บทความวิจัย

การเติมค่าสูญหายข้อมูลฝนรายวันด้วยวิธีควอนไทล์

ศรีสุนี วุฒิวงศ์โยธิน*
ภาควิชาวิศวกรรมโยธา คณะวิศวกรรมศาสตร์ มหาวิทยาลัยบูรพา

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 0 3810 2222 ต่อ 3356 อีเมล: srisunee.wu@eng.buu.ac.th
DOI: 10.14416/j.kmutnb.2021.05.021
รับเมื่อ 26 มีนาคม 2563 แก้ไขเมื่อ 8 มิถุนายน 2563 ตอบรับเมื่อ 24 มิถุนายน 2563 เผยแพร่ออนไลน์ 25 พฤษภาคม 2564

บทคัดย่อ

ข้อมูลฝนเป็นข้อมูลพื้นฐานที่สำคัญในการศึกษาใดๆ ที่เกี่ยวข้องกับทรัพยากรน้ำ โดยเฉพาะอย่างยิ่งข้อมูลฝนรายวันมีลักษณะของข้อมูลแบบต่อเนื่อง และแบบไม่ต่อเนื่อง มีการแจกแจงความถี่ไม่ใช่การแจกแจงปกติ วิธีการเติมค่าสูญหายข้อมูลฝนมักใช้วิธีอย่างง่าย เช่น วิธีค่าเฉลี่ย หรือวิธีระยะทางผกผัน (Inverse Distance Weighting Method; IDW) วิธีการเหล่านี้มักมีข้อจำกัด 1) ให้ค่าปริมาณฝนรายวันต่ำกว่าความเป็นจริง 2) จำนวนวันฝนตกมากเกินไป และ 3) ไม่สามารถประมาณค่ากรณีเหตุการณ์ฝนตกหนักได้ การศึกษานี้จึงพัฒนาวิธีการเติมค่าข้อมูลฝนรายวันด้วยวิธีควอนไทล์ (Quantile Approach; QT) โดยใช้การแจกแจงความถี่แบบแบร์นูลี-แกมมา และเปรียบเทียบกับวิธีระยะทางผกผัน ผลการศึกษาพบว่า ค่าทางสถิติพื้นที่ต่างๆ ได้แก่ ค่าฝนรายวันสูงสุด ฝนเฉลี่ยรายวัน ความแปรปรวน การเติมค่าสูญหายด้วยวิธีควอนไทล์ ให้ค่าทางสถิติพื้นฐานดีกว่าวิธีระยะทางผกผัน อีกทั้งให้ค่าฝนที่เปอร์เซ็นต์ไทล์ 95 และ 99 ใกล้เคียงค่าตรวจวัดจริง ดังนั้นวิธีควอนไทล์สามารถประมาณค่ากรณีฝนตกหนักได้ดีกว่าวิธีระยะทางผกผัน นอกจากนี้การประเมินความแม่นยำในการทำนายค่าเหตุการณ์วันฝนตก และวันฝนไม่ตก วิธีควอนไทล์สามารถทำนายค่าได้แม่นยำกว่า ทั้งนี้ การเลือกใช้วิธีควอนไทล์เหมาะสำหรับการศึกษาใดๆ ที่ต้องพิจารณาถึงเหตุการณ์ฝนตกหนักซึ่งวิธีนี้สามารถให้ค่าและผลการศึกษาที่ถูกต้องมากกว่า

คำสำคัญ: การเติมค่าสูญหาย ฝนรายวัน วิธีระยะทางผกผัน วิธีควอนไทล์ ลุ่มน้ำปิงตอนบน

การอ้างอิงบทความ: ศรีสุนี วุฒิวงศ์โยธิน, "การเติมค่าสูญหายข้อมูลฝนรายวันด้วยวิธีควอนไทล์," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 31, ฉบับที่ 4, หน้า 599–613, ต.ค.–ธ.ค. 2564.

Research Article

Imputation of Missing Daily Rainfall Using Quantile Method

Srisunee Wuthiwongyothin*
Civil Engineering Department, Faculty of Engineering, Burapha University, Chon Buri, Thailand

* Corresponding Author, Tel. 0 3810 2222 Ext. 3356, E–mail: srisunee.wu@eng.buu.ac.th DOI: 10.14416/j.kmutnb.2021.05.021
Received 26 March 2020; Revised 8 June 2020; Accepted 24 June 2020; Published online: 25 May 2021

Abstract

Rainfall data is essential for any study related to water resources. Daily rainfall has its specific characteristics which is continuous time series and discrete data with non-normal distribution. Generally, methods to estimate missing daily rainfall data, for examples arithmetic mean, inverse distance weighting method (IDW) still have some limitations. Such founded limitations are: 1) underestimate of average daily rainfall, 2) overestimate of non-zero rainfall events, and 3) underestimate of extreme rainfall magnitude. This study attempts to develop an imputation method for daily rainfall using quantile approach (QT) which is based on Bernoulli-Gamma distribution, and then compare to IDW method. The study results reveal that QT could yield sample statistics such as maximum, mean, and variance of estimated daily rainfall better than IDW. In addition, the 95^{th} and 99^{th} percentiles of rainfall depths from QT method are closer to the observed data. Therefore, QT method is capable to estimate extreme rainfall magnitude superior than IDW approach. Moreover, QT gives a higher accuracy in number of zero and non-zero rainfall events. Using QT method might be appropriate for any study that concerns with extreme rainfall events since QT would give more accurate results.

Keywords: Imputation, Daily Rainfall, Inverse Distance Weighting Method, Quantile Method, Upper Ping River Basin

Please cite this article as: S. Wuthiwongyothin, "Imputation of missing daily rainfall using quantile method," *The Journal of KMUTNB*, vol. 31, no. 4, pp. 599–613, Oct.–Dec. 2021 (in Thai).

1. บทนำ

ข้อมูลฝนเป็นข้อมูลพื้นฐานที่สำคัญในการออกแบบวางแผน ด้านอุทกวิทยา การบริหารจัดการน้ำ การเกษตรและสิ่งแวดล้อม ตลอดจนภัยพิบัติที่เกี่ยวกับน้ำ [1], [2] การมีข้อมูลตรวจวัดฝนที่ดีมีความสำคัญอย่างมากในการเป็นค่าตัวแทนสำหรับการวิเคราะห์และประยุกต์ใช้ในการศึกษาวิจัยต่างๆ ที่เกี่ยวข้องกับทรัพยากรน้ำ ทั้งนี้ การตรวจวัดข้อมูลใดๆ มักประสบปัญหาการมีข้อมูลไม่สมบูรณ์หรือมีค่าสูญหาย (Missing Data หรือ Gap Data) สาเหตุการมีค่าสูญหายข้อมูลฝนรายวันสามารถเกิดได้จากผู้บันทึก อุปกรณ์ตรวจวัดหรือระบบต่างๆ ที่เกี่ยวข้องกับสถานีตรวจวัด ตลอดจนสภาพภูมิประเทศที่ตั้งสถานีและสภาพอากาศ [3] การมีค่าสูญหายข้อมูลฝนรายวันเป็นเรื่องที่เกิดขึ้นได้ทั่วไปและหลีกเลี่ยงได้ยากแม้จะมีความพยายามควบคุมการเกิดความผิดพลาดต่างๆ ในการบันทึกข้อมูลด้วยการออกแบบเครื่องมือเป็นอย่างดี [4] ข้อมูลฝนที่มีค่าสูญหายเมื่อนำไปใช้ศึกษาวิเคราะห์ผลต่างๆ อาจทำให้การศึกษานั้นๆ ได้ผลการวิเคราะห์ที่มีค่าลำเอียง (Bias) ดังนั้นการวิธีประมาณค่าสูญหายจึงมีความสำคัญเพื่อที่จะประมาณค่าและแทนที่ค่าสูญหายฝนรายวันให้ได้ผลลัพธ์ข้อมูลที่น่าเชื่อถือ

วิธีการเติมค่าสูญหาย (Imputation Method) ข้อมูลฝนรายวันที่นิยม ได้แก่ วิธีค่าเฉลี่ย (Arithmetic Averaging Method) วิธีอัตราส่วนปรกติ (Normal Ratio Method) และวิธีระยะทางผกผัน (Inverse Distance Weighting Method; IDW) วิธีการเหล่านี้เป็นแบบดั้งเดิมที่นิยมใช้ทั่วโลกเนื่องจากเป็นวิธีที่ง่ายและคำนวณได้รวดเร็ว (ใช้ทรัพยากรในการคำนวณน้อย) โดยเฉพาะอย่างยิ่งวิธี IDW จัดเป็นวิธีที่นิยมมากที่สุดวิธีหนึ่ง แต่วิธีการเหล่านี้เป็นวิธีการประมาณค่าในช่วงเชิงพื้นที่ (Spatial Interpolation) มีข้อจำกัด ได้แก่ มักให้ค่าปริมาณฝนรายวันที่ต่ำกว่าความเป็นจริง จำนวนวันฝนตกที่มากเกินไป ซึ่งเกิดจากการคำนวณค่าเฉลี่ยปริมาณฝนจากสถานีข้างเคียง ตัวอย่างเช่น ถ้าสถานีเป้าหมายในวันที่ค่าสูญหายไม่มีฝนตก แต่สถานีข้างเคียงมีปริมาณฝนตกแม้เพียงสถานีเดียว ค่าที่คำนวณได้ของสถานีเป้าหมายก็จะมีปริมาณฝนตกด้วยเช่นกัน (ทำให้จำนวนวันฝนตกมากเกินจริง) อีกทั้งไม่สามารถประมาณค่าที่ใกล้เคียงกรณีเกิดเหตุการณ์ฝนสุดขีด (Extreme Storm) ได้ [5] ปัญหาหลักดังกล่าวก่อให้เกิดความคลาดเคลื่อน (Error) เมื่อนำผลที่ได้ไปใช้ศึกษาวิเคราะห์

ข้อมูลฝนรายวันมีลักษณะแตกต่างจากข้อมูลฝนรายเดือน ลักษณะเฉพาะของข้อมูลฝนรายวันสามารถแบ่งได้เป็น 2 ลักษณะ 1) ข้อมูลอนุกรมเวลาแบบต่อเนื่อง (Continuous Time Series) ได้แก่ ปริมาณฝนตกที่วัดค่าได้หรือความลึกฝนตก (Rainfall Depth) และ 2) ข้อมูลแบบไม่ต่อเนื่อง (Discrete Data) ได้แก่ ข้อมูลจำนวนวันฝนตกและวันฝนไม่ตก [5] นอกจากนี้ข้อมูลปริมาณฝนตกมีการแจกแจงความถี่แบบไม่สมมาตร มีความแปรปรวนสูง โดยมีการศึกษาพบว่า การแจกแจกความถี่ที่เหมาะสมสำหรับข้อมูลปริมาณฝนรายวัน ได้แก่ การแจกแจงความถี่แบบแกมมา ดังในงานศึกษาวิจัยของ [5]–[8] เป็นต้น

