KKU SCIENCE JOURNAL
Journal Home Page : https://ph01.tci-thaijo.org/index.php/KKUSciJ
Published by the Faculty of Science, Khon Kaen University, Thailand

การเปรียบเทียบวิธีการใส่ค่าสูญหายสำหรับอนุกรมเวลาเชิงพหุ กรณีศึกษาดัชนีราคากลุ่มอุตสาหกรรมตลาดหลักทรัพย์แห่งประเทศไทย

Comparing Imputation Methods for Multivariate Time Series: A Case Study for Industrial Group Index of Thailand Stock Market

พงษ์พล ยิ่งประทานพร[1*]

Pongpon Yingpratanpron[1*]

[1]สาขาสถิติและวิทยาการข้อมูล คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพมหานคร 10330

[1]Statistics and Data Science, Faculty of Commerce and Accountancy, Chulalongkorn University, Bangkok, 10330, Thailand

บทคัดย่อ

การศึกษานี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการใส่ค่าสูญหายสำหรับอนุกรมเวลาเชิงพหุ และประเมินผลเพื่อเลือกวิธีการใส่ค่าสูญหายที่เหมาะสมที่สุดสำหรับอนุกรมเวลาเชิงพหุ โดยใช้ข้อมูลทุติยภูมิดัชนีราคากลุ่มอุตสาหกรรมของตลาดหลักทรัพย์แห่งประเทศไทย 8 กลุ่มอุตสาหกรรม จากฐานข้อมูล SETSMART ตั้งแต่วันที่ 1 มกราคม พ.ศ. 2547 ถึง 1 มกราคม พ.ศ. 2567 รวมทั้งสิ้น 4,877 วัน ซึ่งได้มีการกำหนดรูปแบบการสูญหายออกเป็น 3 รูปแบบ ได้แก่ การสูญหายรูปแบบสุ่ม การสูญหายรูปแบบช่วงตามลำดับ และการสูญหายรูปแบบบล็อก และกำหนดสัดส่วนการสูญหายของข้อมูลที่ร้อยละ 5 10 20 30 40 และ 50 ตามลำดับ โดยจำแนกวิธีการใส่ค่าสูญหายออกเป็น 3 วิธี ได้แก่ วิธีการใส่ค่าสูญหายด้วยวิธีการเชิงสถิติ ประกอบไปด้วย ค่าเฉลี่ย ค่ามัธยฐาน ข้อมูลสุดท้ายก่อนการสูญหาย (LOCF) ข้อมูลล่าสุดหลังการสูญหาย (NOCB) และการประมาณค่าช่วงเส้นตรง (Linear Interpolation) วิธีการใส่ค่าสูญหายด้วยวิธีการเรียนรู้ของเครื่อง ประกอบไปด้วย ค่าคาดหวังสูงที่สุด (EM) การใส่ค่าสูญหายด้วยการทดแทนแบบพหุคูณด้วยสมการลูกโซ่ (MICE) เพื่อนบ้านใกล้เคียงที่สุด (KNN) และป่าสุ่ม (Random Forest) และวิธีการใส่ค่าสูญหายด้วยวิธีการเรียนรู้เชิงลึก ประกอบไปด้วย GP-VAE USGAN และ SAITS โดยผู้วิจัยเปรียบเทียบประสิทธิภาพของวิธีการใส่ค่าสูญหายด้วยค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองโดยเฉลี่ย (RMSE) ค่าความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย (MAE) และค่าร้อยละความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย (MAPE) ผลการศึกษาพบว่า ทุกรูปแบบการสูญหาย กรณีเมื่อสัดส่วนการสูญหายน้อยกว่าร้อยละ 50 การใส่ค่าสูญหายด้วยวิธีการเชิงสถิติโดยการประมาณค่าช่วงเส้นตรง (Linear Interpolation) จะมีประสิทธิภาพสูงที่สุด ส่วนกรณีสัดส่วนการสูญหายเท่ากับร้อยละ 50 พบว่าการใส่ค่าสูญหายด้วยวิธีการเรียนรู้ของเครื่องโดยวิธีป่าสุ่ม (Random Forest) มีประสิทธิภาพสูงที่สุด

ABSTRACT

This study investigates and compares imputation methods for multivariate time series data to identify the most effective approach. Using secondary data of stock price indices from eight industrial sectors

*Corresponding Author, E-mail: 6680247026@student.chula.ac.th

Received date: 5 February 2025 | Revised date: 7 May 2025 | Accepted date: 13 May 2025
doi: 10.14456/kkuscij.2025.11

in the Stock Exchange of Thailand, retrieved from the SETSMART database, the study covers 4,877 days from January 1, 2004, to January 1, 2024. Missing data patterns were categorized into three types: random missing, sequential missing, and block missing, with proportions set at 5%, 10%, 20%, 30%, 40%, and 50%. The imputation methods were grouped into three categories: statistical methods (mean, median, LOCF, NOCB, and linear interpolation), machine learning methods (EM, MICE, KNN, and random forest), and deep learning methods (GP-VAE, USGAN, and SAITS). Performance was assessed using root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). Results indicated that linear interpolation was the most efficient method for missing proportions below 50% across all patterns, while random forest was the most efficient method for a 50% missing data proportion.

คำสำคัญ: ข้อมูลสูญหาย รูปแบบของการสูญหาย การใส่ค่าสูญหาย อนุกรมเวลาเชิงพหุ การเปรียบเทียบประสิทธิภาพ
Keywords: Missing Data, Missing Pattern, Imputation, Multivariate Time Series, Comparing Performance

บทนำ

การวิเคราะห์ข้อมูลเป็นกระบวนการที่สำคัญต่อการตอบปัญหาและบรรลุวัตถุประสงค์ในหลาย ๆ ด้าน เช่น ด้านเศรษฐกิจ ด้านการเงิน หรือแม้แต่ด้านการแพทย์ โดยข้อมูลที่นำมาวิเคราะห์ประกอบไปด้วยข้อมูลภาคตัดขวาง (Cross-Sectional Data) ข้อมูลอนุกรมเวลา (Time Series Data) และข้อมูลพาแนลหรือข้อมูลตามยาว (Longitudinal Data) และปัญหาที่เกิดขึ้นในลำดับแรกของการจัดการข้อมูลคือการสูญหายไปของข้อมูลอาจเกิดจากความผิดพลาดในการรวบรวมกรณีเก็บแบบสอบถาม ความไม่ต่อเนื่องในการบันทึกข้อมูลตามช่วงเวลาอาจเกิดจากการบันทึกข้อมูลโดยผู้บันทึกที่ต่างกัน และการหลีกเลี่ยงในการบันทึกข้อมูล ซึ่งการสูญหายของข้อมูลที่เกิดขึ้นอาจทำให้เห็นรูปแบบในการสูญหายของข้อมูลไม่ว่าจะเป็นการสูญหายของข้อมูลแบบสุ่มที่เกิดขึ้นจากข้อผิดพลาดในการเก็บรวบรวมข้อมูล การสูญหายของข้อมูลที่หายไปตามช่วงของเวลาซึ่งเกิดจากการหลีกเลี่ยงในการบันทึกข้อมูล และการสูญหายของข้อมูลที่หายไปเป็นช่วงซึ่งอาจเกิดจากความไม่ต่อเนื่องในการบันทึกข้อมูล

จากที่กล่าวมาเห็นได้ว่ารูปแบบการสูญหายที่เกิดขึ้นมีความแตกต่างกันในแง่ของลักษณะการจัดเก็บข้อมูลซึ่งส่งผลต่อกระบวนในการวิเคราะห์ข้อมูลเป็นอย่างมากโดยเฉพาะข้อมูลอนุกรมเวลาเชิงพหุที่หมายถึง ข้อมูลหลายตัวแปรที่มีลักษณะเป็นลำดับตามเวลา ซึ่งส่วนมากมักเป็นข้อมูลด้านการเงินหรือข้อมูลด้านเศรษฐกิจ เช่น ราคาทองคำรายวัน อัตราเงินเฟ้อรายเดือน ผลิตภัณฑ์มวลรวมภายในประเทศ (Gross Domestic Product: GDP) รายไตรมาส และอัตราการว่างงานรายปี ข้อมูลที่กล่าวมาข้างต้นล้วนมีความสำคัญในการวิเคราะห์ที่แตกต่างกันตามความถี่ของเวลาที่ถูกเก็บ ดั้งนั้นหากข้อมูลเกิดการสูญหายไปในช่วงใดช่วงหนึ่งอาจส่งผลให้การวิเคราะห์ผิดพลาดได้

โดยทั่วไปหากเกิดข้อมูลสูญหายกับชุดข้อมูลจะมีวิธีการในการจัดการได้หลากหลายวิธีโดยที่นิยมใช้คือ การนำแถวของข้อมูลที่สูญหายออก การใส่ค่าสูญหายด้วยค่าทางสถิติ การใส่ค่าสูญหายด้วยแบบจำลอง เป็นต้น แต่ด้วยลักษณะข้อมูลอนุกรมเวลาการนำข้อมูลออกนั้นหมายถึงการไม่วิเคราะห์ตามจุดเวลาซึ่งส่งผลให้การวิเคราะห์มีความคลาดเคลื่อนสูง จึงได้มีการศึกษาวิธีการในการใส่ค่าสูญหายขึ้นโดยเริ่มจากการใช้ค่าทางสถิติ เช่น ค่าเฉลี่ย ค่ามัธยฐาน และค่าตำแหน่งข้อมูลก่อนการสูญหายหรือหลังการสูญหาย เพื่อทดแทนการสูญหายไปของข้อมูลและมีการพัฒนาวิธีการใส่ค่าข้อมูลสูญหายมากขึ้นด้วยการใช้แบบจำลองทางสถิติหรือการใช้ค่าประมาณหรือค่าพยากรณ์ เช่น การแทนค่าสูญหายด้วยสมการถดถอยเชิงเส้นตรง การแทนค่าสูญหายด้วยการประมาณค่าช่วงเส้นตรง และการแทนค่าสูญหายด้วยค่าเฉลี่ยเคลื่อนที่

