ขาดหายไป ส่วนตัวแปรที่ไม่ปรากฏในแผนภูมิมีข้อมูลครบถ้วนสมบูรณ์ โดยจะประมาณค่าข้อมูลสูญหายทั้งวิธีแทนค่าสูญหายของตัวแปรด้วยค่าของตัวแปรนั้น (Univariate Imputation) ซึ่งก็คือวิธีค่าเฉลี่ยและวิธี LI และวิธีแทนค่าสูญหายผ่านตัวแปรอื่น ๆ (Multivariate Imputation) ในการนี้ใช้วิธี KNN และวิธี MICE

การพยากรณ์ค่า $PM_{2.5}$ รายวันต้องใช้ข้อมูลเป็นรายวัน ข้อมูลที่ไม่ใช่รายวันให้ใช้ค่าเฉลี่ยต่อวันแทนข้อมูลรายวัน ดังนั้นขนาดข้อมูลที่จะนำไปใช้ในขั้นตอนถัดไปคงเหลือ 1,370 แถวและ 11 คอลัมน์ ข้อมูลเหล่านี้จะถูกแบ่งออกเป็น 3 ชุด ได้แก่ ข้อมูลชุดสอน (Training Set) นำไปสอนให้โครงข่ายเกิดการเรียนรู้ ข้อมูลชุดตรวจสอบ (Validation Set) เพื่อประเมินการทำงานของแบบจำลองที่สร้างขึ้นและข้อมูลชุดทดสอบ (Test Set) ใช้ในการทดสอบประสิทธิภาพการทำงานของแบบจำลอง ในลำดับแรกแบ่งข้อมูลออกเป็น 2 ส่วน 80% สำหรับสอน และ 20% สำหรับทดสอบ โดยให้ข้อมูลล่าสุดเป็นชุดทดสอบ จากนั้นตัดเอาส่วนท้าย 20% ของข้อมูลชุดสอนมาเป็นชุดตรวจสอบ ส่วนที่เหลืออยู่จะเป็นข้อมูลชุดสอนที่แท้จริง

4.3 แบบจำลอง LSTM

แบบจำลองโครงข่าย LSTM มีตัวแปรผลลัพธ์เพียง (Output Variable) 1 ตัวแปร คือค่าความเข้มข้นของ $PM_{2.5}$ และตัวแปรป้อนเข้า (Input Variable) หลายตัว ในที่นี้จะใช้แบบจำลอง 4 รูปแบบตามที่เสนอไว้ในงานวิจัยก่อนหน้า [5] ดังนี้

แบบจำลองที่ 1 ประกอบด้วยตัวแปรป้อนเข้าที่มีค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สันสูงสุด 5 อันดับแรก

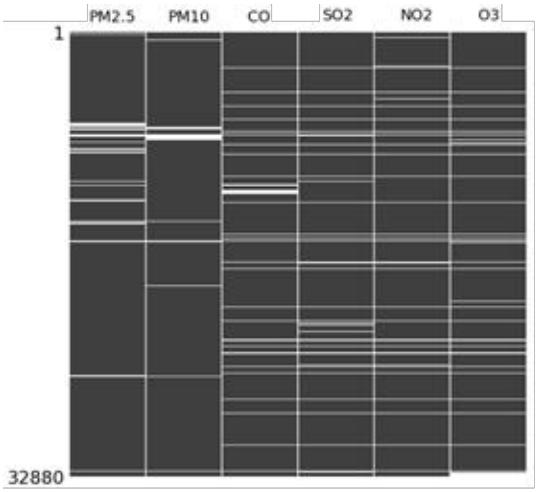
แบบจำลองที่ 2 นำตัวแปรค่า $PM_{2.5}$ ของวันก่อนหน้า (LAGPM2.5) เข้ารวมในแบบจำลองที่ 1