ปัจจุบันมีการศึกษาและพัฒนาวิธีการเติมค่าสูญหายข้อมูลฝนรายวันด้วยวิธีการต่างๆ มากมาย ทั้งวิธีการอย่างง่ายและวิธีการที่ซับซ้อน เช่น การใช้วิธีการถดถอย (Regression Method) วิธีการถดถอยเชิงเส้นแบบพหุ (Multiple Linear Regression; MLR) การใช้วิธีการแจกแจกความถี่ การพัฒนาปรับปรุงวิธีรูปเหลี่ยมธีเอสเสน วิธีถ่วงน้ำหนักค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient Weighting; CCW) วิธีถ่วงน้ำหนักค่าพิกัด (Geographic Coordinate) ซึ่งวิธี CCW และวิธีถ่วงน้ำหนักค่าพิกัดพัฒนามาจากวิธี IDW [2] นอกจากนี้ยังมีการพัฒนาวิธีการอื่นๆ อีกมากมาย

Simolo และคณะ [7] เสนอวิธีปรับปรุงการเติมค่าสูญหายข้อมูลฝนรายวันด้วยวิธีการวิเคราะห์ถดถอยเชิงเส้นพหุคูณ (Multi-Linear Regression; MLR) ร่วมกับการแจกแจงความถี่ โดยทำการจำแนกจำนวนวันฝนตก และฝนไม่ตก และหาค่าเฉลี่ยโดยมีการถ่วงน้ำหนักจากสถานีข้างเคียงประมาณค่าปริมาณฝนตกด้วยวิธี MLR และนำค่าที่ได้จาก MLR ไปปรับแก้อีกครั้งด้วยการแจกแจงความถี่แบบแกมมา การศึกษาดังกล่าวทำการทดสอบกับสถานีตรวจวัดข้อมูลฝนรายวันจำนวน 36 สถานี ในลุ่มน้ำ Reno ตั้งอยู่ทางตอนเหนือของประเทศอิตาลี ตั้งแต่ ค.ศ. 1916–2004 โดยที่แต่ละ

สถานีมีค่าสูญหายข้อมูลไม่เกิน 8% ผลการศึกษาสรุปว่าวิธีการนี้สามารถสร้างข้อมูลน้ำฝนรายวันขึ้นมาใหม่ โดยให้ค่าที่น่าเชื่อถือทั้งเวลาในการเกิดและปริมาณฝนที่ตก และสามารถคงคุณลักษณะค่าทางสถิติของชุดข้อมูลได้ วิธีดังกล่าวเป็นวิธีที่ใช้เวลาในการคำนวณน้อยซึ่งเป็นประโยชน์ต่อการนำไปใช้สร้างชุดข้อมูลที่มีขนาดใหญ่ได้

การศึกษาเปรียบเทียบวิธีการเติมค่าสูญหายข้อมูลฝนรายวันด้วยวิธีทางสถิติกับวิธีระยะทางผกผัน (IDW) โดย [5] ศึกษาในประเทศอินเดีย ใช้สถานีตรวจวัดฝนจำนวน 20 สถานี ตั้งอยู่ในลุ่มน้ำ Brahmani เมือง Rachi ประเทศอินเดีย วิธีทางสถิติใช้การแจกแจงความถี่แบบปัวซง-แกมมา (The Poisson-Gamma; PG) ผลการศึกษาพบว่า ค่าเฉลี่ยและเปอร์เซ็นต์จำนวนวันฝนไม่ตกระหว่างข้อมูลตรวจวัดและข้อมูลที่ทำการเติมมีค่าใกล้เคียงกัน อย่างไรก็ตาม การแจกแจงแบบปัวซง-แกมมา ให้ค่าที่ประมาณได้สูงกว่าในช่วงเปอร์เซ็นต์ไทล์ที่ 95% แต่จะให้ค่าต่ำกว่าในช่วงเปอร์เซ็นต์ไทล์ที่ 99% บ่งชี้ได้ว่าการเติมข้อมูลฝนด้วยวิธีการแจกแจงข้อมูลแบบปัวซง-แกมมา ไม่สามารถประมาณค่าฝนกรณีเหตุการณ์ฝนตกสูงๆ เรียก ฝนสุดขีด (Extremely Heavy Rainfall) ได้ อย่างไรก็ตาม เมื่อเปรียบเทียบการเติมข้อมูลฝนรายวันที่ศึกษาทั้ง 2 วิธี พบว่า วิธีการเติมข้อมูลฝนโดยใช้การแจกแจงแบบปัวซง-แกมมาให้ผลที่ดีกว่าการประมาณค่าฝนด้วยวิธีระยะทางผกผัน (IDW)

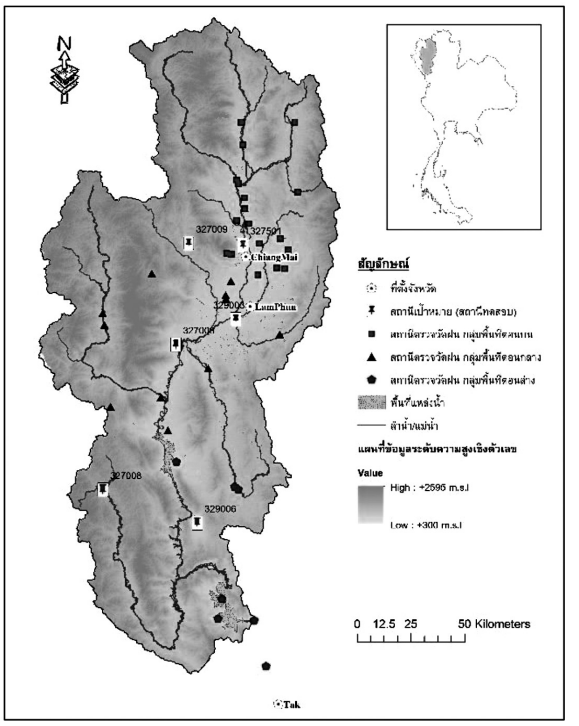
Kim และ Ryu [6] ศึกษาวิธีการเพื่อปรับปรุงการเติมข้อมูลฝนรายวันโดยใช้วิธีการแจกแจงความถี่แบบแกมมาร่วมกับการหาความสัมพันธ์ทางสถิติ แบ่งเป็น 2 ขั้นตอนหลัก ได้แก่ ขั้นตอนแรกทำการวิเคราะห์การจัดกลุ่ม (Cluster Analysis) สถานีตรวจวัด และขั้นตอนที่สองทำการประมาณค่าและเติมค่าสูญหายข้อมูลฝนรายวัน โดยการสร้างกราฟการแจกแจกความถี่สะสม (Cumulative Distribution Function; CDF) ของทั้งสถานีอ้างอิงและสถานีเป้าหมาย และวิธีประมาณค่าฝนรายวันแบบเดิมอีก 3 วิธี ได้แก่ วิธีระยะทางผกผัน (IDW) วิธี Ordinary Kriging (OK) และวิธีประมาณค่าเฉลี่ยจากสถานีข้างเคียง (Gauge Mean Estimator; GME) ศึกษาในรัฐไอดาโฮ ประเทศสหรัฐอเมริกา ซึ่งมีลักษณะอากาศแตกต่างกันระหว่างพื้นที่ฝั่งตะวันตกและตะวันออก มีสถานีตรวจอากาศมากกว่า 150 สถานี และเลือกใช้เฉพาะสถานีที่มีค่าสูญหายน้อยกว่า 15% ในการศึกษามีการเปรียบเทียบผลการศึกษาทั้งแบบมีการวิเคราะห์การจัดกลุ่มและไม่มีการจัดกลุ่มสถานี ผลการศึกษาชี้ได้ชัดเจนว่าการวิเคราะห์การจัดกลุ่มสถานีสามารถปรับปรุงประสิทธิภาพในการประมาณค่าเพื่อเติมข้อมูลได้ดีกว่าในทุกวิธี โดยเฉพาะอย่างยิ่งวิธีการแจกแจงความถี่แบบแกมมาร่วมกับการวิเคราะห์การจัดกลุ่มเป็นวิธีที่มีประสิทธิภาพสูงกว่าวิธีอื่น และเมื่อเปรียบเทียบจำนวนวันฝนตกและจำนวนวันฝนไม่ตกในพื้นที่ศึกษาให้ผลใกล้เคียงกว่าวิธีการแบบเดิมอีก 3 วิธี

ในประเทศไทยมีการศึกษาการเติมค่าสูญหายข้อมูลฝนรายวันไม่มากนัก การศึกษาส่วนใหญ่มักจะเน้นการเติมค่าสูญหายโดยใช้ค่าเฉลี่ยจากสถานีข้างเคียง มีข้อจำกัดในการให้ค่าฝนเฉลี่ยที่น้อยกว่าค่าจริง จำนวนวันฝนตกมากกว่าความจริง และไม่สามารถประมาณค่ากรณีเหตุการณ์ฝนสุดขีด (Extreme Rainfall) ได้ดังที่กล่าวมาแล้วนั้น ดังนั้นการศึกษานี้มีวัตถุประสงค์เพื่อ 1) ศึกษาวิธีการเติมข้อมูลฝนโดยวิธีควอนไทล์ ซึ่งใช้การวิเคราะห์แจกแจงความถี่ผสมแบบแบร์นูลี-แกมมา (Bernoulli-Gamma Distribution) ร่วมกับการวิเคราะห์เพื่อจัดกลุ่มข้อมูลสถานี (Cluster Analysis; CA) และ 2) เปรียบเทียบการประมาณค่าสูญหายข้อมูลฝนรายวันวิธีควอนไทล์กับวิธีที่นิยมใช้ทั่วไป ได้แก่ วิธีระยะทางผกผัน (Inverse Distance Weighting) ในพื้นที่ลุ่มน้ำปิงตอนบน