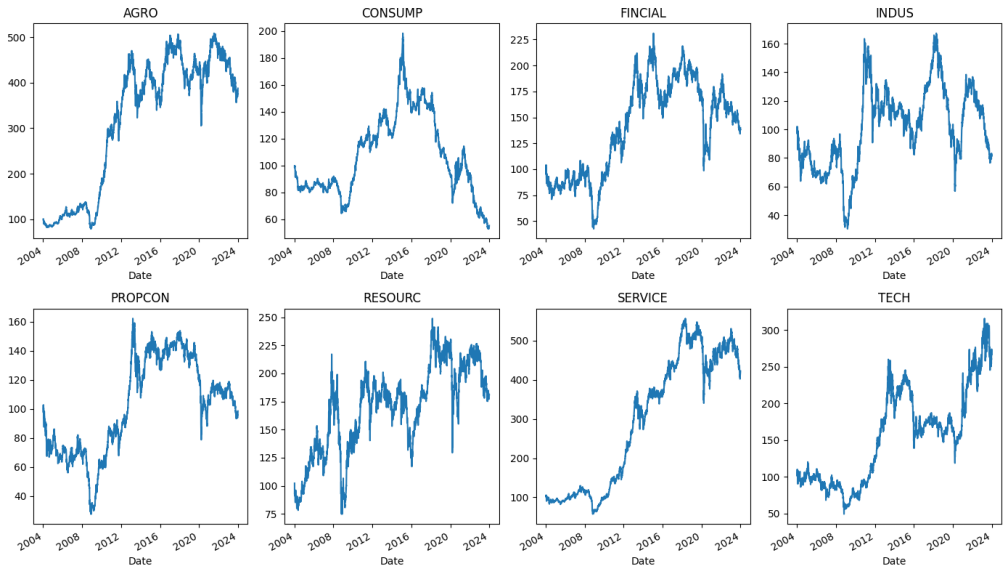
ซึ่งในปัจจุบันได้มีการคิดค้นและศึกษาวิธีการในการใส่ค่าสูญหายให้กับข้อมูลอนุกรมเวลาในหลายงานศึกษา เพื่อพัฒนาวิธีการใส่ค่าสูญหายใหม่ที่ให้ประสิทธิภาพมากกว่าเดิมและประหยัดทรัพยากรในการคำนวณมากยิ่งขึ้นในกรณีที่มีการ

สูญหายของข้อมูลที่มีสัดส่วนมากของข้อมูลอนุกรมเวลาขนาดใหญ่ โดยนำวิธีการเรียนรู้ของเครื่อง (Machine Learning) และวิธีการเรียนรู้เชิงลึก (Deep Learning) มาประยุกต์กับวิธีการใส่ค่าโดยในหลายงานวิจัยได้มีการนำเสนอวิธีการใหม่และทำการเปรียบเทียบกับวิธีการใส่ค่าสูญหายแบบดั้งเดิม จึงเกิดการศึกษาที่เปรียบเทียบประสิทธิภาพของการใส่ค่าสูญหายในแต่ละวิธีการทั้งวิธีการแบบดั้งเดิมและวิธีการที่ถูกคิดค้นขึ้นมาใหม่

โดยมีงานวิจัยที่เกี่ยวข้องกับการศึกษาเปรียบเทียบวิธีการใส่ค่าสูญหายสำหรับอนุกรมเวลาเชิงพหุกับข้อมูลทางการเงิน โดยจำแนกวิธีการใส่ค่าสูญหายออกเป็น 2 วิธีการหลัก ประกอบไปด้วย การใส่ค่าสูญหายกลุ่มวิธีการเชิงสถิติ (Statistical Based) และการใส่ค่าสูญหายกลุ่มวิธีการเรียนรู้ของเครื่อง (Machine Learning Based) (Goldani, 2024) และงานวิจัยที่พัฒนาวิธีการใส่ค่าสูญหายหรือคิดค้นใหม่ ซึ่งเป็นการนำวิธีการเรียนรู้ของเครื่อง (Deep Learning) มาประยุกต์กับการใส่ค่าสูญหาย (Du *et al.*, 2023; Fortuin *et al.*, 2020; Miao *et al.*, 2021) รวมไปถึงงานวิจัยที่ศึกษาการใส่ค่าสูญหายโดยทำการกำหนดรูปแบบการสูญหายออกเป็น 2 รูปแบบได้แก่ การสูญหายรูปแบบทับซ้อน (Overlap) และการสูญหายรูปแบบไม่ต่อกัน (Disjoint) ที่สัดส่วนการสูญหายน้อยกว่าเท่ากับร้อยละ 50 (Almeida *et al.*, 2024) งานวิจัยในครั้งนี้จึงมุ่งเน้นการเปรียบเทียบวิธีการใส่ค่าสูญหายสำหรับข้อมูลอนุกรมเวลาเชิงพหุโดยจำแนกตามวิธีการ ประกอบไปด้วย วิธีการเชิงสถิติ (Statistical Based) วิธีการเรียนรู้ด้วยเครื่อง (Machine Learning Based) และวิธีการเรียนรู้เชิงลึก (Deep Learning Based) ทั้งนี้สาเหตุที่เลือกศึกษาโดยจำแนกตามวิธีการเพื่อให้เห็นถึงการเปรียบเทียบเครื่องมือการใส่ค่าสูญหายของแต่ละวิธีการได้อย่างชัดเจน และหาวิธีการใส่ค่าสูญหายที่เหมาะสมที่สุดกับอนุกรมเวลาเชิงพหุที่มีรูปแบบการสูญหายที่แตกต่างกัน

วิธีการดำเนินการวิจัย

1. ใช้ข้อมูลทุติยภูมิดัชนีราคากลุ่มอุตสาหกรรมในตลาดหลักทรัพย์แห่งประเทศไทย (SET) ประกอบไปด้วย กลุ่มเกษตรและอุตสาหกรรมอาหาร กลุ่มสินค้าอุปโภคบริโภค กลุ่มธุรกิจการเงิน กลุ่มสินค้าอุตสาหกรรม กลุ่มอสังหาริมทรัพย์และก่อสร้าง กลุ่มทรัพยากร กลุ่มบริการ และกลุ่มเทคโนโลยี โดยเป็นข้อมูลรายวันระยะเวลา 20 ปี นับตั้งแต่วันที่ 1 มกราคม พ.ศ. 2547 ถึงวันที่ 1 มกราคม พ.ศ. 2567 รวมทั้งสิ้น 4,877 วัน และใช้ดัชนีราคาปิดในการศึกษา ซึ่งภาพรวมของดัชนีราคาปิดทั้ง 8 อุตสาหกรรมมีลักษณะเป็นข้อมูลอนุกรมเวลาที่มีแนวโน้มขาขึ้นและขาลงตามช่วงเวลาของตลาดหลักทรัพย์ ตามธรรมชาติของข้อมูลด้านการเงินที่มีความถี่เป็นรายวันจะมีลักษณะเป็นแนวโน้มมีความคลาดเคลื่อน และอาจมีหรือไม่มีฤดูกาลก็ได้ (Kotu and Deshpande, 2019) ดังรูปที่ 1



รูปที่ 1 ดัชนีราคา (ราคาปิด) ทั้ง 8 กลุ่มอุตสาหกรรม

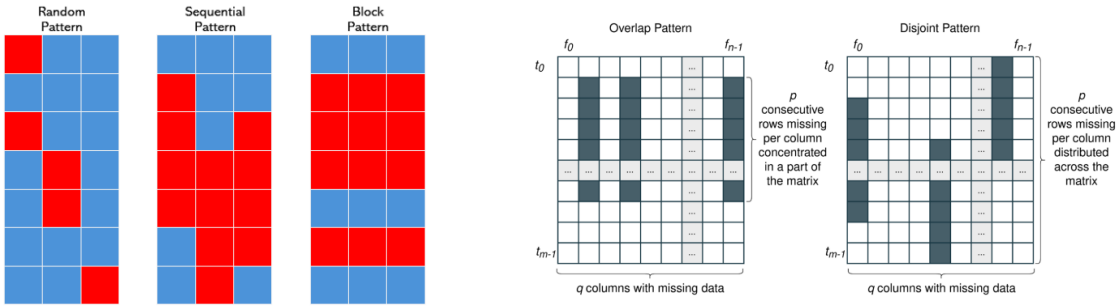
2. กำหนดรูปแบบการสูญหายและกำหนดสัดส่วนการสูญหาย โดยที่ผู้วิจัยได้กำหนดรูปแบบการสูญหายทั้งหมด 3 รูปแบบ ได้แก่ การสูญหายรูปแบบสุ่ม การสูญหายรูปแบบช่วงตามลำดับ และการสูญหายรูปแบบบล็อกสามารถอธิบายได้ดังนี้

2.1 การสูญหายรูปแบบสุ่ม หมายถึง การที่ข้อมูลเกิดการสูญหายไปเป็นจุดแบบสุ่มตามลำดับเวลาซึ่งอาจสุ่มได้เป็นจุดต่อกันในกรณีที่มีสัดส่วนการสูญหายมีค่าสูง