แบบจำลองที่ 3 เกิดจากการใช้ตัวแปรป้อนเข้าทั้งหมดมาพยากรณ์ตัวแปรผลลัพธ์

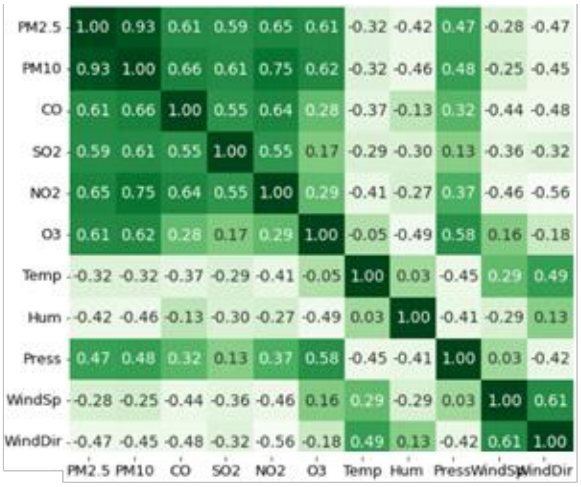
แบบจำลองที่ 4 ใส่ตัวแปร LAGPM2.5 เข้ามาในแบบจำลองที่ 3

ซึ่งข้อมูลที่ได้จากเทคนิควิธีแทนค่าสูญหายทั้ง 4 วิธี ให้ผลการคัดเลือกตัวแปรป้อนเข้าที่มีความสัมพันธ์ในระดับสูงสุด 5 อันดับเหมือนกัน ตัวแปรเหล่านี้คือ PM10, CO, SO2, NO2 และ O3 จะถูกนำเข้ามาในแบบจำลอง ดังนั้นแต่ละวิธีการประมาณค่าข้อมูลสูญหายจะใช้แบบจำลองที่สร้างขึ้นจากตัวแปรชุดเดียวกัน ตัวอย่างค่าสัมประสิทธิ์สหสัมพันธ์ของชุดข้อมูลที่

ถูกแทนที่ค่าหายไปด้วยวิธี MICE แสดงโดยแผนภูมิความร้อน (Heat Map) ดังรูปที่ 3 ยิ่งมีความเข้มของสีมากยิ่งมีขนาดความสัมพันธ์มากตามไปด้วย ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์แสดงผลเป็นตัวเลขในช่องของตาราง



รูปที่ 2 แผนภาพเมทริกซ์การกระจายตัวของข้อมูลสูญหาย



รูปที่ 3 แผนภูมิความร้อนของค่าสัมประสิทธิ์สหสัมพันธ์

แบบจำลองการพยากรณ์ [5] สร้างขึ้นโดยใช้ Keras และได้นำเอาวิธีการ MinMaxScaler มาปรับขอบเขตของข้อมูลให้อยู่ในช่วงเดียวกัน (Normalization) คือระหว่าง 0 กับ 1 มีการใช้งานฟังก์ชันการกระตุ้นรีลู (ReLU Function) และอัลกอริทึมอดัม (Adaptive Moment Estimation: Adam) โดยกำหนดค่าไฮเปอร์พารามิเตอร์ (Hyperparameter) ดังนี้ จำนวนชั้นซ่อน 2 ชั้น จำนวนโหนดในแต่ละชั้นซ่อนเท่ากับ 32 โหนด ขนาดข้อมูลที่ใช้เพื่อปรับค่าน้ำหนักในแต่ละครั้ง (Batch Size)เท่ากับ 32 จำนวนรอบการสอน (Epoch) 100 รอบ และการสุ่มลดจำนวนโหนด (Dropout Rate) 20%