2. วัสดุ อุปกรณ์และวิธีการวิจัย

2.1 พื้นที่ศึกษา และข้อมูลที่ใช้ในการศึกษา

พื้นที่ศึกษา ได้แก่ ลุ่มน้ำปิงตอนบน มีพื้นที่ประมาณ 26,111 ตารางกิโลเมตร (รูปที่ 1) ปริมาณน้ำจากลุ่มน้ำปิงตอนบน คิดเป็น 75 เปอร์เซ็นต์ของทั้งลุ่มน้ำปิง และไหลลงเขื่อนภูมิพล สภาพภูมิประเทศของลุ่มน้ำปิงส่วนใหญ่เป็นภูเขา ปกคลุมด้วยป่าไม้กึ่งเขตร้อน มีแนวเทือกเขาทางด้านทิศตะวันออกและทิศตะวันตกเป็นแหล่งกำเนิดแม่น้ำสาขาที่สำคัญของแม่น้ำปิง ระดับความสูงสูงสุดประมาณ +2,595



รูปที่ 1 พื้นที่ศึกษา และตำแหน่งสถานีตรวจวัดน้ำฝนที่คัดเลือกสำหรับเติมค่าสูญหาย

ม.รทก ที่ดอยอินทนนท์ในจังหวัดเชียงใหม่ ระดับความสูงต่ำสุดประมาณ +300 ม.รทก บริเวณอ่างเก็บน้ำเขื่อนภูมิพล ซึ่งเป็นแหล่งเก็บน้ำหลักเพื่ออุปโภค-บริโภค การเกษตร อุตสาหกรรม และอื่นๆ ในพื้นที่ภาคกลางเขตลุ่มน้ำเจ้าพระยา

ข้อมูลที่ใช้ศึกษา ได้แก่ ข้อมูลฝนรายวันจากสถานีตรวจวัดภาคพื้นดินที่ตั้งอยู่ในลุ่มน้ำปิงตอนบน รวบรวมได้จำนวน 92 สถานี เป็นสถานีตรวจวัดที่อยู่ภายใต้การดูแลรับผิดชอบโดยกรมอุตุนิยมวิทยา กรมชลประทาน และกรมทรัพยากรน้ำ (รูปที่ 1) ช่วงเวลาที่ศึกษา ได้แก่ มกราคม 2496 ถึงธันวาคม 2560 (65 ปี)

2.2 วิธีการและขั้นตอน

การศึกษาแบ่งเป็น 4 ขั้นตอนหลัก ได้แก่ 1) การวิเคราะห์จัดกลุ่มสถานีด้วยวิธี K-means (K-means Cluster Analysis) เพื่อแบ่งกลุ่มสถานี 92 สถานี และคัดเลือกสถานีตัวแทนเพื่อศึกษา 2) สร้างค่าสูญหายข้อมูลให้กับสถานีเป้าหมาย 3) ทดสอบการเติมค่าสูญหายข้อมูลฝนรายวันด้วยวิธีควอนไทล์ (QT) และวิธี IDW ให้กับสถานีเป้าหมาย และ 4) ประยุกต์การเติมค่าสูญหายให้กับสถานีฝนในลุ่มน้ำปิงตอนบนที่มีค่าสูญหายจริงไม่เกิน 50% (แสดงตำแหน่งสถานีในรูปที่ 1) รายละเอียดแต่ละขั้นตอนการศึกษามีดังนี้

2.2.1 การวิเคราะห์จัดกลุ่มสถานี (Cluster Analysis) และการคัดเลือกสถานีเป้าหมาย สถานีอ้างอิง

การวิเคราะห์จัดกลุ่มสถานี พิจารณาความคล้ายคลึงของสถานีจากระยะห่างระหว่างสถานีเป็นหลัก ในการศึกษานี้ใช้วิธีการจัดกลุ่มแบบ K-means (K-mean Clustering) เนื่องจากมีผลการศึกษาพบว่า การจัดกลุ่มด้วยวิธี K-means สามารถปรับปรุงประสิทธิภาพในการประมาณค่า เพื่อเติมค่าสูญหายข้อมูลฝนให้ดีขึ้นได้ดังเช่นในงานวิจัยของ Kim และ Ryu [6] และที่เสนอโดย Teegavarapu [9]

K-means เป็นการวิเคราะห์การจัดกลุ่มแบบไม่เป็นขั้นตอน (Nonhierarchical Cluster Analysis) โดยอัลกอลิทึม K-Means จะตัดแบ่งสถานีออกเป็น K กลุ่ม จากนั้นแทนค่าแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่ม ซึ่งใช้เป็นจุดศูนย์กลาง (Centroid) ของกลุ่ม โดยวัดจากระยะห่างของสถานีในกลุ่มเดียวกัน ทำการแบ่งสถานีให้อยู่ในแต่ละกลุ่มและคำนวณหาจุดศูนย์กลางใหม่ ทำซ้ำเรื่อยๆ จนสมาชิกในแต่ละกลุ่มและค่าจุดศูนย์กลางไม่มีการเปลี่ยนแปลง

สมการการวิเคราะห์จัดกลุ่มข้อมูลด้วยวิธี K-means แสดงดังสมการที่ (1)

$$SSE = \sum_{i=1}^{K} \sum_{n \in C_i} |x - m_i|^2 \quad (1)$$

เมื่อ SSE คือ ค่าผลต่างกำลังสอง (Sum of Squared Error)
X คือ ข้อมูลในกลุ่ม C_i
m_i คือ จุด Centroid สำหรับกลุ่ม C_i
K คือ จำนวนกลุ่ม

การแบ่งกลุ่มสถานีฝนทั้ง 92 สถานี ในพื้นที่ลุ่มน้ำปิงตอนบน กำหนดให้แบ่งเป็น 3 กลุ่ม (K=3) ได้แก่ พื้นที่ตอนบน ตอนกลาง และตอนล่าง จากนั้นทำการคัดเลือกสถานีตัวแทนกลุ่มละ 2 สถานี รวม 6 สถานีเป้าหมาย (ดังรูปที่ 1) เพื่อใช้ทดสอบการเติมค่าสูญหาย คัดเลือกจากสถานีที่มีจำนวน

ข้อมูลมากที่สุดในกลุ่ม (ม.ค. 2496 ถึง ธ.ค. 2560) กำหนดให้สถานีทดสอบสำหรับเติมค่าสูญหายเรียกว่า “สถานีเป้าหมาย” (Target Station; TS) และสถานีข้างเคียงที่จะนำค่ามาใช้ในการคำนวณ เรียกว่า “สถานีอ้างอิง” (Source Station; SS)

2.2.2 การสร้างค่าสูญหายให้กับสถานีเป้าหมาย

สถานีเป้าหมายที่คัดเลือก 6 สถานี นำมากรองค่าข้อมูลให้เหลือเฉพาะวันที่มีค่าตรวจวัดจริงโดยตัดค่าสูญหายจริงออก จากนั้นทำการสร้างค่าสูญหาย (NA) แบบสุ่ม (Missing Random, MAR) เพื่อจำลองการสูญหายให้กับค่าข้อมูลฝนรายวัน โดยทั่วไปกลไกการสูญหายข้อมูล (Missingness Mechanism) [10] หรือรูปแบบการสูญหายข้อมูล (Pattern of Missing Data) [11] มีผลต่อวิธีการเติมค่าสูญหาย แต่จากงานศึกษาของ Presti และคณะ [4] พบว่า ลักษณะการสูญหายข้อมูลฝนรายวันจากสถานีตรวจวัดภาคพื้นดินมีลักษณะการสูญหายแบบสุ่ม (MAR) การศึกษาวิจัยนี้ทำการทดสอบการเติมค่าสูญหายแบบสุ่ม (MAR) ที่เปอร์เซ็นต์การสูญหายที่แตกต่างกัน ได้แก่ 5%, 10%, 20%, 30%, 40% และ 50%

2.2.3 การเติมค่าสูญหายด้วยวิธีระยะทางผกผัน (IDW)

วิธี IDW เป็นการประมาณค่าในช่วงเชิงพื้นที่ (Spatial Interpolation Method) นิยมใช้ค่าจากสถานีที่ทราบค่าที่อยู่ใกล้เคียงจำนวน 3–4 สถานี มาประมาณค่า ซึ่งการใช้ค่าจาก 4 สถานี จะให้ค่าที่ดีกว่า [12] ดังนั้นในการศึกษาวิจัยนี้จึงเลือกใช้ 3 สถานี ในการประมาณค่า การคำนวณด้วยวิธี IDW แสดงดังสมการที่ (2)

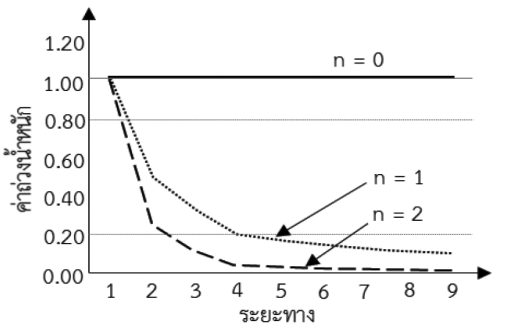
$$\hat{x} = \frac{\sum_{i=1}^{m}(\frac{x_i}{d_i^n})}{\sum_{i=1}^{m}(\frac{1}{d_i^n})} \quad (2)$$

เมื่อ d_i คือ ระยะระหว่างสถานีเป้าหมายกับสถานีอ้างอิง

x_i คือ ปริมาณน้ำฝนจากสถานีอ้างอิงที่ทราบค่า (มม.)

$\hat{x}$ คือ ปริมาณน้ำฝนที่ประมาณค่าได้ (มม.)