2.2 การสูญหายรูปแบบช่วงตามลำดับ หมายถึง การที่ข้อมูลเกิดการสูญหายไปเป็นช่วงตามลำดับเวลาซึ่งจะเห็นรูปแบบได้ชัดขึ้นเมื่อช่วงที่เกิดการสูญหายกว้างขึ้น

2.3 การสูญหายรูปแบบบล็อก หมายถึง การที่ข้อมูลเกิดการสูญหายไปเป็นช่วงตามลำดับเวลาและเกิดการสูญหายกับตัวแปรทั้งหมดในชุดข้อมูลในลำดับเวลาดังกล่าว

โดยเมื่อเปรียบเทียบกับงานวิจัยของ Almeida *et al.* (2024) ที่ศึกษาวิธีการใส่ค่าสูญหายด้วยวิธีเพื่อนบ้านใกล้เคียงที่สุดแบบมุ่งเน้น (Focalized KNN) โดยกำหนดรูปแบบการสูญหายออกเป็น 2 รูปแบบ คือ การสูญหายรูปแบบทับซ้อน (Overlap) และการสูญหายแบบไม่ต่อกัน (Disjoint) พบว่ารูปแบบการสูญหาย 2 จาก 3 รูปแบบที่ใช้ศึกษามีความสอดคล้องกันโดยที่การสูญหายรูปแบบช่วงตามลำดับในการศึกษาครั้งนี้มีลักษณะที่ใกล้เคียงกับการสูญหายแบบไม่ต่อกัน และการสูญหายรูปแบบบล็อกในการศึกษาครั้งนี้มีลักษณะที่ใกล้เคียงกับการสูญหายรูปแบบทับซ้อนดังรูปที่ 2 ตามคำอธิบายในข้อที่ 2.2 และ 2.3 และกำหนดสัดส่วนการใส่ค่าสูญหายที่ร้อยละ 5 10 20 30 40 และ 50 ตามลำดับ ใช้โปรแกรม Visual Studio Code ภาษาไพธอน (Python) เครื่องมือ PyGrinder ในการกำหนดรูปแบบและสัดส่วนการสูญหายและกำหนดค่าควบคุมการสุ่ม (Random Seed) ที่ 55 โดยกำหนดให้เกิดการสูญหายกับตัวแปร 4 ตัวจากทั้งหมด 8 ตัว และสำหรับการสูญหายรูปแบบช่วงตามลำดับ และการสูญหายรูปแบบบล็อคกำหนดให้ช่วงเกิดการสูญหายประมาณ 10 วัน เพื่อให้สอดคล้องกับการสูญหายในทางปฏิบัติ



รูปที่ 2 เปรียบเทียบรูปแบบการสูญหายจากงานวิจัยของ Almeida *et al.* (2024)

3. ใส่ค่าสูญหายด้วย 3 วิธีการ

3.1 การใส่ค่าสูญหายด้วยวิธีการเชิงสถิติ ประกอบไปด้วย ค่าเฉลี่ย ค่ามัธยฐาน ข้อมูลสุดท้ายก่อนการสูญหาย (Last Observation Carried Forward: LOCF) ข้อมูลล่าสุดหลังการสูญหาย (Next Observation Carried Backward: NOCB) และการประมาณค่าช่วงเส้นตรง (Linear Interpolation) (Almeida *et al.*, 2024)

3.1.1 การใส่ค่าสูญหายด้วยค่าเฉลี่ย เป็นวิธีการใส่ค่าสูญหายที่นำข้อมูลที่สังเกตได้หรือปรากฏในชุดข้อมูลมาหาค่าเฉลี่ยเพื่อแทนที่ข้อมูลที่สูญหายไป

3.1.2 การใส่ค่าสูญหายด้วยค่ามัธยฐาน เป็นวิธีการใส่ค่าสูญหายที่นำค่ากลางของข้อมูลที่สังเกตได้หรือปรากฏในชุดข้อมูลเพื่อแทนที่ข้อมูลที่สูญหายไป

3.1.3 การใส่ค่าสูญหายด้วยข้อมูลสุดท้ายก่อนการสูญหาย (LOCF) เป็นการใส่ค่าสูญหายด้วยข้อมูล ณ จุดเวลาสุดท้ายก่อนการสูญหายเพื่อทดแทนการสูญหายของข้อมูล ณ จุดเวลาถัดไป โดยมีข้อจำกัดเมื่อข้อมูล ณ จุดเวลาเริ่มต้น

เกิดการสูญหายวิธีการนี้ไม่สามารถใส่ค่าสูญหายให้ชุดข้อมูลสมบูรณ์ได้ ซึ่งในการวิจัยครั้งนี้ผู้วิจัยได้ใช้ข้อมูลล่าสุดหลังการสูญหาย (NOCB) เพื่อแทนค่าสูญหายให้ชุดข้อมูลสมบูรณ์

3.1.4 การใส่ค่าสูญหายด้วยข้อมูลล่าสุดหลังการสูญหาย (NOCB) เป็นการใส่ค่าสูญหายด้วยข้อมูล ณ จุดเวลาแรกหลังจากการเกิดการสูญหายเพื่อทดแทนการสูญหายของข้อมูล ณ จุดเวลาก่อนหน้า โดยมีข้อจำกัดเมื่อข้อมูล ณ จุดเวลาสุดท้ายเกิดการสูญหายวิธีการนี้ไม่สามารถใส่ค่าสูญหายให้ชุดข้อมูลสมบูรณ์ได้ ซึ่งผู้วิจัยได้ใช้ข้อมูลสุดหท้ายหลังการสูญหาย (LOCF) เพื่อแทนค่าสูญหายให้ชุดข้อมูลสมบูรณ์

3.1.5 การใส่ค่าสูญหายด้วยการประมาณค่าช่วงเส้นตรง (Linear Interpolation) เป็นการใส่ค่าสูญหายด้วยการใช้ค่าประมาณในช่วงเส้นตรงเพื่อแทนค่าข้อมูลที่สูญหายไป โดยมีข้อจำกัดเมื่อมีข้อมูล ณ จุดเวลาเริ่มต้น และข้อมูล ณ จุดเวลาสุดท้ายสูญหายไปวิธีการนี้ไม่สามารถใส่ค่าสูญหายให้ชุดข้อมูลสมบูรณ์ได้ ซึ่งผู้วิจัยได้ใช้ข้อมูลสุดท้ายก่อนการสูญหาย (LOCF) และข้อมูลสุดท้ายหลังการสูญหาย (NOCB) เพื่อแทนค่าสูญหายให้ชุดข้อมูลสมบูรณ์

โดยที่การใส่ค่าสูญหายด้วยวิธีการทางสถิติทั้ง 5 วิธี ผู้วิจัยได้ใช้ภาษาไพธอน (Python) เครื่องมือ Pandas ผ่านโปรแกรม Visual Studio Code ในการจัดการ

3.2 การใส่ค่าสูญหายด้วยวิธีการเรียนรู้ของเครื่อง ประกอบไปด้วย ค่าคาดหวังสูงที่สุด (Expectation Maximization: EM) การใส่ค่าสูญหายด้วยการทดแทนแบบพหุคูณด้วยสมการลูกโซ่ (Multiple Imputation by Chained Equation: MICE) เพื่อนบ้านใกล้เคียงที่สุด (K-Nearest Neighbor: KNN) และป่าสุ่ม (Random Forest) (Ahn *et al.*, 2022; Goldani, 2024)

3.2.1 การใส่ค่าสูญหายด้วยวิธีการค่าคาดหวังสูงที่สุด (EM) เป็นการใช้อัลกอริทึม EM ซึ่งเป็นอัลกอริทึมเชิงวนซ้ำ (Iterative Algorithm) โดยในแต่ละรอบประกอบด้วยสองขั้นตอน คือ ขั้นตอนคาดหมาย (Expectation Step: E-step) และขั้นตอนการเพิ่มค่าสูงสุด (Maximization Step: M-step) ภายใต้เงื่อนไขที่จะบรรลุการลู่เข้าสู่ค่าที่ดีที่สุด (Global Convergence) (Ng *et al.*, 2012)

3.2.2 การใส่ค่าสูญหายด้วยการทดแทนแบบพหุคูณด้วยสมการลูกโซ่ (MICE) คือการสุ่มเลือกค่าจากข้อมูลที่สังเกตได้แล้วเติมค่าที่ขาดหายไปในแต่ละตัวแปรอย่างเป็นลำดับทีละตัว ซึ่งอัลกอริทึม MICE จัดเป็นวิธีการของ Markov Chain Monte Carlo (MCMC) โดยวิธีการใส่ค่าสูญหายที่ได้รับความนิยมวิธีหนึ่ง คือ แบบจำลองการจับคู่ค่าเฉลี่ยเชิงพยากรณ์แบบกึ่งพาราเมตริก (Predictive Mean Matching: PMM) มีจุดมุ่งหมายเพื่อแทนที่ค่าที่สูญหายด้วยค่าที่สังเกตได้จากข้อมูลที่มีการพยากรณ์ใกล้เคียงกัน (Alruhaymi and Kim, 2021)

3.2.3 การใส่ค่าสูญหายด้วยวิธีเพื่อนบ้านใกล้เคียงที่สุด (KNN) เป็นการใช้ "ความคล้ายคลึงของคุณลักษณะ" (Feature Similarity) ในการคาดการณ์ค่าของจุดข้อมูลใหม่ โดยจุดข้อมูลใหม่จะถูกจัดประเภทด้วยป้ายกำกับ (Label) ที่พบมากที่สุดในบรรดาตัวอย่างฝึก k ตัวอย่างที่อยู่ใกล้ที่สุด โดยทั่วไปมักใช้ระยะทางแบบยุคลิด (Euclidean distance) หรือสัมประสิทธิ์สหสัมพันธ์ในการวัดระยะทาง (Dong *et al.*, 2024)