5. ผลการวิจัย

แบบจำลอง LSTM ถูกนำไปใช้ในการทำนายความหนาแน่นของ $PM_{2.5}$ ของวันรุ่งขึ้น โดยอาศัยข้อมูลย้อนหลัง 1 วัน, 3 วัน และ 5 วัน ซึ่งความแม่นยำของการพยากรณ์สำหรับแต่ละวิธีประมาณค่าสูญหายวัดด้วยตัวชี้วัด RMSE และ MAE ดังตารางที่ 2 ค่าที่น้อยสุดในแถวตามแนวนอนแสดงผลเป็นตัวหนา คำนวณได้ด้วยสมการด้านล่าง (3) และ (4) โดยที่ y_i คือค่าข้อมูลจริง $\hat{y}_i$ คือค่าพยากรณ์ และ n คือจำนวนข้อมูล

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (4)$$

การพิจารณาค่าพารามิเตอร์ k ของวิธี KNN หาได้จากการลองผิดลองถูกปรับเปลี่ยนค่าไปเรื่อย ๆ แล้วเลือกค่า k ที่ให้ความผิดพลาดของการพยากรณ์น้อยที่สุด ในกรณีนี้ทดสอบจำนวนเลขคี่ทั้งหมดระหว่าง 1 ถึง 20 ได้ k มีค่าเท่ากับ 11 สำหรับวิธี MICE ใช้งาน IterativeImputer กับ RandomForestRegressor ของ Scikit-learn ให้เงื่อนไขการหยุดการทำงาน Tolerance เป็น 1e-3 จากผลประเมินประสิทธิภาพการพยากรณ์พบว่าค่าความคลาดเคลื่อนต่ำสุดส่วนมากได้จากวิธี KNN และ MICE โดยวิธี KNN ให้ค่าความคลาดเคลื่อนน้อยที่สุด 9 กรณี น้อยที่สุดอันดับสอง 9 กรณีจากทั้งหมด 24 กรณี ขณะที่วิธี MICE ให้ค่าความคลาดเคลื่อนน้อยที่สุด 9 กรณี น้อยที่สุดอันดับสอง 7 กรณี และวิธี KNN ให้ค่าเฉลี่ยความคลาดเคลื่อน RMSE น้อยที่สุด (RMSE = 8.2149) ส่วนวิธี MICE ให้ค่าเฉลี่ยความคลาดเคลื่อน MAE น้อยสุด (MAE = 5.8680) ซึ่งแสดงให้เห็นว่าทั้งสองวิธีไม่แตกต่างกันมากนัก มีความถูกต้องใกล้เคียงกัน ดังนั้นจึงสรุปได้ว่าวิธี KNN และ MICE ประมาณค่าสูญหายได้ดีที่สุดเมื่อเทียบกับ

ตารางที่ 2 ผลวิเคราะห์เปรียบเทียบประสิทธิภาพการพยากรณ์

แบบจำลองที่	ข้อมูลย้อนหลัง (วัน)	วิธีค่าเฉลี่ย		วิธี LI		วิธี KNN (k=11)		วิธี MICE	
		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
1	1	8.1786	5.9897	8.1974	6.0536	8.0989	5.8737	**8.0459**	**5.7203**
	3	8.6244	6.3999	8.5359	6.2024	**8.1633**	**5.9775**	8.4469	6.1222
	5	8.8853	6.6605	9.4231	6.8087	**8.5266**	**6.2718**	8.8007	6.3984
2	1	7.7127	5.3021	**7.6812**	**5.1964**	7.7067	5.2953	7.7721	5.3086
	3	7.7523	5.5212	7.6457	5.3828	7.8449	5.4533	**7.5889**	**5.2834**
	5	8.2646	5.7581	8.7674	6.1022	**8.0618**	**5.6123**	8.3627	5.8665
3	1	8.4806	6.2604	8.9818	6.8289	**8.1947**	**5.8680**	8.2216	6.0583
	3	9.0918	6.8451	8.7415	6.3734	**8.2782**	6.1733	8.3089	**5.9324**
	5	**9.1076**	**6.6341**	9.5423	6.9109	9.5226	7.0639	9.7006	6.9777
4	1	**7.6318**	**5.3213**	7.7938	5.5334	7.6897	5.5107	7.7101	5.4289
	3	7.8389	5.6263	7.8634	5.6647	7.7954	5.5133	**7.5859**	**5.4196**
	5	8.7016	6.0475	8.6816	6.0154	8.6960	6.1742	**8.4108**	**5.9000**
ค่าเฉลี่ย		8.3559	6.0305	8.4879	6.0894	8.2149	5.8989	8.2463	5.8680
ส่วนเบี่ยงเบนมาตรฐาน		0.5396	0.5348	0.6557	0.5741	0.5161	0.4861	0.5976	0.4944