จากสมการที่ (2) วิธี IDW ให้ความสำคัญกับสถานีที่อยู่ใกล้มากกว่าสถานีที่อยู่ห่างไกลออกไปด้วยการใช้ค่าถ่วงน้ำหนัก ได้แก่ ส่วนกลับของระยะทางระหว่างสถานีเป้าหมายกับสถานีอ้างอิง การเพิ่มขึ้นหรือลดลงของค่าถ่วงน้ำหนักขึ้นอยู่กับค่ายกกำลัง n ถ้า n=0 แสดงถึงไม่มีการลดลงของระยะทาง



รูปที่ 2 ความสัมพันธ์ระหว่างค่าถ่วงน้ำหนักและระยะทาง

ซึ่งได้แก่ วิธีค่าเฉลี่ยนั่นเอง แต่ถ้า n เพิ่มขึ้น ค่าถ่วงน้ำหนักสำหรับระยะห่างระหว่างสถานีจะลดลงอย่างรวดเร็วหรือมีความสำคัญน้อยนั่นเอง ดังรูปที่ 2 ดังนั้นในงานศึกษานี้เลือกใช้ค่าเลขยกกำลัง n=2 ซึ่งนิยมใช้โดยทั่วไป

2.2.4 การเติมค่าสูญหายด้วยวิธีควอนไทล์ (QT)

วิธีควอนไทล์ที่เสนอในการศึกษาวิจัยนี้ กระทำโดยการสร้างฟังก์ชันความน่าจะเป็น (Probability Distribution Function; PDF) และฟังก์ชันความน่าจะเป็นสะสม (Cumulative Distribution Function; CDF) ของสถานีเป้าหมาย (CDF-TS) และสถานีอ้างอิง (CDF-SS) จากข้อมูลฝนรายวันของแต่ละสถานีดังกล่าว โดยสถานีอ้างอิง ทำการเลือกจากสถานีข้างเคียงที่มีช่วงเวลาข้อมูลครอบคลุมช่วงเวลาค่าสูญหายของสถานีเป้าหมายมากที่สุด และมีค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient; r) ของฝนรายเดือนระหว่างสถานีอ้างอิงและสถานีเป้าหมายที่ดีที่สุด การแจกแจงความถี่ข้อมูลน้ำฝนรายวันที่ใช้ในการศึกษานี้เลือกใช้การแจกแจงความถี่แบบผสม ได้แก่

- การแจกแจงแบร์นูลี (Bernoulli Distribution) และ
- การแจกแจงแกมมา (Gamma Distribution)

เมื่อ x คือ ปริมาณฝนรายวัน กำหนดให้ฝนตกมากกว่า 0.1 มม. เป็นวันที่มีฝนตก การแจกแจงแบบแบร์นูลีใช้สำหรับการหาความน่าจะเป็นของโอกาสการเกิดวันฝนตก (π) และวันฝนไม่ตก ($1-\pi$) ส่วนการแจกแจงความถี่แบบแกมมาใช้สำหรับวิเคราะห์รูปแบบการกระจายตัวของฝนรายวันของวันที่ฝนตกซึ่งเป็นที่นิยมใช้ในด้านอุทกวิทยาและสำหรับ

ข้อมูลฝน [5], [8], [13] ฟังก์ชันการแจกแจงความถี่ (PDF) แบร์นูลี-แกมมา แสดงดังสมการที่ (3)–(5)

$$g(x) = \begin{cases} \pi * \gamma(x); \text{ ถ้า } x > 0.1 \\ 1 - \pi; \text{ ถ้า } x \leq 0.1 \end{cases} \quad (3)$$

โดยที่ π คือ ความน่าจะเป็นของวันฝนตก
$1\text{-}\pi$ คือ ความน่าจะเป็นของวันฝนไม่ตก
$\gamma(x)$ คือ การแจกแจงความถี่แบบแกมมา

ซึ่ง $\gamma(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{(-\frac{x}{\beta})}$ (4)

และ $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$ เมื่อ $\alpha > 0$ (5)

หรือ $\Gamma(\alpha)=(\alpha-1)!$
เมื่อ α คือ พารามิเตอร์รูปร่าง (Shape Parameter)
และ β คือ พารามิเตอร์มาตราส่วน (Scale Parameter)

สมการการแจกแจงความถี่แบบสะสม (Cumulative Distribution Function; CDF) แบบแบร์นูลี-แกมมาดังสมการที่ (6)

$$G(x) = \begin{cases} 1 - \pi + \pi * \Gamma(x) \text{ ; ถ้า } x > 0.1 \\ 1 - \pi \qquad \text{ ; ถ้า } x \leq 0.1 \end{cases} \quad (6)$$

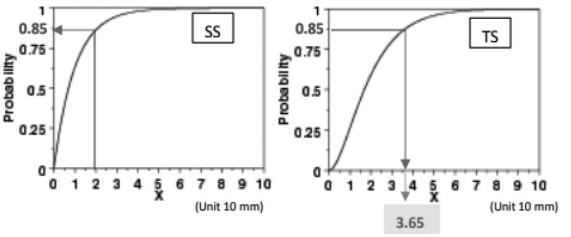
เมื่อ $\Gamma(x)$ คือ ฟังก์ชันแกมมาของฝนรายวัน

ฟังก์ชันควอนไทล์ หรือส่วนกลับของ CDF (Inverse CDF) เขียนเป็นสมการดังสมการที่ (7)

$$G^{-1}(x) = \begin{cases} \Gamma^{-1}\left(\frac{p-1+\pi}{\pi}\right) \text{ ; ถ้า } \pi > 1 - p \\ 0 \qquad \text{ ; ถ้า } p \leq 1 - p \end{cases} \quad (7)$$

เมื่อ $\Gamma^{-1}(p)$ คือ ส่วนกลับของ CDF แบบแกมมา
p คือ ค่าความน่าจะเป็น

การประมาณค่าสูญหายข้อมูลฝนรายวันโดยวิธี QT เพื่อเติมค่าให้กับสถานีเป้าหมายมีวิธีการดังนี้ กำหนดให้วันที่มีค่าสูญหายของสถานีเป้าหมาย (TS) คือ วันที่ 30 เม.ย. 2543 สมมุติให้ปริมาณฝนตกในวันที่ 30 เม.ย. 2543



รูปที่ 3 ตัวอย่างการอ่านค่าเพื่อเติมค่าสูญหายฝนรายวันด้วยวิธีควอนไทล์

ของสถานี SS เท่ากับ 20 มิลลิเมตร นำไปอ่านค่าความน่าจะเป็นหรือโอกาสการเกิดฝนตกจากกราฟ CDF-SS ได้ 0.85 (85%) จากนั้นนำค่าความน่าจะเป็น 0.85 ดังกล่าวไปอ่านค่าปริมาณน้ำฝนจากกราฟการแจกแจงความถี่แบบสะสมของสถานีเป้าหมาย (CDF-TS) จะได้ค่าน้ำฝนรายวัน 36.5 มม. และนำค่านี้ไปเติมให้กับสถานีเป้าหมายในวันที่ 30 เม.ย. 2543 ดังตัวอย่างรูปที่ 3

2.2.5 การประเมินประสิทธิภาพและความแม่นยำ (Performance and Accuracy Assessment)

การประเมินประสิทธิภาพทำโดยการเปรียบเทียบค่าสถิติพื้นฐานระหว่างข้อมูลที่ประมาณค่าได้กับข้อมูลตรวจวัดจริง ได้แก่ ค่าฝนรายวันสูงสุด ค่าเฉลี่ยฝนรายวัน ความแปรปรวน ค่าฝนที่เปอร์เซ็นต์ไทล์ 95 และ 99 การวัดค่าความคลาดเคลื่อนกระทำโดยวิธีค่ารากที่สองความคลาดเคลื่อนเฉลี่ยกำลังสอง (Root Mean Square Error; RMSE) และค่าความคลาดเคลื่อนเฉลี่ยสมบูรณ์ (Mean Absolute Error; MAE) ดังสมการที่ (8) และ (9)

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

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |x_i - \hat{x}_i| \quad (9)$$

โดยที่ x_i คือ ปริมาณฝนจากค่าตรวจวัดจริง
$\hat{x}_i$ คือ ปริมาณฝนที่ได้จากการประมาณค่า
n คือ จำนวนข้อมูลทั้งหมด (ที่เติมค่าสูญหาย)

นอกจากนี้การประเมินความแม่นยำ ทำการวัดจาก

คะแนนทักษะ (Skill Scores) ซึ่งคำนวณจากการนับความแม่นยำของจำนวนเหตุการณ์วันฝนตก และวันที่ฝนไม่ตก ($x \leq 0.1$) กำหนดให้

$C_{0,0}$ คือ เหตุการณ์ที่ค่าตรวจวัดจริงฝนไม่ตก และค่าที่คำนวณได้ฝนไม่ตก

$C_{1,1}$ คือ เหตุการณ์ที่ค่าตรวจวัดจริงมีฝนตก และค่าที่คำนวณได้มีฝนตกเช่นกัน

$C_{0,1}$ คือ เหตุการณ์ที่ค่าตรวจวัดจริงฝนไม่ตก แต่ค่าที่คำนวณได้มีฝนตก

$C_{1,0}$ คือ เหตุการณ์ที่ค่าตรวจวัดจริงมีฝนตก แต่ค่าที่คำนวณได้ฝนไม่ตก

ตัวเลขห้อยตำแหน่งแรก หมายถึง ค่าตรวจวัดจริง

ตัวเลขห้อยตำแหน่งที่สอง หมายถึง ค่าคำนวณ

เลข 0 หมายถึง เหตุการณ์วันฝนไม่ตก

เลข 1 หมายถึง เหตุการณ์วันฝนตก

การวัดความแม่นยำด้วยคะแนนทักษะ (Skill Scores) ต่างๆ มีดังนี้

1) ความแม่นยำในทำนายวันฝนไม่ตก (Correct State for Dry Time Intervals; P_d) P_d = 1 แม่นยำมากที่สุด

$$P_d = \frac{\sum C_{0,0}}{n_0} \tag{10}$$

เมื่อ n_0 คือ จำนวนวันที่ฝนไม่ตกทั้งหมดจากค่าตรวจวัด

2) ความแม่นยำในการทำนายเหตุการณ์ได้ถูกต้องทั้งหมด (Correct State for All Time Intervals; P_a) P_a = 1 แม่นยำหรือถูกต้องทั้งหมด

$$P_a = \frac{\sum(C_{0,0}+C_{1,1})}{n} \tag{11}$$

3) ความคลาดเคลื่อน หรืออัตราความคลาดเคลื่อน (Error Rate) ในการทำนายเหตุการณ์

$$Error\ rate = \frac{\sum(C_{0,1}+C_{1,0})}{n} \tag{12}$$

อัตราการคลาดเคลื่อนยิ่งน้อยยิ่งดี

โดยที่ n จากสมการที่ (11) และ (12) คือ จำนวนข้อมูลทั้งหมดที่เติมค่าซึ่งเท่ากับจำนวนค่าสูญหาย

4) คะแนนความลำเอียง (Bias Score; BS) ของจำนวนเหตุการณ์วันฝนตกแสดงดังสมการที่ (13)

$$Bias\ score = \frac{forecast\ events}{observed\ events}$$

$$= \frac{\sum(C_{1,1}+C_{0,1})}{\sum(C_{1,1}+C_{1,0})} \tag{13}$$

ค่า *BS* = 1 สามารถทำนายจำนวนวันฝนตกได้ไม่ลำเอียง หรือถูกต้องทั้งหมด ค่า *BS* < 1 ทำนายเหตุการณ์วันฝนตกได้ต่ำกว่าความเป็นจริง และ *BS* > 1 ทำนายเหตุการณ์ได้มากกว่าเหตุการณ์จริง