3.2.4 การใส่ค่าสูญหายด้วยวิธีป่าสุ่ม (Random Forest) เป็นการสุ่มตัวอย่างแบบบูตสแตรป (Bootstrap) จากชุดข้อมูล S โดยที่ S(i) แทนบูตสแตรปที่ i จากนั้นสร้างต้นไม้ตัดสินใจ (Decision Tree) โดยใช้ขั้นตอนการเรียนรู้ที่ถูกปรับแต่งให้แตกต่างจากอัลกอริทึมการเรียนรู้ต้นไม้ตัดสินใจแบบดั้งเดิม กล่าวคือในแต่ละโหนดของต้นไม้มีการสุ่มเลือกคุณลักษณะ (Feature) บางส่วนจากชุด f ที่เป็นสับเซตของคุณลักษณะทั้งหมดใน F แล้วแบ่งโหนดโดยใช้คุณลักษณะที่ดีที่สุดใน f แทนที่จะใช้ทั้งหมดใน F และนำผลลัพธ์ของต้นไม้เหล่านี้มารวมกันผ่านค่าเฉลี่ยหรือการโหวต (Bernstein, 2017) มาแทนที่ข้อมูลที่สูญหาย

โดยที่การใส่ค่าสูญหายด้วยวิธีการเรียนรู้ของเครื่อง วิธีค่าคาดหวังสูงที่สุด (EM) วิธีการทดแทนทดแทนแบบพหุคูณด้วยสมการลูกโซ่ (MICE) และวิธีป่าสุ่ม (Random Forest) ผู้วิจัยได้ใช้โปรแกรม Visual Code Studio ภาษา R ขณะที่การใส่

ค่าสูญหายด้วยวิธีเพื่อนบ้านใกล้เคียงที่สุด (KNN) ผู้วิจัยได้ใช้ภาษาไพธอน (Python) ในการจัดการโดยการกำหนดค่าพารามิเตอร์ในการเขียนโค้ดเป็นค่าเริ่มต้น (Default) ยกเว้นจำนวนเพื่อนบ้าน (n_neighbors) ที่ผู้วิจัยได้กำหนดเท่ากับ 2 นอกจากนี้ได้กำหนดค่าควบคุมการสุ่ม (Random Seed) ที่ 55

3.3 การใส่ค่าสูญหายด้วยวิธีการเรียนรู้เชิงลึก ประกอบไปด้วย Gaussian Process-Variation Autoencoder (GP-VAE) Unsupervised Generative Adversarial Network (USGAN) และ Self-Attention Imputation for Time Series (SAITS)

3.3.1 การใส่ค่าสูญหายด้วยวิธี GP-VAE คือการใส่ค่าสูญหายสำหรับอนุกรมเวลาด้วยความน่าจะเป็นเชิงลึก โดยเป็นการใส่ค่าสูญหายที่พัฒนามาจากวิธี (Variational Autoencoder: VAE) ประกอบกับ (Gaussian Process Prior) ใน (Latent Space) (Fortuin *et al.*, 2020)

3.3.2 การใส่ค่าสูญหายด้วยวิธี USGAN เป็นวิธีที่พัฒนามาจากโครงข่ายปราสาทเทียมวนกลับแบบสองทาง (Bidirectional Recurrent Neural Networks: BiRNNs) ประกอบไปด้วย 3 องค์ประกอบ คือ ตัวกำเนิด (Generator) ตัวกำหนด (Discriminator) และตัวจำแนก (Classifier) (Miao *et al.*, 2021)

3.3.3 การใส่ค่าสูญหายด้วยวิธี SAITS เป็นวิธีการที่พัฒนามาจากวิธี (State of The Art: SOTA) ที่เรียนรู้ค่าสูญหายจากการฝึกฝนที่เหมาะสมร่วม (Joint-Optimization) ของการใส่ค่าสูญหายและการสร้างค่าใหม่ ซึ่งมีลักษณะเป็นการรวมกันแบบถ่วงน้ำหนักของ (Diagonally-Masked Self-Attention: DMSA) (Du *et al.*, 2023)

โดยที่การใส่ค่าสูญหายด้วยวิธีการเรียนรู้เชิงลึกทั้ง 3 วิธีผู้วิจัยได้ใช้ภาษาไพธอน (Python) ผ่านเครื่องมือ PyPOTS (Du, 2023) โดยใช้ค่าพารามิเตอร์และไฮเปอร์พารามิเตอร์เป็นค่าเริ่มต้น (Default) ทั้งหมดและได้กำหนดค่าควบคุมการสุ่ม (Random Seed) ที่ 55

4. การประเมินผลและวัดประสิทธิภาพของการใส่ค่าสูญหายแต่ละวิธีทั้ง 3 เกณฑ์ (Hua *et al.*, 2024) ได้แก่

4.1 ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองโดยเฉลี่ย (RMSE) เป็นการวัดความแตกต่างระหว่างค่าจริงและค่าประมาณยกกำลังสองและหาค่าเฉลี่ยก่อนใส่รากที่สองมีลักษณะดังสมการที่ 1

$$RMSE = \sqrt{\frac{1}{n}\sum_{d=1}^{D}\sum_{t=1}^{T}(Y_{d,t} - \hat{Y}_{d,t})^2} \quad (1)$$

4.2 ค่าความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย (MAE) เป็นการวัดความแตกต่างระหว่างค่าจริงและค่าประมาณสัมบูรณ์โดยเฉลี่ยมีลักษณะดังสมการที่ 2

$$MAE = \frac{1}{n}\sum_{d=1}^{D}\sum_{t=1}^{T}\left|Y_{d,t} - \hat{Y}_{d,t}\right| \quad (2)$$

4.3 ค่าร้อยละความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย (MAPE) เป็นการวัดสัดส่วนระหว่างความแตกต่างของค่าจริงกับค่าประมาณและค่าจริงสัมบูรณ์โดยเฉลี่ยมีหน่วยเป็นร้อยละ และมีลักษณะดังสมการที่ 3

$$MAPE = \frac{1}{n}\sum_{d=1}^{D}\sum_{t=1}^{T}\left|\frac{Y_{d,t} - \hat{Y}_{d,t}}{Y_{d,t}}\right| \times 100 \quad (3)$$

โดยที่ $Y_{d,t}$ คือ ค่าข้อมูลจริงตัวแปรที่ d ณ เวลาที่ t

โดยที่ $\hat{Y}_{d,t}$ คือ ค่าแทนที่ข้อมูลสูญหายตัวแปรที่ d ณ เวลาที่ t

โดยที่ D คือ จำนวนตัวแปรที่มีการสูญหาย

โดยที่ T คือ เวลาของข้อมูลที่มีการสูญหาย

โดยที่ n คือ จำนวนข้อมูลที่สูญหาย

5. เปรียบเทียบประสิทธิภาพ โดยที่วิธีการใส่ค่าสูญหายวิธีใดให้ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองโดยเฉลี่ย ค่าความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย และค่าร้อยละความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยต่ำที่สุดหมายความว่าวิธีการใส่ค่าสูญหายวิธีนั้นมีประสิทธิภาพสูงที่สุด

ผลการวิจัยและวิจารณ์ผล

จากการใส่ค่าสูญหายด้วยวิธีการเชิงสถิติ ประกอบด้วย ค่าเฉลี่ย ค่ามัธยฐาน ข้อมูลสุดท้ายก่อนการสูญหาย (LOCF) ข้อมูลล่าสุดหลังการสูญหาย (NOCB) และการประมาณค่าช่วงเส้นตรง (Linear Interpolation) การใส่ค่าสูญหายด้วยวิธีการเรียนรู้ของเครื่อง ประกอบด้วย ค่าคาดหวังสูงที่สุด (EM) การใส่ค่าสูญหายด้วยการทดแทนแบบพหุคูณด้วยสมการลูกโซ่ (MICE) เพื่อนบ้านใกล้เคียงที่สุด (KNN) และป่าสุ่ม (Random Forest) และการใส่ค่าสูญหายด้วยวิธีการเรียนรู้เชิงลึก ประกอบด้วย GP-VAE USGAN และ SAITS แล้วนำมาคำนวณค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองโดยเฉลี่ย ค่าความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย และค่าร้อยละความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย สำหรับการสูญหายรูปแบบสุ่ม การสูญหายรูปแบบช่วงตามลำดับ และการสูญหายรูปแบบบล็อก ดังตารางที่ 1 2 และ 3 ตามลำดับ