วิธีค่าเฉลี่ยและวิธี LI ทั้งวิธี KNN (k=11) และ MICE เป็น Multivariate Imputation ต่างก็ทำงานได้ดีกว่าวิธีค่าเฉลี่ยและวิธี LI ซึ่งเป็น Univariate Imputation นอกจากนี้ยังพบว่าการใช้ข้อมูลล่าสุดในระยะเวลาที่สั้นจะคาดการณ์แม่นยำมากยิ่งขึ้นและแบบจำลองที่ 2 ทำนายได้ถูกต้องมากกว่าแบบจำลองอื่นสำหรับทั้ง 4 วิธีการแทนค่าสูญหาย

6. สรุปและอภิปรายผล

เนื่องจากการสูญหายของข้อมูลเป็นประเด็นสำคัญที่นักวิจัยไม่ควรละเลย เพราะอาจจะส่งผลต่อผลการวิเคราะห์ที่ได้หรือการตีความจากการนำข้อมูลไปใช้ งานวิจัยหลายประเภทจำเป็นอย่างยิ่งที่ต้องใช้ข้อมูลที่สมบูรณ์เท่านั้น ซึ่งการแก้ไขปัญหาการสูญหายของข้อมูลสามารถใช้วิธีการลบทิ้งหรือแทนที่ด้วยค่าใหม่ การตัดทิ้งอาจทำให้ข้อมูลเกิดความเอนเอียงหากข้อมูลที่ขาดหายมีเป็นจำนวนมาก จึงมักเติมเต็มค่าที่ขาดหายไปโดยการประมาณค่าจากข้อมูลตัวอย่างที่มีอยู่ งานวิจัยนี้เปรียบเทียบวิธีการจัดการข้อมูลสูญหาย 4 วิธี คือวิธีค่าเฉลี่ย วิธี LI วิธี KNN และวิธี MICE ด้วยค่าความคลาดเคลื่อนของการพยากรณ์ฝุ่น $PM_{2.5}$ ในพื้นที่กรุงเทพมหานคร โดยใช้ค่า RMSE และ MAE มาช่วยในการประเมินวิธีการประมาณค่าสูญหาย ความคลาดเคลื่อนยิ่งมีค่าน้อยแสดงว่าวิธีการแทนค่าข้อมูลสูญหายยิ่งให้ผลดี ผลวิจัยสรุปว่าวิธี KNN และ MICE มีประสิทธิภาพสูงสุด ทั้งสองวิธีให้ผลลัพธ์ดีกว่าวิธีค่าเฉลี่ยและวิธี LI เนื่องจากวิธี Multivariate Imputation คำนวณค่าประมาณจากหลายตัวแปรที่มีความสัมพันธ์เกี่ยวข้องกันทำให้ได้ค่าที่แม่นยำมากกว่าวิธี Univariate Imputation และให้ข้อสรุปที่สอดคล้องกับงานวิจัยที่ผู้วิจัยทำมาก่อนหน้านี้ [5] ซึ่งใช้วิธีการดั้งเดิมในการจัดการข้อมูลสูญหายด้วยค่าเฉลี่ย ถึงกระนั้นก็ตามการแทนค่าด้วยค่าเฉลี่ยเป็นวิธีที่ไม่ดีนักแม้ว่าเป็นวิธีที่ใช้กันมาก จึงควรพิจารณาเลือกใช้วิธีการทดแทนค่าข้อมูลสูญหายที่เหมาะสมกับลักษณะข้อมูลสูญหายที่เกิดขึ้น เพื่อให้ได้ข้อมูลที่มีคุณภาพจะช่วยเพิ่มความถูกต้องแม่นยำของผลการวิเคราะห์ นอกจากนี้การพยากรณ์ล่วงหน้าโดยใช้ข้อมูลย้อนหลัง 1 วันมีความถูกต้องมากสุด ด้วยเหตุที่ข้อมูลล่าสุดเป็นข้อมูลที่ใกล้เคียงปัจจุบันมากที่สุด และแบบจำลองที่ 2 ได้พัฒนาขึ้นโดยอาศัยข้อมูลคุณภาพอากาศและ $PM_{2.5}$ ของวันก่อนหน้าให้ผลการทำนายแม่นยำที่สุดเพราะเลือกใช้เฉพาะตัวแปรที่มีความสำคัญมากต่อการพยากรณ์