3. ผลการทดลอง

3.1 จำนวนข้อมูลค่าสูญหายของสถานีเป้าหมาย

จากสถานีที่คัดเลือกเพื่อใช้ทดสอบการเติมค่าจำนวน 6 สถานีเป้าหมาย ได้แก่ สถานี 327008, 329006, 327003, 329003, 327009 และ 327501 ทำการตัดค่าสูญหายจริงคงเหลือเฉพาะวันที่มีค่าตรวจวัด จากนั้นนำมาสร้างค่าสูญหายแบบสุ่ม (MAR) และไม่ต่อเนื่องให้มีเปอร์เซ็นต์ค่าสูญหายแตกต่างกันที่ 5%, 10%, 20%, 30%, 40% และ 50% แสดงดังตารางที่ 1

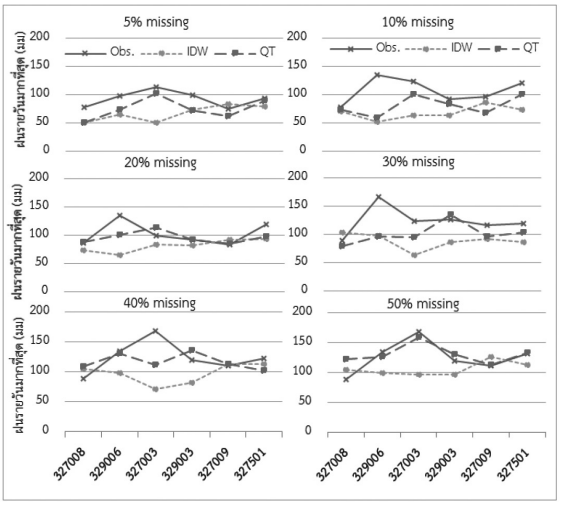
3.2 การประเมินค่าทางสถิติและความคลาดเคลื่อน

เมื่อทำการเติมค่าสูญหายด้วยวิธี IDW และ QT แล้วนำค่าที่ประมาณค่าได้มาแทนที่ค่าสูญหาย และนำมาเปรียบเทียบกับค่าตรวจวัดจริง ด้วยการประเมินค่าทางสถิติต่างๆ ได้แก่ ค่าฝนรายวันสูงสุด ฝนรายวันเฉลี่ย ความแปรปรวน ปริมาณฝนที่เปอร์เซ็นต์ไทล์ที่ 95 และ 99 และการประเมินความคลาดเคลื่อนด้วยค่า *RMSE* และ *MAE* ผลที่ได้แสดงดังนี้

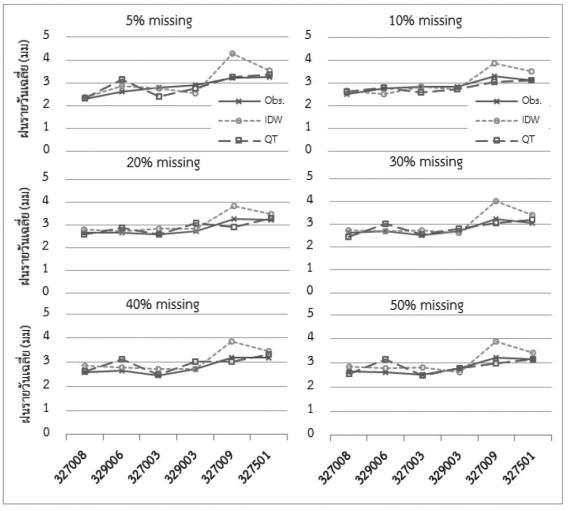
จากกราฟรูปที่ 4 เปรียบเทียบฝนรายวันมากที่สุด

ตารางที่ 1 จำนวนข้อมูลที่ใช้ศึกษาการประมาณค่าสูญหายที่เปอร์เซ็นต์การสูญหายต่างๆ

จำนวน	จำนวนข้อมูลที่เปอร์เซ็นต์ค่าสูญหาย (NA) ต่างๆ					
	5%	10%	20%	30%	40%	50%
สถานีเป้าหมาย (TS): LTS1 (327008)						
ค่าสูญหาย	1,075	2,151	4,302	6,453	8,604	10,756
ข้อมูลที่มีค่า	20,438	19,362	17,211	15,060	12,909	10,757
ข้อมูลทั้งหมด	21,513	21,513	21,513	21,513	21,513	21,513
สถานีเป้าหมาย (TS): LTS2 (329006)						
ค่าสูญหาย	1,044	2,088	4,176	6,264	8,352	10,440
ข้อมูลที่มีค่า	19,837	18,793	16,705	14,617	12,529	10,441
ข้อมูลทั้งหมด	20,881	20,881	20,881	20,881	20,881	20,881
สถานีเป้าหมาย (TS): MTS1(327003)						
ค่าสูญหาย	1,165	2,331	4,662	6,993	9,325	11,656
ข้อมูลที่มีค่า	22,149	20,983	18,652	16,321	13,989	11,658
ข้อมูลทั้งหมด	23,314	23,314	23,314	23,314	23,314	23,314
สถานีเป้าหมาย (TS): MTS2(329003)						
ค่าสูญหาย	1,106	2,213	4,426	6,640	8,853	11,067
ข้อมูลที่มีค่า	21,029	19,922	17,709	15,495	13,282	11,068
ข้อมูลทั้งหมด	22,135	22,135	22,135	22,135	22,135	22,135
สถานีเป้าหมาย (TS): UTS1(327009)						
ค่าสูญหาย	1,170	2,340	4,681	7,021	9,361	11,702
ข้อมูลที่มีค่า	22,236	21,066	18,725	16,385	14,045	11,704
ข้อมูลทั้งหมด	23,406	23,406	23,406	23,406	23,406	23,406
สถานีเป้าหมาย (TS) UTS2(327501)						
ค่าสูญหาย	1,159	2,318	4,637	6,954	9,274	11,591
ข้อมูลที่มีค่า	22,027	20,868	18,549	16,232	13,912	11,595
ข้อมูลทั้งหมด	23,186	23,186	23,186	23,186	23,186	23,186



รูปที่ 4 ฝนรายวันสูงสุดระหว่างค่าตรวจวัดจริง (Obs) และค่าที่ทำการเติมด้วยวิธี IDW และ QT



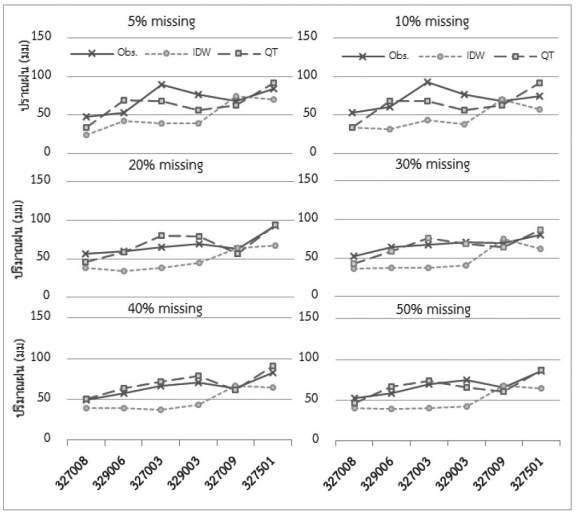
รูปที่ 5 ฝนเฉลี่ยรายวันระหว่างค่าตรวจวัดจริง (Obs) และค่าที่ทำการเติมด้วยวิธี IDW และ QT

ระหว่างค่าตรวจวัดจริงและวิธีการเติมค่าสูญหาย สังเกตได้ว่าวิธี QT (เส้นประห่าง) มักให้ค่าฝนรายวันมากที่สุดสูงกว่าวิธี IDW (เส้นประถี่) และใกล้เคียงกับค่าตรวจวัดจริงมากกว่า

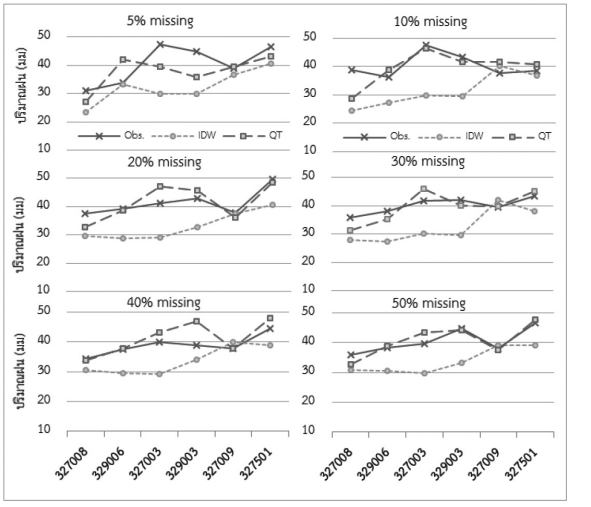
รูปที่ 5 เปรียบเทียบฝนรายวันเฉลี่ยระหว่างวิธี QT (เส้นประห่าง) และ IDW (เส้นประถี่) พบว่า วิธี QT ให้ค่าสอดคล้องใกล้เคียงกับค่าตรวจวัดจริงมากกว่า

เมื่อทำการเปรียบเทียบค่าความแปรปรวนฝนรายวันที่ทำการเติมค่า ดังรูปที่ 6 พบว่า วิธี QT ให้ค่าความแปรปรวนสูงและสอดคล้องกับค่าตรวจวัดจริง ขณะที่วิธี IDW มักให้ค่าความแปรปรวนของฝนรายวันต่ำกว่าค่าจริง

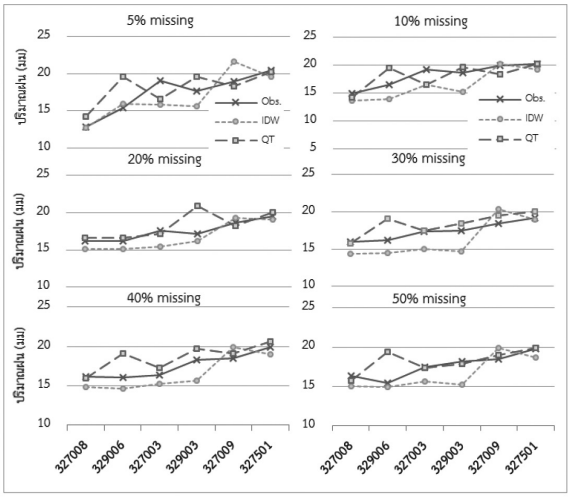
นอกจากนี้การพิจารณาค่าเปอร์เซ็นต์ไทล์ 95 และ 99 ระหว่างวิธีการเติมค่าสูญหายและค่าตรวจวัดฝนรายวันจริง แสดงผลดังกราฟรูปที่ 7 และ 8 โดยที่ถ้าปริมาณฝนที่