ตารางที่ 1 ประสิทธิภาพของการใส่ค่าสูญหายรูปแบบสุ่ม

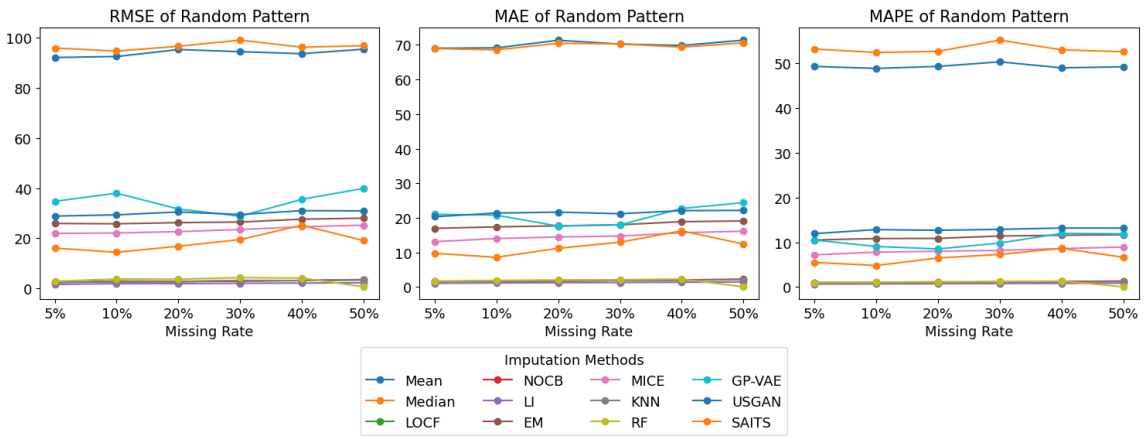
วิธีการใส่ค่าสูญหาย	สัดส่วนการสูญหาย					
	5%	10%	20%	30%	40%	50%
Mean	(92.164/ 69.019/ 49.319%)	(92.529/ 69.130/ 48.883%)	(95.331/ 71.318/ 49.326%)	(94.462/ 70.228/ 50.350%)	(93.671/ 69.823/ 49.004%)	(95.478/ 71.355/ 49.241%)
Median	(95.923/ 68.932/ 53.214%)	(94.700/ 68.541/ 52.432%)	(96.620/ 70.450/ 52.697%)	(99.072/ 70.261/ 55.197%)	(96.248/ 69.325/ 53.035%)	(96.861/ 70.593/ 52.608%)
LOCF	(2.412/ 1.598/ 0.966%)	(2.950/ 1.742/ 1.012%)	(2.820/ 1.832/ 1.068%)	(3.051/ 1.946/ 1.162%)	(3.170/ 2.064/ 1.221%)	(3.395/ 2.217/ 1.306%)
NOCB	(2.394/ 1.639/ 0.963%)	(2.686/ 1.734/ 1.023%)	(2.674/ 1.774/ 1.049%)	(2.922/ 1.906/ 1.127%)	(3.179/ 2.057/ 1.227%)	(3.450/ 2.250/ 1.319%)
Linear Interpolation	(1.611/ 1.099/ 0.655%)	(1.903/ 1.194/ 0.693%)	(1.909/ 1.264/ 0.735%)	(2.041/ 1.326/ 0.784%)	(2.149/ 1.386/ 0.816%)	(2.252/ 1.489/ 0.879%)
EM	(25.916/ 16.968/ 10.517%)	(25.788/ 17.398/ 10.836%)	(26.252/ 17.698/ 10.867%)	(26.497/ 17.999/ 11.401%)	(27.624/ 18.905/ 11.574%)	(28.042/ 19.112/ 11.687%)

ตารางที่ 1 ประสิทธิภาพของการใส่ค่าสูญหายรูปแบบสุ่ม (ต่อ)

วิธีการใส่ค่าสูญหาย	สัดส่วนการสูญหาย					
	5%	10%	20%	30%	40%	50%
MICE	(21.950/ 13.118/ 7.180%)	(22.117/ 14.029/ 7.807%)	(22.668/ 14.473/ 7.996%)	(23.502/ 14.743/ 8.185%)	(24.621/ 15.696/ 8.592%)	(25.210/ 16.157/ 8.939%)
KNN	(2.480/ 1.535/ 0.886%)	(2.638/ 1.564/ 0.902%)	(2.685/ 1.635/ 0.941%)	(3.202/ 1.927/ 1.114%)	(3.122/ 1.943/ 1.102%)	(3.585/ 2.167/ 1.235%)
Random Forest	(2.857/ 1.743/ 1.008%)	(3.777/ 1.922/ 1.091%)	(3.693/ 2.052/ 1.181%)	(4.305/ 2.121/ 1.217%)	(4.129/ 2.321/ 1.334%)	(0.654/ 0.085/ 0.049%)
GP-VAE	(34.767/ 21.030/ 10.498%)	(38.002/ 20.784/ 9.065%)	(31.693/ 17.620/ 8.511%)	(28.803/ 17.976/ 9.835%)	(35.549/ 22.697/ 11.983%)	(39.921/ 24.418/ 11.909%)
USGAN	(28.865/ 20.338/ 11.941%)	(29.364/ 21.381/ 12.833%)	(30.506/ 21.660/ 12.682%)	(29.507/ 21.195/ 12.898%)	(31.032/ 22.092/ 13.191%)	(30.944/ 22.160/ 13.174%)
SAITS	(16.088/ 9.771/ 5.507%)	(14.425/ 8.596/ 4.806%)	(16.843/ 11.208/ 6.494%)	(19.499/ 12.960/ 7.276%)	(25.220/ 16.247/ 8.654%)	(19.084/ 12.439/ 6.647%)

หมายเหตุ: ภายในวงเล็บแสดงถึงค่า RMSE MAE และ MAPE ตามลำดับ

จากตารางที่ 1 พบว่าวิธีการใส่ค่าสูญหายด้วยข้อมูลสุดท้ายก่อนการสูญหาย (LOCF) ข้อมูลล่าสุดหลังการสูญหาย (NOCB) การประมาณค่าช่วงเส้นตรง (Linear Interpolation) เพื่อนบ้านใกล้เคียงที่สุด (KNN) และป่าสุ่ม (Random Forest) มีประสิทธิภาพที่ใกล้เคียงกันในทุกสัดส่วนการสูญหาย โดยมีค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองโดยเฉลี่ย ค่าความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย และค่าร้อยละความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยอยู่ในระดับที่ต่ำ ในขณะที่การใส่ค่าสูญหายด้วยวิธีการเรียนรู้เชิงลึกทั้ง 3 วิธี และการใส่ค่าสูญหายด้วยวิธีค่าคาดหวังสูงที่สุด (EM) กับการใส่ค่าสูญหายด้วยการทดแทนแบบพหุคูณด้วยสมการลูกโซ่ (MICE) มีประสิทธิภาพที่ต่ำกว่า 5 วิธีข้างต้น แต่มีประสิทธิภาพที่มากกว่าการใส่ค่าสูญหายด้วยค่าเฉลี่ยและค่ามัธยฐานที่ประสิทธิภาพต่ำที่สุด ซึ่งเมื่อเปรียบเทียบกลุ่มที่มีประสิทธิภาพที่ดีที่สุด 5 วิธีแรกพบว่าที่สัดส่วนการสูญหายร้อยละ 5 10 20 30 40 วิธีการใส่ค่าสูญหายด้วยการประมาณค่าช่วงเส้นตรง (Linear Interpolation) มีประสิทธิภาพสูงที่สุดโดยมีค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองโดยเฉลี่ยอยู่ที่ 1.611 1.903 1.909 2.041 และ 2.149 ตามลำดับ มีค่าความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยอยู่ที่ 1.099 1.194 1.264 1.326 และ 1.386 ตามลำดับ และมีค่าร้อยละความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยอยู่ที่ 0.655 0.693 0.735 0.784 และ 0.816 ตามลำดับ ขณะที่สัดส่วนการสูญหายที่ร้อยละ 50 การใส่ค่าสูญหายด้วยวิธีป่าสุ่ม (Random Forest) มีประสิทธิภาพสูงที่สุด โดยมีค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองโดยเฉลี่ย ค่าความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย และค่าร้อยละความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยอยู่ที่ 0.654 0.085 0.049 ตามลำดับ สามารถแสดงการเปรียบเทียบได้ดังรูปที่ 3