7. กิตติกรรมประกาศ

งานวิจัยนี้ได้รับทุนอุดหนุนการวิจัยจากกองทุนสนับสนุนการวิจัย นวัตกรรม และการสร้างสรรค์ มหาวิทยาลัยศิลปากร ปี 2562 และได้รับความอนุเคราะห์อย่างดียิ่งด้านข้อมูลเพื่อการวิจัยจากกองจัดการคุณภาพอากาศและเสียง กรมควบคุมมลพิษ กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม และกรมอุตุนิยมวิทยา กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม ผู้วิจัยจึงขอขอบพระคุณอย่างสูงมา ณ โอกาสนี้

8. เอกสารอ้างอิง

[1] World Health Organization, Switzerland (n.d.)., Air Pollution, URL: https://www.who.int/health-topics/air- pollution#tab=tab_1, accessed on 1/05/2022.

[2] กรมควบคุมมลพิษ, กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม, 2563, รายงานสถานการณ์มลพิษของประเทศไทย ปี 2562.

[3] คณะกรรมการสิ่งแวดล้อมแห่งชาติ, 2553, เรื่องกำหนดมาตรฐานฝุ่นละอองขนาดไม่เกิน 2.5 ไมครอนในบรรยากาศโดยทั่วไป, [ระบบออนไลน์], แหล่งที่มา https://www.pcd.go.th/laws/ประกาศคณะกรรมการสิ่งแวดล้อม, เข้าดูเมื่อวันที่ 5 พฤษภาคม 2564.

[4] ชัยยศ ยงค์เจริญชัย, 2562, ฝุ่น: $PM_{2.5}$ ในเชียงใหม่ขึ้นสูงแตะอันดับหนึ่งของโลก, บีบีซีไทย, [ระบบออนไลน์], แหล่งที่มา https://www.bbc.com/thai/47534868, เข้าดูเมื่อวันที่ 20 เมษายน 2564.

[5] กรีนพีซ, 2565, ความเหลื่อมล้ำใต้ท้องฟ้าเดียวกัน: เมื่อต่างจังหวัดฝุ่นเยอะพอๆ กับกรุงเทพฯ แต่เข้าถึงเครื่องวัดคุณภาพอากาศได้ยากกว่า, [ระบบออนไลน์], แหล่งที่มา https://www.greenpeace.org/thailand/story/25299/climate-airpollution-inquity-of-airpollutin, เข้าดูเมื่อวันที่ 30 มีนาคม 2566.

[6] Ahn, H., Sun, K. and Kim, K.P., 2021, Comparison of Missing Data Imputation Methods in Time Series Forecasting, Computers, Materials and Continua, Vol. 70, September 2021, pp. 767–779.

[7] Little, R.J.A. and Rubin, D.B., 2002, Statistical Analysis with Missing Data, 2nd edition, John Wiley & Sons, Hoboken.