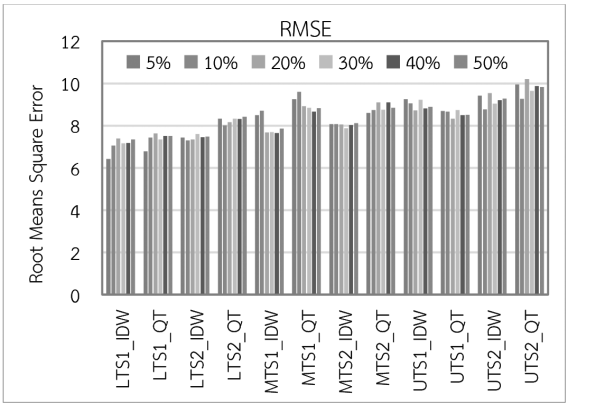
รูปที่ 6 ความแปรปรวนระหว่างค่าตรวจวัดจริง (Obs) และค่าที่ทำการเติมด้วยวิธี IDW และ QT



รูปที่ 8 ค่าฝนเปอร์เซ็นต์ไทล์ที่ 99 ระหว่างค่าตรวจวัดจริง (Obs) และค่าที่ทำการเติมด้วยวิธี IDW และ QT

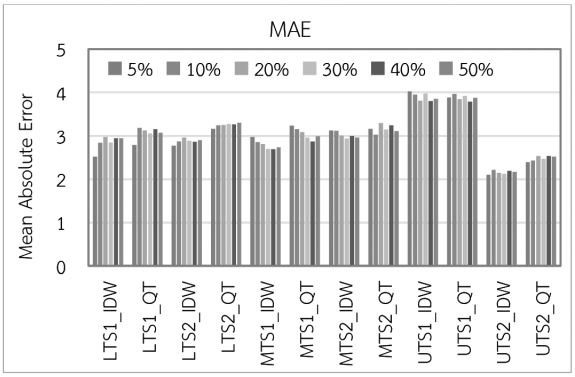


รูปที่ 7 ค่าฝนเปอร์เซ็นต์ไทล์ที่ 95 ระหว่างค่าตรวจวัดจริง (Obs) และค่าที่ทำการเติมด้วยวิธี IDW และ QT

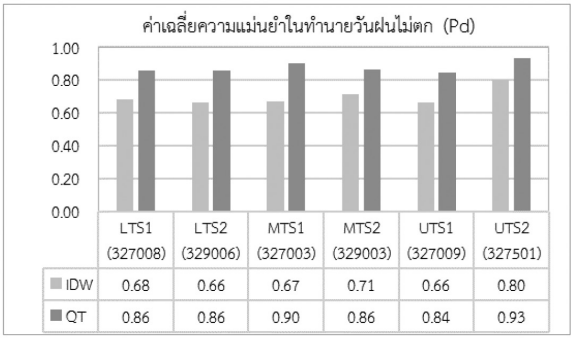


รูปที่ 9 ค่ารากที่สองความคลาดเคลื่อนเฉลี่ยกำลังสอง (RMSE) ที่เปอร์เซ็นต์การสูญหายและสถานีต่างๆ ของวิธี IDW และ QT

เปอร์เซ็นต์ไทล์ 99 มีฝนตกเท่ากับ 38 มม. หมายความว่าโอกาสที่จะมีฝนตกมากกว่า 38 มม. มีเพียง 1% ทำนองเดียวกับเปอร์เซ็นต์ไทล์ที่ 95 ที่ค่าฝนที่จะตกหนักกว่ามีเพียง 5% ดังนั้นค่าเปอร์เซ็นต์ไทล์ที่ 95 และ 99 จึงมีประโยชน์ในการพิจารณาค่ากรณีฝนตกหนัก (เปอร์เซ็นต์ไทล์ที่ 95) หรือฝนสุดขีด (เปอร์เซ็นต์ไทล์ที่ 99) ซึ่งจากกราฟรูปที่ 7 และ 8 พบว่า ค่าเปอร์เซ็นต์ไทล์ที่ 95 และ 99 ของวิธี QT จะให้ค่าใกล้เคียงสอดคล้องกับค่าตรวจวัดจริงมากกว่าวิธี IDW ส่วนวิธี IDW มักให้ค่าฝนของทั้งสองเปอร์เซ็นต์ไทล์ต่ำกว่าค่าฝนจริงซึ่งสอดคล้องกับสมมุติฐานที่ว่า วิธี IDW มักจะให้ค่าฝนตกหนักต่ำกว่าค่าจริง โดยเฉพาะอย่างยิ่งกรณีเหตุการณ์ฝนสุดขีด สรุปได้ว่า วิธี IDW ไม่สามารถประมาณค่าดังกล่าวได้

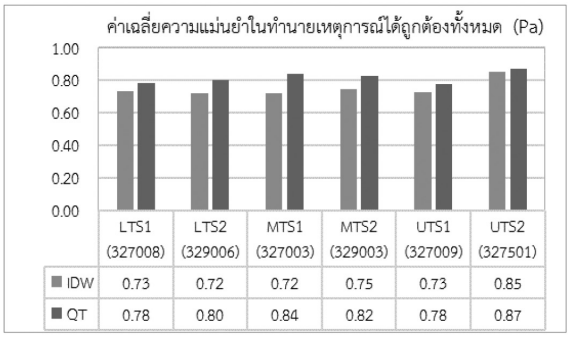


รูปที่ 10 ค่าความคลาดเคลื่อนเฉลี่ยสมบูรณ์ (MAE) ที่เปอร์เซ็นต์การสูญหายและสถานีต่างๆ ของวิธี IDW และ QT

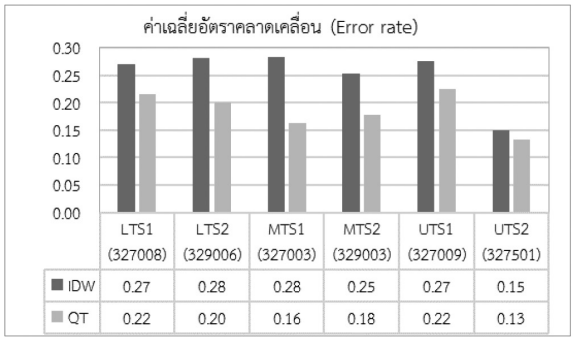


รูปที่ 11 ค่าเฉลี่ยความแม่นยำในการทำนายวันฝนไม่ตก (P_d) ของวิธี IDW และ QT

ส่วนการประเมินค่าความคลาดเคลื่อนด้วย *RMSE* และ *MAE* แสดงในรูปที่ 9 และ 10 พบว่า วิธี IDW มีความคลาดเคลื่อนทั้ง *RMSE* และ *MAE* น้อยกว่าวิธี QT ในทุกกรณี ทั้งนี้ เนื่องจากวิธี QT มักให้ค่าความแปรปรวนมากกว่าวิธี IDW จึงส่งผลให้ค่าความคลาดเคลื่อนมากกว่า อย่างไรก็ตาม ทั้งสองวิธีดังกล่าวให้ค่าความคลาดเคลื่อนไม่แตกต่างกันมากนัก จากกราฟพบว่า ค่า *RMSE* มีค่าเพิ่มขึ้นเล็กน้อยเมื่อมีเปอร์เซ็นต์ค่าสูญหายมากขึ้น ขณะที่ *MAE* ไม่มีทิศทางชัดเจนในทางเพิ่มหรือลดตามเปอร์เซ็นต์การสูญหาย จึงไม่สามารถสรุปได้ว่าที่เปอร์เซ็นต์ค่าสูญหายมาก จะส่งผลให้เกิดค่าความคลาดเคลื่อนมากขึ้นตามไปด้วย



รูปที่ 12 ค่าเฉลี่ยความแม่นยำในการทำนายเหตุการณ์ได้ถูกต้อง (P_a) ระหว่างวิธี IDW และ QT



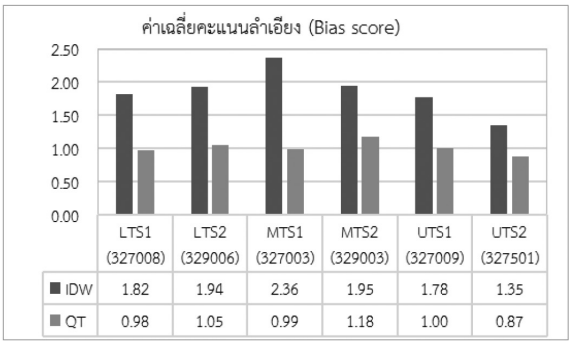
รูปที่ 13 ค่าเฉลี่ยอัตราความคลาดเคลื่อน (Error Rate) ระหว่างวิธี IDW และ QT

3.3 ค่าคะแนนทักษะ

การคำนวณความแม่นยำในการทำนายค่าวันฝนไม่ตก (P_d) ความแม่นยำในการทำนายเหตุการณ์ได้ถูกต้องทั้งหมด (P_a) อัตราคลาดเคลื่อน (Error Rate) และคะแนนความลำเอียง (Bias Score; BS) ของการเติมค่าด้วยวิธี IDW และ QT โดยแสดงค่าเฉลี่ยความแม่นยำของทุกเปอร์เซ็นต์ค่าสูญหายของแต่ละสถานีดังรูปที่ 11–14

จากค่าเฉลี่ยความแม่นยำในการทำนายวันฝนไม่ตก (P_d) พบว่า วิธี QT ให้ความแม่นยำมากกว่าในทุกกรณี

เช่นเดียวกับค่าเฉลี่ยความแม่นยำในการทำนายวันฝนไม่ตก (P_d) ค่าเฉลี่ยความแม่นยำในการทำนายเหตุการณ์ได้ถูกต้องทั้งหมด (P_a) ได้แก่ ทั้งเหตุการณ์วันฝนตก และวันฝนไม่ตก วิธี QT ทำนายได้แม่นยำกว่า



	LTS1 (327008)	LTS2 (329006)	MTS1 (327003)	MTS2 (329003)	UTS1 (327009)	UTS2 (327501)
IDW	1.82	1.94	2.36	1.95	1.78	1.35
QT	0.98	1.05	0.99	1.18	1.00	0.87