รูปที่ 3 แสดงการเปรียบเทียบประสิทธิภาพวิธีการใส่ค่าสูญหายรูปแบบสุ่ม

ตารางที่ 2 ประสิทธิภาพของการใส่ค่าสูญหายรูปแบบช่วงตามลำดับ

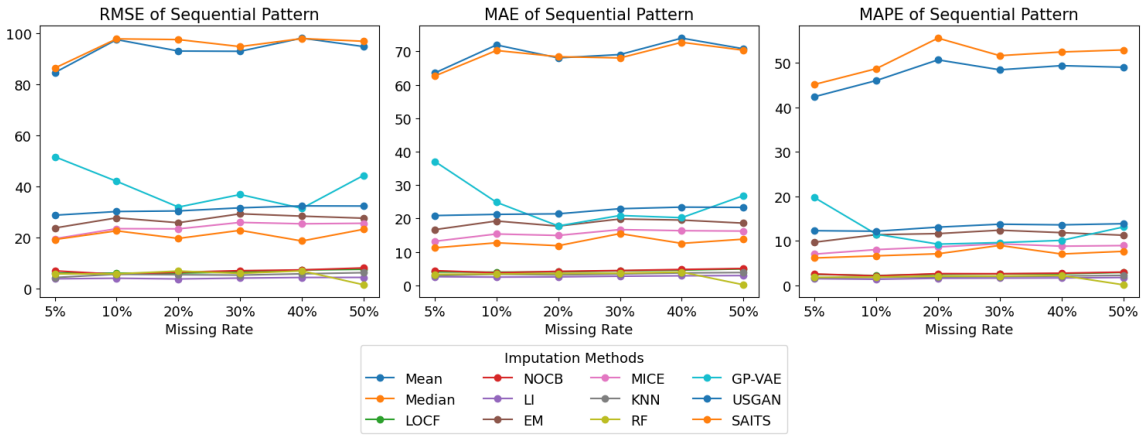
วิธีการใส่ค่าสูญหาย	สัดส่วนการสูญหาย					
	5%	10%	20%	30%	40%	50%
Mean	(84.726/ 63.54/ 42.39%)	(97.637/ 71.935/ 46.025%)	(93.157/ 68.051/ 50.681%)	(93.055/ 69.06/ 48.449%)	(98.211/ 73.966/ 49.377%)	(94.91/ 70.777/ 49.021%)
Median	(86.486/ 62.629/ 45.152%)	(97.926/ 70.272/ 48.694%)	(97.654/ 68.413/ 55.565%)	(94.897/ 68.044/ 51.631%)	(98.053/ 72.722/ 52.458%)	(96.987/ 70.355/ 52.913%)
LOCF	(6.046/ 4.047/ 2.437%)	(6.071/ 3.971/ 2.174%)	(6.115/ 4.043/ 2.459%)	(6.494/ 4.362/ 2.557%)	(7.307/ 4.590/ 2.575%)	(7.572/ 4.943/ 2.878%)
NOCB	(6.970/ 4.429/ 2.527%)	(5.628/ 3.794/ 2.110%)	(6.442/ 4.215/ 2.603%)	(7.048/ 4.464/ 2.582%)	(7.343/ 4.775/ 2.707%)	(8.127/ 5.073/ 2.954%)
Linear Interpolation	(3.855/ 2.595/ 1.472%)	(4.070/ 2.491/ 1.338%)	(3.798/ 2.553/ 1.561%)	(4.125/ 2.756/ 1.600%)	(4.381/ 2.885/ 1.644%)	(4.408/ 2.955/ 1.737%)
EM	(23.779/ 16.68/ 9.670%)	(27.752/ 19.268/ 11.365%)	(25.845/ 17.746/ 11.608%)	(29.339/ 19.878/ 12.373%)	(28.419/ 19.551/ 11.827%)	(27.617/ 18.631/ 11.244%)
MICE	(19.578/ 13.197/ 7.003%)	(23.508/ 15.387/ 8.011%)	(23.431/ 14.954/ 8.611%)	(25.96/ 16.698/ 9.316%)	(25.422/ 16.379/ 8.779%)	(25.618/ 16.257/ 8.911%)

ตารางที่ 2 ประสิทธิภาพของการใส่ค่าสูญหายรูปแบบช่วงตามลำดับ (ต่อ)

วิธีการใส่ค่าสูญหาย	สัดส่วนการสูญหาย					
	5%	10%	20%	30%	40%	50%
KNN	(4.370/ 2.917/ 1.674%)	(5.627/ 3.342/ 1.771%)	(5.44/ 3.148/ 1.868%)	(5.312/ 3.426/ 1.935%)	(5.806/ 3.739/ 2.069%)	(6.254/ 3.858/ 2.213%)
Random Forest	(5.765/ 3.323/ 1.785%)	(5.796/ 3.432/ 1.834%)	(6.887/ 3.552/ 2.049%)	(5.993/ 3.654/ 2.060%)	(6.956/ 3.995/ 2.202%)	(1.507/ 0.193/ 0.098%)
GP-VAE	(51.695/ 37.119/ 19.707%)	(42.083/ 24.880/ 11.480%)	(31.975/ 17.797/ 9.244%)	(36.890/ 20.916/ 9.578%)	(31.484/ 20.264/ 10.081%)	(44.422/ 26.862/ 13.086%)
USGAN	(28.792/ 20.901/ 12.267%)	(30.229/ 21.253/ 12.139%)	(30.468/ 21.421/ 13.063%)	(31.677/ 22.947/ 13.723%)	(32.434/ 23.422/ 13.576%)	(32.383/ 23.344/ 13.834%)
SAITS	(19.256/ 11.250/ 6.129%)	(22.655/ 12.758/ 6.598%)	(19.686/ 11.882/ 7.071%)	(22.856/ 15.497/ 8.949%)	(18.684/ 12.579/ 7.024%)	(23.276/ 13.843/ 7.622%)

หมายเหตุ: ภายในวงเล็บแสดงถึงค่า RMSE MAE และ MAPE ตามลำดับ

จากตารางที่ 2 สามารถสรุปได้ในลักษณะเดียวกับตารางที่ 1 ซึ่งเมื่อเปรียบเทียบกลุ่มที่มีประสิทธิภาพที่ดีที่สุด 5 วิธีแรกพบว่าที่สัดส่วนการสูญหายร้อยละ 5 10 20 30 40 วิธีการใส่ค่าสูญหายด้วยการประมาณค่าช่วงเส้นตรง (Linear Interpolation) มีประสิทธิภาพสูงที่สุดโดยมีค่ารากที่สองของความคลาดเคลื่อนกำลังสองโดยเฉลี่ยอยู่ที่ 3.855 4.070 3.798 4.125 และ 4.381 ตามลำดับ มีค่าความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยอยู่ที่ 2.595 2.491 2.553 2.756 และ 2.885 ตามลำดับ และมีค่าร้อยละความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยอยู่ที่ 1.472 1.338 1.561 1.600 และ 1.644 ตามลำดับ ขณะที่สัดส่วนการสูญหายที่ร้อยละ 50 การใส่ค่าสูญหายด้วยวิธีป่าสุ่ม (Random Forest) มีประสิทธิภาพสูงที่สุด โดยมีค่ารากที่สองของความคลาดเคลื่อนกำลังสองโดยเฉลี่ย ค่าความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย และค่าร้อยละความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยอยู่ที่ 1.507 0.193 0.098 ตามลำดับ สามารถแสดงการเปรียบเทียบได้ดังรูปที่ 4



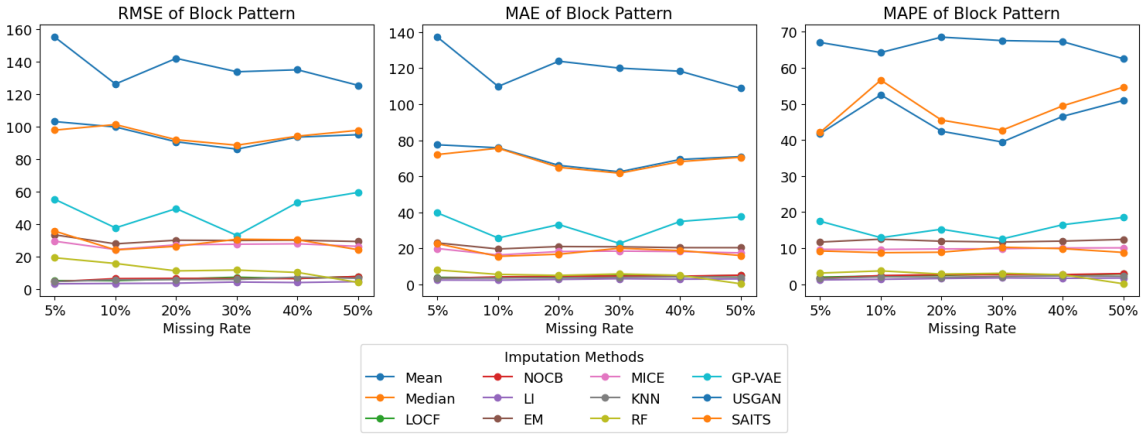
รูปที่ 4 แสดงการเปรียบเทียบประสิทธิภาพวิธีการใส่ค่าสูญหายรูปแบบช่วงตามลำดับ

ตารางที่ 3 ประสิทธิภาพของการใส่ค่าสูญหายรูปแบบบล็อก

วิธีการใส่ค่าสูญหาย	สัดส่วนการสูญหาย					
	5%	10%	20%	30%	40%	50%
Mean	(103.194/ 77.480/ 41.784%)	(99.862/ 75.805/ 52.526%)	(90.759/ 66.038/ 42.441%)	(86.241/ 62.474/ 39.456%)	(93.656/ 69.332/ 46.584%)	(95.126/ 70.919/ 51.002%)
Median	(97.962/ 72.086/ 42.135%)	(101.358/ 75.559/ 56.615%)	(92.035/ 64.971/ 45.527%)	(88.655/ 61.704/ 42.725%)	(94.217/ 68.137/ 49.497%)	(97.816/ 70.550/ 54.690%)
LOCF	(5.329/ 3.867/ 1.957%)	(5.314/ 3.642/ 2.307%)	(6.370/ 4.211/ 2.455%)	(7.425/ 5.076/ 2.845%)	(6.650/ 4.417/ 2.526%)	(7.914/ 4.931/ 2.846%)
NOCB	(4.710/ 3.320/ 1.637%)	(6.567/ 4.121/ 2.385%)	(6.730/ 4.567/ 2.671%)	(6.533/ 4.576/ 2.640%)	(6.603/ 4.498/ 2.673%)	(7.905/ 5.091/ 2.953%)
Linear Interpolation	(3.392/ 2.364/ 1.194%)	(3.601/ 2.322/ 1.359%)	(3.772/ 2.719/ 1.588%)	(4.479/ 3.105/ 1.755%)	(4.154/ 2.862/ 1.647%)	(4.732/ 3.003/ 1.725%)
EM	(33.499/ 22.998/ 11.690%)	(27.970/ 19.634/ 12.489%)	(30.191/ 20.984/ 11.936%)	(30.047/ 20.848/ 11.699%)	(30.212/ 20.358/ 11.961%)	(29.406/ 20.350/ 12.432%)
MICE	(29.585/ 19.881/ 9.680%)	(24.468/ 16.283/ 9.657%)	(27.367/ 18.211/ 9.787%)	(27.731/ 18.458/ 9.847%)	(28.008/ 18.175/ 10.052%)	(26.399/ 17.481/ 10.093%)
KNN	(4.757/ 3.230/ 1.559%)	(5.852/ 3.563/ 2.054%)	(5.970/ 3.567/ 1.960%)	(6.061/ 3.890/ 2.137%)	(7.443/ 4.143/ 2.300%)	(6.798/ 3.902/ 2.199%)
Random Forest	(19.401/ 7.938/ 3.095%)	(15.891/ 5.521/ 3.696%)	(11.335/ 4.964/ 2.822%)	(11.818/ 5.766/ 3.024%)	(10.367/ 5.016/ 2.631%)	(4.252/ 0.302/ 0.120%)
GP-VAE	(55.355/ 39.651/ 17.468%)	(37.794/ 25.727/ 12.917%)	(49.647/ 33.181/ 15.257%)	(33.125/ 22.719/ 12.567%)	(53.442/ 34.903/ 16.505%)	(59.633/ 37.517/ 18.546%)
USGAN	(155.143/ 137.239/ 67.067%)	(126.316/ 109.873/ 64.275%)	(142.091/ 123.875/ 68.540%)	(133.848/ 120.046/ 67.591%)	(135.064/ 118.318/ 67.275%)	(125.470/ 108.807/ 62.521%)
SAITS	(35.792/ 22.577/ 9.226%)	(24.242/ 15.539/ 8.747%)	(26.436/ 16.721/ 8.882%)	(30.797/ 20.042/ 10.265%)	(30.451/ 18.747/ 9.868%)	(24.366/ 15.992/ 8.817%)