[8] Pollice, A. and Lasinio, G.J., 2009, Two Approaches to Imputation and Adjustment of Air Quality Data from a Composite Monitoring Network, Journal of Data Science, Vol. 7, January 2009, pp. 43–59.

[9] Marwala, T., 2009, Computational Intelligence for Missing Data Imputation, Estimation and Management: Knowledge Optimization Techniques, Information Science Reference, Hershey.

[10] Junger, W.L. and Leon, P.D., 2015, Imputation of Missing Data in Time Series for Air Pollutants, Atmospheric Environment, Vol. 102, February 2015, pp. 96–104.

[11] Saeipourdizaj, K.P., Sarbakhsh, P. and Gholampour, A., 2021, Application of Imputation Methods for Missing Values of PM_{10} and O_3 Data: Interpolation, Moving Average and K-nearest Neighbor Methods, Environmental Health Engineering and Management Journal, Vol. 8, September 2021, pp. 215–226.

[12] Rumaling, M.I., Chee, F.P., Dayou, J. and Chang, J., 2020, Missing Value Imputation for PM_{10} Concentration in Sabah Using Nearest Neighbour Method (NNM) and Expectation-Maximization (EM) Algorithm, Asian Journal of Atmospheric Environment, Vol. 14, March 2020, pp. 62–72.

[13] Wijesekara, W.M.L.K.N. and Liyanage, L., 2020, Comparison of Imputation Methods for Missing Values in Air Pollution Data: Case Study on Sydney Air Quality Index, In: Arai, K., Kapoor, S. and Bhatia, R. (eds) Advances in Information and Communication, FICC 2020, Advances in Intelligent Systems and Computing, Vol. 1130, Springer, Cham.

[14] Zainuri, N.A., Jemain, A.A. and Muda, N., 2015, A Comparison of Various Imputation Methods for Missing Values in Air Quality Data, Sains Malaysiana, Vol. 44, March 2015, pp. 449–456.

[15] Donders, A.R., Van Der Heijden, G.J., Stijnen, T., and Moons, K.G., 2006, A Gentle Introduction to Imputation of Missing Values. Journal of Clinical Epidemiology, Vol. 59, October 2006, pp. 1087–1091.

[16] Quinteros, M.E, Lu, S., Blazquez, C., Cárdenas-R, J.P., Ossa, X., Delgado-Saborit, J.M., Harrison, R.M. and Ruiz-Rudolph, P., 2019, Use of Data Imputation Tools to Reconstruct Incomplete Air Quality Datasets: A Case-Study in Temuco, Chile. Atmospheric Environment, Vol. 200, March 2019, pp. 40–49.

[17] Wilson, S., 2021, Miceforest: Fast Imputation with Random Forests in Python, Virginia, USA, URL: https:// morioh.com/p/e19cd87c66e3, Accessed on 10 May 2022.

[18] McCulloch, W.S. and Pitts, W., 1943, A Logical Calculus of the Ideas Immanent in Nervous Activity, The Bulletin of Mathematical Biophysics, Vol. 5, December 1943, pp. 115–133.

[19] Hochreiter, S. and Schmidhuber, J., 1997, Long Short-Term Memory, Neural Computation, Vol. 9, November 1997, pp. 1735–1780.

[20] Olah, C., 2015, Understanding LSTM Networks, Anthropic, California, USA, URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs, Accessed on 10 May 2022.

[21] McCulloch, W.S. and Pitts, W., 1943, A Logical Calculus of the Ideas Immanent in Nervous Activity, The Bulletin of Mathematical Biophysics, Vol. 5, December 1943, pp. 115–133.

[22] Hochreiter, S. and Schmidhuber, J., 1997, Long Short-ธerm Memory, Neural Computation, Vol. 9, November 1997, pp. 1735–1780.

[23] Olah, C., 2015, Understanding LSTM Networks, Anthropic, California, USA, URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs, Accessed on 10 May 2022.