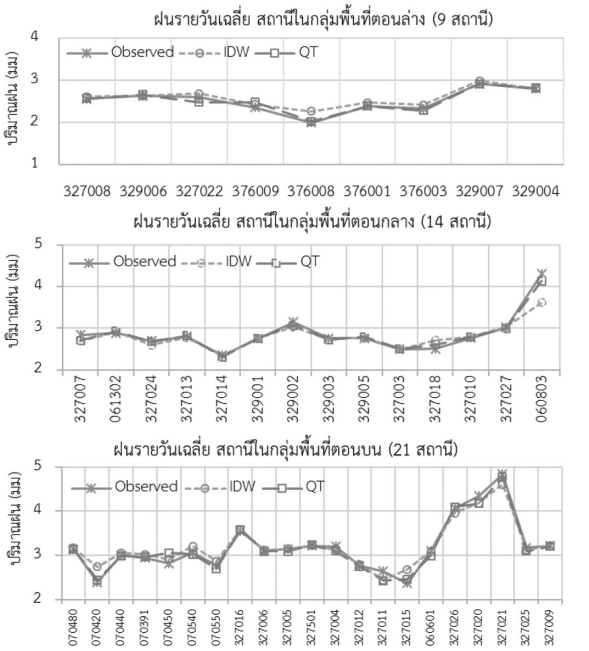
รูปที่ 14 ค่าเฉลี่ยคะแนนลำเอียง (Bias Score) ระหว่างวิธี IDW และ QT

ค่าเฉลี่ยอัตราการคลาดเคลื่อน (Error Rate) ดังรูปที่ 13 ค่ามากกว่าแสดงว่ามีความคลาดเคลื่อนมาก ซึ่งพบว่า วิธี IDW ให้อัตราความคลาดเคลื่อนมากกว่าวิธี QT ในทุกกรณีเช่นกัน

รูปที่ 14 แสดงค่าเฉลี่ยคะแนนลำเอียง (Bias Score; BS) ถ้าค่า *BS*=1 แสดงว่าทำนายเหตุการณ์ได้แม่นยำทุกเหตุการณ์ *BS* < 1 แสดงว่าทำนายเหตุการณ์วันฝนตกน้อยกว่าความเป็นจริง *BS* > 1 แสดงว่าทำนายเหตุการณ์วันฝนตกได้มากกว่าความเป็นจริง จากกราฟรูปที่ 14 จะพบว่า ค่าเฉลี่ยคะแนนลำเอียงจากทุกกรณี (ทุกเปอร์เซ็นต์ความคลาดเคลื่อน) ของวิธี QT จะให้ค่าใกล้เคียง 1 มากกว่าวิธี IDW ดังนั้นจึงสรุปได้ว่าวิธี QT สามารถประมาณค่าเหตุการณ์และเติมค่าสูญหายฝนรายวันได้แม่นยำกว่าวิธี IDW

3.4 ผลการประยุกต์การเติมค่าสูญหายให้กับสถานีฝนในลุ่มน้ำปิงตอนบน

การประยุกต์วิธีการเติมค่าสูญหายให้กับสถานีฝนในลุ่มน้ำปิงตอนบนเพื่อให้ได้ชุดข้อมูลฝนรายวันที่สมบูรณ์นั้น ในการศึกษานี้ทำการเติมเฉพาะสถานีที่มีค่าสูญหายไม่เกิน 50% พบว่า กลุ่มสถานีในพื้นที่ตอนล่างมีจำนวน 9 สถานี พื้นที่ตอนกลางมี 14 สถานี และพื้นที่ตอนบนมีจำนวน 21 สถานี รวมทั้งสิ้น 44 สถานี ผลการเติมค่าสูญหายจริงให้กับสถานีดังกล่าว เมื่อนำชุดข้อมูลฝนในช่วงเวลาที่ศึกษามาหาค่าฝนรายวันเฉลี่ย (เฉลี่ยจากข้อมูลทั้งหมด) แสดงผลดังรูปที่ 15 และเมื่อหาค่าฝนที่เปอร์เซ็นต์ไทล์ 99 แสดงได้ดังรูปที่ 16

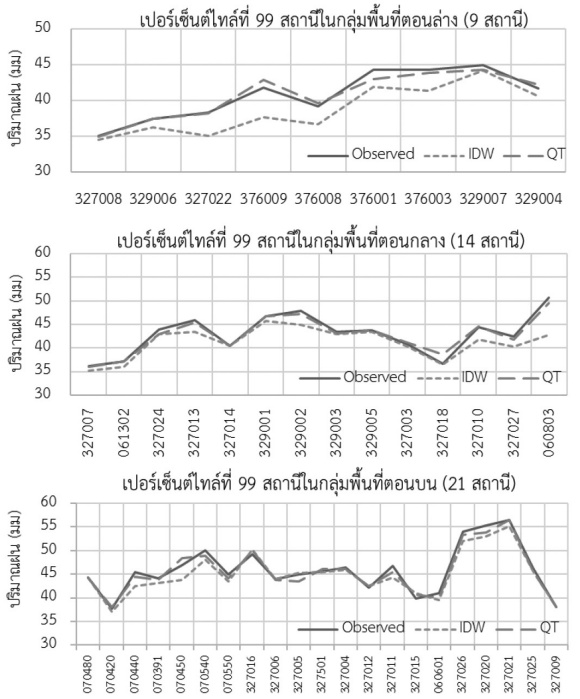


รูปที่ 15 ฝนรายวันเฉลี่ย ของสถานีตรวจวัดฝนในลุ่มน้ำปิงตอนบนที่ทำการเติมค่าสูญหาย

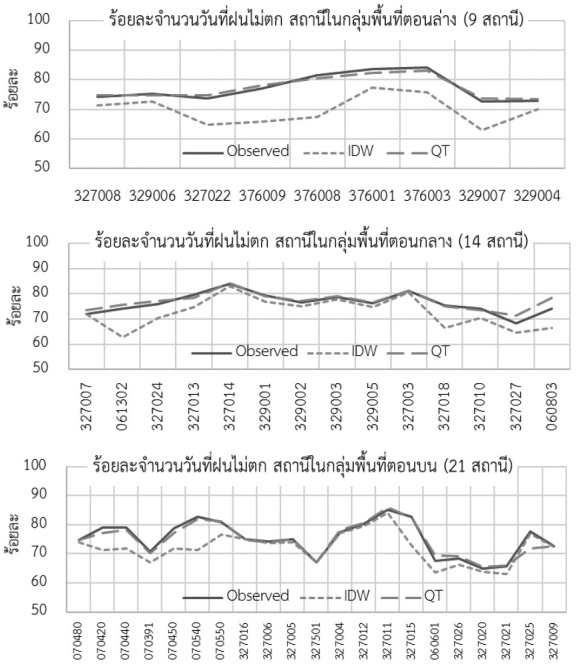
และร้อยละของจำนวนวันที่ฝนไม่ตก แสดงดังรูปที่ 17 โดยที่ Obs หมายถึง ข้อมูลตรวจวัดที่มีค่าสูญหายจริง ส่วน IDW และ QT หมายถึง ชุดข้อมูลที่ทำการเติมค่าสูญหายแล้วด้วยวิธี IDW และ QT ตามลำดับ

จากรูปที่ 15 พบว่า ฝนรายวันเฉลี่ยก่อนทำการเติมค่าและหลังทำการเติมค่าด้วยวิธี IDW และ QT มีความแตกต่างกันไม่มากนัก (เนื่องจากเป็นการเฉลี่ยค่า) ทั้งนี้วิธี QT จะให้ค่าฝนรายวันเฉลี่ยใกล้เคียงกับค่าตรวจวัดจริงมากกว่าวิธี IDW

จากรูปที่ 16 ค่าฝนที่เปอร์เซ็นต์ไทล์ที่ 99 ซึ่งสามารถใช้แสดงกรณีเหตุการณ์ฝนสุดขีด พบว่า วิธี QT ให้ค่าสอดคล้องกับค่าตรวจวัดจริงมากกว่าวิธี IDW นอกจากนี้จะสังเกตเห็นว่าสถานีในกลุ่มพื้นที่ตอนล่างวิธี IDW มีแนวโน้มของค่าเปอร์เซ็นต์ไทล์ที่ 99 ต่ำกว่าค่าตรวจวัดจริงอย่างชัดเจน ทั้งนี้เนื่องจากสถานีในกลุ่มพื้นที่ตอนล่างเป็นสถานีที่มีจำนวนข้อมูลสูญหาย (เปอร์เซ็นต์การสูญหาย) มากกว่าสถานีในกลุ่มพื้นที่ตอนกลาง และตอนบน (ตามลำดับ)



รูปที่ 16 ค่าฝนที่เปอร์เซ็นต์ไทล์ 99 ของสถานีตรวจวัดฝนในลุ่มน้ำปิงตอนบนที่ทำการเติมค่าสูญหาย



รูปที่ 17 ร้อยละของจำนวนวันฝนไม่ตก ของสถานีตรวจวัดฝนในลุ่มน้ำปิงตอนบนที่ทำการเติมค่าสูญหาย

และจากกราฟรูปที่ 17 แสดงถึงร้อยละของจำนวนวันฝนไม่ตกจากค่าข้อมูลทั้งหมด ผลการเติมค่าด้วยวิธี IDW และ QT พบว่า วิธี IDW มักให้จำนวนร้อยละของวันฝนไม่ตกต่ำกว่าค่าตรวจวัดจริง ซึ่งสามารถยืนยันผลที่สอดคล้องกับข้อจำกัดของวิธีการเติมค่าสูญหายแบบดั้งเดิม (เช่น วิธีค่าเฉลี่ย และวิธี IDW) ที่มักให้จำนวนวันฝนตกที่มากเกินจริง นอกจากนี้จากกราฟพบว่า สถานีในกลุ่มพื้นที่ตอนล่าง การเติมค่าสูญหายด้วยวิธี IDW ให้ผลแตกต่างจากค่าตรวจวัดจริงอย่างชัดเจน ทั้งนี้ เนื่องจากสถานีในกลุ่มพื้นที่ตอนล่างมีจำนวนค่าสูญหายมากกว่าสถานีในกลุ่มพื้นที่ตอนกลางและตอนบน ดังนั้นจึงสรุปได้ว่ายิ่งข้อมูลมีค่าสูญหายมากการเติมค่าด้วยวิธี IDW จะให้ผลที่แตกต่างจากค่าตรวจวัดจริงเพิ่มขึ้น