หมายเหตุ: ภายในวงเล็บแสดงถึงค่า RMSE MAE และ MAPE ตามลำดับ

จากตารางที่ 3 สังเกตได้ว่าวิธีการใส่ค่าสูญหายด้วยวิธี USGAN นั้นมีประสิทธิภาพต่ำที่สุดโดยมีค่ารากที่สองของความคลาดเคลื่อนกำลังสองโดยเฉลี่ยโดย ค่าความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยมากกว่า 100 และค่าร้อยละความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยมากกว่าร้อยละ 60 ในขณะที่วิธีการอื่นยังคงสรุปได้ในลักษณะเดียวกับตารางที่ 1 และ 2 ซึ่งเมื่อเปรียบเทียบกลุ่มที่มีประสิทธิภาพที่ดีที่สุด 5 วิธีแรกพบว่าที่สัดส่วนการสูญหายร้อยละ 5 10 20 30 40 วิธีการใส่ค่าสูญหายด้วยการประมาณค่าช่วงเส้นตรง (Linear Interpolation) มีประสิทธิภาพสูงที่สุดโดยมีค่ารากที่สองของความคลาดเคลื่อนกำลังสองโดยเฉลี่ยอยู่ที่ 3.392 3.601 3.772 4.479 และ 4.154 ตามลำดับ มีค่าความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยอยู่ที่ 2.364 2.322 2.719 3.105 และ 2.862 ตามลำดับ และมีค่าร้อยละความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยอยู่ที่ 1.194 1.359 1.588 1.755 และ 1.647 ตามลำดับ ขณะที่สัดส่วนการสูญหายที่ร้อยละ 50 การใส่ค่าสูญหายด้วยวิธีป่าสุ่ม (Random Forest) มีประสิทธิภาพสูงที่สุด โดยมีค่ารากที่สองของความคลาดเคลื่อนกำลังสองโดยเฉลี่ย ค่าความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย และค่าร้อยละความคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยอยู่ที่ 4.252 0.302 0.120 ตามลำดับ สามารถแสดงการเปรียบเทียบได้ดังรูปที่ 5



รูปที่ 5 แสดงการเปรียบเทียบประสิทธิภาพวิธีการใส่ค่าสูญหายรูปแบบบล็อก

ผลการวิจัยบ่งชี้ว่าการใส่ค่าสูญหายด้วยวิธีการประมาณค่าช่วงเส้นตรง (Linear Interpolation) มีประสิทธิภาพสำหรับการใส่ค่าสูญหายทั้ง 3 รูปแบบการสูญหายที่สัดส่วนการสูญหายร้อยละ 5 10 20 30 และ 40 ในขณะที่สัดส่วนการสูญหายร้อยละ 50 การใส่ค่าสูญหายด้วยวิธีป่าสุ่ม (Random Forest) มีประสิทธิภาพสูงที่สุด ทั้งนี้การใส่ค่าสูญหายด้วยวิธีการเรียนรู้เชิงลึกทั้ง 3 วิธีอาจไม่เหมาะสมกับชุดข้อมูลที่ได้ทำการวิจัยในครั้งนี้ อาจเกิดจากความซับซ้อนที่ต่ำหรือขนาดของข้อมูลรวมไปถึงความถี่ของช่วงเวลาในข้อมูลที่ต่ำกล่าวคือขนาดของข้อมูลที่เหมาะสมกับการใช้แบบจำลองทางด้านการเรียนรู้เชิงลึกจำเป็นต้องมีขนาดที่ใหญ่ที่ทำให้เกิดการเรียนรู้อย่างมีประสิทธิภาพ และการที่ช่วงการสูญหายที่น้อยอาจทำให้แบบจำลองไม่สามารถจับรูปแบบเพื่อการใส่ค่าสูญหายได้อย่างมีประสิทธิภาพ จึงทำให้ประสิทธิภาพในการใส่ค่าสูญหายค่อนข้างต่ำ และการวิจัยในครั้งนี้แสดงให้เห็นว่าการใส่ค่าสูญหายในบางวิธีการอาจไม่เหมาะสมกับรูปแบบการสูญหายด้วยอย่างในกรณีของการใส่ค่าสูญหายด้วยวิธี USGAN กับรูปแบบการสูญหายแบบบล็อกที่มีประสิทธิภาพต่ำที่สุด

อภิปรายผลการวิจัย

จากผลการวิจัยพบที่ว่าการใส่ค่าสูญหายด้วยวิธีการประมาณค่าช่วงเส้นตรง (Linear Interpolation) มีประสิทธิภาพสำหรับการใส่ค่าสูญหายทั้ง 3 รูปแบบการสูญหายที่สัดส่วนการสูญหายร้อยละ 5 10 20 30 และ 40 เมื่อเปรียบเทียบวิจัยของ Almeida *et al.* (2024) ที่ศึกษาการใส่ค่าสูญหายสำหรับข้อมูลอนุกรมเวลาด้วยวิธีเพื่อนบ้านใกล้เคียงที่สุดแบบมุ่งเน้น (Focalized KNN) โดยใช้ ค่าเฉลี่ย ค่ามัธยฐาน และการประมาณค่าช่วงเส้นตรง (Linear Interpolation) เป็นวิธีการใส่ค่าสูญ

หายพื้นฐาน (Baseline) ที่ดีที่สุดเพื่อเปรียบเทียบกับวิธีเพื่อนบ้านใกล้เคียงที่สุด (KNN) และวิธีเพื่อนบ้านใกล้เคียงที่สุดแบบมุ่งเน้น (Focalized KNN) กับชุดข้อมูลอนุกรมเวลา ENTSO-e และ OpenWeather โดยที่ผลการศึกษาพบว่าการใส่ค่าสูญหายด้วยวิธีเพื่อนบ้านใกล้เคียงที่สุดแบบมุ่งเน้น (Focalized KNN) มีประสิทธิภาพสูงที่สุด และเหมาะสมกับรูปแบบการสูญหายแบบไม่ต่อกัน (Disjoint) มากกว่ารูปแบบทับซ้อน (Overlap) พบว่าผลการวิจัยมีความขัดแย้งกัน อาจเนื่องมาจากลักษณะของข้อมูลอนุกรมเวลาที่มีลักษณะแตกต่างกันหรือการกำหนดให้เกิดการสูญหายจากเครื่องมือที่แตกต่างกัน

ในที่สัดส่วนการสูญหายร้อยละ 50 การใส่ค่าสูญหายด้วยวิธีป่าสุ่ม (Random Forest) มีประสิทธิภาพสูงที่สุด เมื่อเปรียบเทียบกับงานวิจัยของ Goldani (2024) ที่ศึกษาการวิเคราะห์เชิงเปรียบเทียบของวิธีการใส่ค่าสูญหายกรณีศึกษาสำหรับข้อมูลทางการเงินโดยใช้ดัชนี S&P500 กับชุดข้อมูลมูลค่าบิตคอยน์ เป็นชุดข้อมูลในการศึกษา ซึ่งทางผู้ทำวิจัยได้มีการจำแนกวิธีการใส่ค่าสูญหายออกเป็น 2 ประเภท ได้แก่ การใส่ค่าสูญหายด้วยวิธีทางสถิติ ประกอบไปด้วย การใส่ค่าสูญหายด้วยค่าเฉลี่ยและค่าฐานนิยม การใส่ค่าสูญหายด้วยสมการถดถอย การใส่ค่าสูญหายด้วยวิธี Hot Deck และการใส่ค่าสูญหายด้วยวิธี EM ประเภทที่ 2 การใส่ค่าสูญหายด้วยวิธีการเรียนรู้ของเครื่อง ประกอบไปด้วย การใส่ค่าสูญหายด้วยวิธีเพื่อนบ้านใกล้เคียงที่สุด (KNN) การใส่ค่าสูญหายด้วยวิธีการต้นไม้ตัดสินใจ (Decision Tree) การใส่ค่าสูญหายด้วยวิธีป่าสุ่ม (Random Forest) และการใส่ค่าสูญหายด้วยวิธีการเรียนรู้เชิงลึก (Deep Learning) โดยในการศึกษามีการสุ่มนำข้อมูลออก 45 จุดข้อมูลจาก 1,900 จุดข้อมูลคิดเป็นสัดส่วนการสูญหายร้อยละ 2.37 ก่อนทำการใส่ค่าสูญหายด้วยวิธีการข้างต้น และใช้ค่าร้อยละความเคลื่อนสัมบูรณ์โดยเฉลี่ย (MAPE) เป็นเกณฑ์ในการวัดประสิทธิภาพของวิธีการ ผลการศึกษาพบว่าการใส่ค่าสูญหายด้วยวิธีป่าสุ่มมีประสิทธิภาพมากที่สุดในขณะที่การใส่ค่าสูญหายด้วยค่าเฉลี่ยให้ประสิทธิภาพที่ต่ำที่สุด พบว่าวิธีการใส่ค่าสูญหายมีความสอดคล้องกันจริงแต่สัดส่วนการสูญหายที่เกิดขึ้นมีความแตกต่างกันมากและหากเปรียบเทียบจริงจะเปรียบเทียบได้เพียงแค่รูปแบบการสูญหายแบบสุ่มเท่านั้น ดังนั้นจึงไม่สามารถสรุปได้ว่าผลการวิจัยมีความสอดคล้องกัน

สรุปผลการวิจัย

ผลการวิจัยพบว่า การใส่ค่าสูญหายด้วยวิธีการประมาณค่าช่วงเส้นตรง (Linear Interpolation) มีประสิทธิภาพสูงที่สุดสำหรับรูปแบบการสูญหายทั้ง 3 รูปแบบ โดยที่สัดส่วนของการสูญหายของข้อมูลต่ำกว่าร้อยละ 50 ขณะที่สัดส่วนการสูญหายที่ร้อยละ 50 การใส่ค่าสูญหายด้วยวิธีป่าสุ่ม (Random Forest) มีประสิทธิภาพสูงที่สุด อย่างไรก็ตามการเลือกใช้วิธีการประมาณค่าช่วงเส้นตรงในการใส่ค่าสูญหายมีข้อจำกัดในการใช้งาน เมื่อข้อมูลที่จุดเวลาแรกและข้อมูลที่จุดเวลาสุดท้ายเกิดการสูญหายการประมาณค่าช่วงเส้นตรงอาจไม่ครอบคลุมในการเติมชุดข้อมูลให้สมบูรณ์ซึ่งต้องใช้วิธีการใส่ค่าสูญหายวิธีการอื่นในการเติมชุดข้อมูลให้สมบูรณ์ ขณะที่การใส่ค่าสูญหายด้วยวิธีป่าสุ่มเป็นอีกวิธีที่มีความเหมาะสมสำหรับอนุกรมเวลาเชิงพหุที่มีข้อได้เปรียบจากวิธีการประมาณค่าช่วงเส้นตรงคือไม่มีข้อจำกัดด้านข้อมูลที่จุดเวลาแรกและข้อมูลจุดเวลาสุดท้าย แต่มีระยะเวลาในการประมวลผลที่นานกว่าและให้ผลลัพธ์ในการใส่ค่าสูญหายแต่ละครั้งไม่เหมือนกันจึงจำเป็นต้องกำหนดค่าควบคุมการสุ่ม (Random Seed) ซึ่งการกำหนดค่าควบคุมการสุ่มอาจส่งผลให้ผลลัพธ์ในการใส่ค่าสูญหายดีหรือไม่ดีก็ได้ และการเลือกใช้วิธีการใส่ค่าสูญหายด้วยวิธีการเรียนรู้เชิงลึกอาจไม่เหมาะสมกับทุกข้อมูลอนุกรมเวลา และอาจไม่เหมาะสมกับทุกรูปแบบการสูญหายของข้อมูล ซึ่งอาจส่งผลให้ประสิทธิภาพในการใส่ค่าสูญหายของข้อมูลต่ำลงได้

ข้อเสนอแนะในการวิจัยครั้งถัดไป

1. ควรเพิ่มจำนวนข้อมูลอนุกรมเวลาเชิงพหุ เพื่อผลลัพธ์ที่ดีขึ้นเมื่อใช้การใส่ค่าสูญหายด้วยวิธีการเรียนรู้เชิงลึกเนื่องจากแบบจำลองด้านการเรียนรู้เชิงลึกจะทำงานได้ดีเมื่อข้อมูลมีขนาดใหญ่

2. ควรเพิ่มข้อมูลอนุกรมเวลาที่มีองค์ประกอบของฤดูกาล เพื่อศึกษาเพิ่มเติมว่าวิธีการใส่ค่าสูญหายวิธีการใดเหมาะสมกับข้อมูลอนุกรมเวลาที่มีองค์ประกอบของฤดูกาล

3. เนื่องจากในงานวิจัยครั้งนี้ได้ทำการกำหนดรูปแบบ และสัดส่วนการสูญหายจากข้อมูลจริงที่ไม่ได้สูญหายโดยธรรมชาติ จึงสามารถประเมินประสิทธิภาพโดยการเปรียบเทียบระหว่างค่าแทนที่การสูญหายด้วยวิธีการต่าง ๆ กับค่าจริงได้ ดังนั้นหากนำวิธีการประเมินประสิทธิภาพในงานวิจัยครั้งนี้ไปปรับใช้กับข้อมูลที่มีการสูญหายโดยธรรมชาติ ควรนำข้อมูลที่ผ่านการใส่ค่าสูญหายไปทำการพยากรณ์ด้วยแบบจำลองสำหรับข้อมูลอนุกรมเวลาเชิงพหุ เช่น VAR ARIMAX และ LSTM เป็นต้น เพื่อประเมินประสิทธิภาพในการพยากรณ์จากวิธีการใส่ค่าสูญหายวิธีการต่าง ๆ เพื่อหาวิธีการใส่ค่าสูญหายที่เหมาะสมที่สุด

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณอาจารย์ที่ปรึกษาจากภาควิชาสถิติ สาขาสถิติและวิทยาการข้อมูล คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยที่ให้คำแนะนำในการดำเนินงานวิจัยให้ประสบผลสำเร็จ และขอขอบคุณฐานข้อมูล SETSMART ของตลาดหลักทรัพย์แห่งประเทศไทย ที่ให้ความอนุเคราะห์ข้อมูลดัชนีราคาอุตสาหกรรมสำหรับงานวิจัยในครั้งนี้

เอกสารอ้างอิง

Ahn, H., Sun, K. and Kim, K.P. (2022). Comparison of Missing Data Imputation Methods in Time Series Forecasting. Computers, Materials and Continua 70(1): 767 - 779. doi: 10.32604/cmc.2022.019369.

Almeida, A., Brás, S., Sargento, S. and Pinto, F.C. (2024). Focalize K-NN: an imputation algorithm for time series datasets. Pattern Analysis and Applications 27(2): 39. doi: 10.1007/s10044-024-01262-3.

Alruhaymi, A.Z. and Kim, C.J. (2021). Why Can Multiple Imputations and How (MICE) Algorithm Work?. Open Journal of Statistics 11(5): 759 - 777. doi: 10.4236/ojs.2021.115045.

Bernstein, M.N. (2017). Random Forest. Source: https://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/ensembles/RandomForests. Retrieved date 12 September 2024.

Dong, Q., Chen, X. and Huang, B. (2024). Data Analysis in Pavement Engineering Methodologies and Applications. Cambridge: Woodhead Publishing. 181 - 196.

Du, W. (2023). PyPOTS: A Python Toolbox for Data Mining on Partially-Observed Time Series. arXiv: 2305.18911 doi: 10.48550/arXiv.2305.18811.

Du, W., Côté, D. and Liu, Y. (2023). SAITS: Self-attention-based imputation for time series. Expert Systems With Applications 219: 119619. doi: 10.1016/j.eswa.2023.119619.

Fortuin, V., Baranchuk, D., Rätsch, G. and Mandt S. (2020). GP-VAE: Deep Probabilistic Time Series Imputation. Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS) 108: 1651 - 1661.

Goldani, M. (2024). Comparative Analysis of Missing Values Imputation Methods: A Case Study in Financial Series (S&P500 and Bitcoin Value Data Sets). Iranian Journal of Finance 1(8): 47 - 70. doi: 10.61186/ijf.2024.414027.1427.

Hua, V., Nguyen, T., Dao, M., Nguyen, H.D. and Nguyen, B.T. (2024). The impact of data imputation on air quality prediction problem. PLoS ONE 19(9): e0306303. doi: 10.1371/journal.pone.0306303.

Kotu, V. and Deshpande, B. (2019). Data Science Concepts and Practice. (2nd ed.). Cambridge: Morgan Kaufmann Publishers. 395 - 443.

Little, R. and Rubin, D. (2019). Statistical Analysis with Missing Data. (3rd ed.). Hoboken: Wiley Series in Probability and Statistics. 3 - 23.

Miao, X., Wu, Y., Wang, J., Gao, Y., Mao, X. and Yin J. (2021). Generative Semi-supervised Learning for Multivariate Time Series Imputation. Proceedings of the AAAI Conference on Artificial Intelligence 35(10): 8983 - 8991. doi: 10.1609/aaai.v35i10.17086.

Ng, S.K., Krishnan, T. and McLachlan, G.J. (2012). The EM Algorithm. In J.E. Gentle, W.K. Härdle and Y. Mori (Eds.). Handbook of Computational Statistics: Concepts and Methods. Heidelberg: Springer-Verlag. 139 - 172.

❑❑❑❑❑