4. อภิปรายผลและสรุป

ข้อมูลฝนเป็นข้อมูลพื้นฐานที่สำคัญในการศึกษาใดๆ ที่เกี่ยวข้องกับทรัพยากรน้ำ โดยเฉพาะอย่างยิ่งข้อมูลฝนรายวันซึ่งมีลักษณะของข้อมูลแบบต่อเนื่อง และแบบไม่ต่อเนื่อง และการแจกแจงความถี่ข้อมูลฝนรายวันไม่ใช่การแจกแจงปกติ โดยทั่วไปวิธีการเติมค่าสูญหายข้อมูลฝนมักใช้วิธีอย่างง่าย ได้แก่ วิธีค่าเฉลี่ย วิธีอัตราส่วนปรกติ และวิธีระยะทางผกผัน (IDW) แต่เนื่องจากวิธีการเหล่านี้มักมีข้อจำกัด 1) ให้ค่าปริมาณฝนรายวันต่ำกว่าความเป็นจริง 2) จำนวนวันฝนกตกที่มากเกินไป และ 3) ไม่สามารถประมาณค่ากรณีเหตุการณ์ฝนสุดขีดได้

งานวิจัยศึกษานี้จึงพัฒนาวิธีการเติมค่าข้อมูลฝนรายวันด้วยวิธีควอนไทล์ (QT) โดยใช้การแจกแจงความถี่แบบแบร์นูลี-แกมมา และเปรียบเทียบกับวิธี IDW ผลการทดสอบการเติมค่าและตรวจสอบด้วยค่าทางสถิติพื้นฐานต่างๆ ได้แก่ ค่าฝนรายวันสูงสุด ฝนรายวันเฉลี่ย และความแปรปรวน การเติมค่าสูญหายด้วยวิธี QT ให้ค่าทางสถิติพื้นฐานโดยรวมและส่วนใหญ่ดีกว่าวิธี IDW ที่ทุกเปอร์เซ็นต์การสูญหาย

ปริมาณฝนที่ค่าเปอร์เซ็นต์ไทล์ 95 และ 99 วิธี QT ให้ค่า

เปอร์เซ็นต์ไทล์ดังกล่าวใกล้เคียงกับค่าตรวจวัด บ่งชี้ได้ชัดเจนว่าวิธี QT สามารถประมาณค่ากรณีเหตุการณ์ฝนสุดขีดได้ดีกว่าวิธี IDW

ค่าคะแนนทักษะ ทั้งกรณีการทำนายเหตุการณ์วันฝนไม่ตก (P_d) และความแม่นยำในการทำนายเหตุการณ์ได้ถูกต้องทั้งหมด $C_{0,0}$ และ $C_{1,1}$ (P_a) วิธี QT ให้ผลการทำนายเหตุการณ์แม่นยำกว่า โดยเฉพาะอย่างยิ่งความแม่นยำในการทำนายวันฝนไม่ตก (P_d) ดังนั้นวิธี QT สามารถแก้ข้อจำกัดการประมาณค่าของวิธีดั้งเดิม หรือวิธี IDW ในประเด็นการให้จำนวนวันฝนตกที่มากเกินไปได้ ส่งผลให้วิธี QT ให้ค่าอัตราความคลาดเคลื่อนและคะแนนค่าลำเอียงดีกว่าวิธี IDW

การประเมินประสิทธิภาพด้วยค่ารากที่สองความคลาดเคลื่อนเฉลี่ยกำลังสอง (RMSE) และค่าความคลาดเคลื่อนเฉลี่ยสมบูรณ์ (MAE) วิธี IDW ให้ค่าความคลาดเคลื่อนในการประมาณค่าฝนรายวันน้อยกว่าวิธี QT ทั้งนี้เนื่องจากวิธี QT จะมีค่าความแปรปรวนของค่าที่ประมาณได้มากกว่า (แต่สอดคล้องกับค่าตรวจวัดจริงมากกว่า) จึงคำนวณค่า *RMSE* และ *MAE* ได้ค่ามากกว่า อย่างไรก็ตาม ความคลาดเคลื่อนดังกล่าวของทั้งสองวิธีมีค่าแตกต่างกันไม่มากนัก

การทดสอบการจำลองค่าสูญหายแบบสุ่ม (MAR) ที่มีเปอร์เซ็นต์ค่าสูญหายแตกต่างกัน (5%, 10%, 20%, 30%, 40% และ 50%) ของสถานีทดสอบ 6 สถานี พบว่า การเติมค่าสูญหายที่เปอร์เซ็นต์แตกต่างกันไม่มีผลชัดเจนต่อค่าทางสถิติต่างๆ รวมถึงค่าคะแนนทักษะ ยกเว้นค่าความคลาดเคลื่อน RMSE มีแนวโน้มเพิ่มขึ้นเมื่อมีค่าเปอร์เซ็นต์การสูญหายสูงขึ้น

นอกจากนี้การประยุกต์วิธีการเติมค่าสูญหายให้กับสถานีตรวจวัดฝนที่มีค่าสูญหายจริงไม่เกิน 50% จำนวน 44 สถานี ในพื้นที่ศึกษา ผลการเติมค่าเพื่อให้ได้ชุดข้อมูลที่สมบูรณ์สามารถยืนยันได้ชัดเจนว่าวิธี QT ให้ค่าฝนรายวันเฉลี่ย ค่าฝนที่เปอร์เซ็นต์ไทล์ 99 และจำนวนร้อยละวันฝนไม่ตกใกล้เคียงกับค่าตรวจวัดมากกว่าวิธี IDW ทั้งนี้ ยังสังเกตพบว่าสถานีที่มีจำนวนค่าสูญหายมาก ถ้าใช้การเติมค่าด้วยวิธี IDW จะยิ่งให้ผลที่แตกต่างจากค่าตรวจวัดมากขึ้นด้วย

ดังนั้นผลการศึกษานี้จึงสรุปได้ว่าวิธี QT สามารถใช้ประมาณค่าและเติมค่าสูญหายข้อมูลฝนรายวันได้ดีกว่าวิธี IDW โดยสามารถลดข้อจำกัดต่างๆ ของวิธีดั้งเดิมได้ดังกล่าวมาแล้ว อย่างไรก็ตามวิธี QT มีความยุ่งยากในการคำนวณมากกว่าวิธี IDW ดังนั้นในการเลือกใช้การเติมค่าสูญหายข้อมูลฝนรายวันสามารถเลือกใช้โดยขึ้นกับวัตถุประสงค์ในการศึกษา เช่น กรณีใช้เฉพาะฝนรายเดือนหรือไม่ได้พิจารณาฝนสูงสุดสามารถใช้วิธี IDW ประมาณค่าได้ แต่ในกรณีที่ต้องพิจารณาเหตุการณ์ฝนตกหนักร่วมด้วยควรใช้วิธี QT ซึ่งจะให้ค่าฝนตกหนักได้ดีกว่า

งานศึกษาวิจัยนี้ยังสามารถพัฒนาเพิ่มเติมได้อีก โดยมีประเด็นที่ต้องพิจารณาเพิ่มเติม เช่น ประยุกต์ใช้กับกรณีมีข้อมูลสั้น ได้แก่ เช่น 10 ปี 20 ปี หรือ 30 ปี หรือน้อยกว่าข้อมูลที่ใช้ในการศึกษานี้ (65 ปี) ซึ่งการเติมค่าด้วยวิธี QT ยังจะให้ผลดีหรือไม่ นอกจากนี้ยังต้องมีการพิจารณากรณีมีค่าสูญหายแบบต่อเนื่อง เช่น ค่าข้อมูลหายติดต่อกันเป็นช่วงเวลาหลายวัน ซึ่งผู้วิจัยจะได้ทำการพัฒนาและศึกษาเพิ่มเติมต่อไป

5. กิตติกรรมประกาศ

ขอขอบคุณนายณัฐพงศ์ บุญประเสริฐ นางสาวสุธินี วงศ์แสง นายคมกฤษ โสภา และนางสาวกุลสตรี ศรีจุมปา ที่มีส่วนช่วยในการจัดเตรียมข้อมูลและทดลอง ส่งผลให้งานวิจัยนี้สำเร็จลุล่วงได้เป็นอย่างดี

เอกสารอ้างอิง

[1] R. P. De Silva, N. D. K. Dayawansa, and M. D. Ratnasiri, "A comparison of methods used in estimating missing rainfall data," *The Journal of Agricultural Sciences,* vol. 3, no. 2, pp. 101–108, 2007.

[2] R. S. V. Teegavarapu and V. Chandramouli, "Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records," *Journal of Hydrology,* vol. 312, no. 1–4, pp. 191–206, 2005.

[3] B. I. Lozada Garcia, G. Sparovek, P. C. Sentelhas,

and L. Tapia, "Filling in missing rainfall data in the Andes region of Venezuela, based on a cluster analysis approach," *Revista Brasileira de Agrometeorologia,* vol. 14, no. 2, pp. 225–233, 2006.

[4] L. R. Presti, E. Barca, and G. Passarella, "A methodology for treating missing data applied to daily rainfall data in the Candelaro river basin (Italy)," *Environmental Monitoring and Assessment,* vol. 160, no. 1, pp. 1, 2010.

[5] M. M. Hasan and B. F. W. Crokea, "Filling gaps in daily rainfall data: A statistical approach," presented at 20th International Congress on Modelling and Simulation, Adelaide, Australia, December 1–6, 2013.

[6] J. Kim and J. H. Ryu, "A heuristic gap filling method for daily precipitation series," *Water Resources Management,* vol. 30, no. 7, pp. 2275–2294, 2016.

[7] C. Simolo, M. Brunetti, M. Maugeri, and T. Nanni, "Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach," *International Journal of Climatology,* vol. 30, no. 10, pp. 1564–1576, 2010.

[8] H. Aksoy, "Use of gamma distribution in hydrological analysis," *Turkish Journal of Engineering and Environmental Sciences*, vol. 24, no. 6, pp. 419–428, 2000.

[9] R. S. V. Teegavarapu, "Missing precipitation data estimation using optimal proximity metric-based imputation, nearest-neighbour classification and cluster-based interpolation methods," *Hydrological Sciences Journal,* vol. 59, no. 11, pp. 2009–2026, 2014.

[10] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data,* John Wiley & Sons Inc., 1987.

[11] S. Srisuttiyakorn, "Missing data analysis," *Journal of Education Studies,* vol. 42, no. 1, pp. 217–223, 2014.

[12] S. Wuthiwongyothin, "Evaluating inverse distance weighting and correlation coefficient weighting methods on daily rainfall time series," *SNRU Journal of Science and Technology,* vol. 13, no. 2, pp. 71–79, 2021.

[13] D. A. Mooley, "Gamma distribution probability model for Asian summer monsoon monthly rainfall," *Monthly Weather Review,* vol. 101, no. 2, pp. 160–176, 1973.