การเปรียบเทียบวิธีการเติมข้อมูลสูญหายในตัวแปรตามที่เกิดการสูญหายแบบสุ่ม สำหรับการถดถอยเชิงเส้นพหุคูณ

Comparison of Missing Data Imputation Methods in Dependent Variable with Missing at Random for Multiple Linear Regression

สุปรียา สระโสม[1] และ ธิดาเดียว มยุรีสวรรค์[1*]

[1]สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น อ.เมือง จ.ขอนแก่น

[*]Corresponding Author, E-mail: mtidadeaw@gmail.com

Received: 5 April 2019 | Revised17 July 2019 | Accepted: 29 October 2019

บทคัดย่อ

งานวิจัยนี้ได้พัฒนาวิธีการเติมข้อมูลสูญหายในตัวแปรตามสำหรับการถดถอยเชิงเส้นพหุคูณ เมื่อตัวแปรตามมีการสูญหายแบบสุ่ม (Missing at random) โดยวิธีที่พัฒนาได้แก่วิธี Mean Regression Imputation (MRI) วิธี Expectation Maximization with Multiple Imputation (EMMI) และวิธี Nearest Average Regression Imputation (NARI) โดยเปรียบเทียบประสิทธิภาพวิธีที่พัฒนาขึ้นกับอีก 6 วิธี ได้แก่วิธี Regression Imputation (RI) วิธี Stochastic Regression Imputation (SRI) วิธี K Nearest Neighbor Imputation (KNN) วิธี Expectation Maximization Algorithm (EM) วิธี Multiple Imputation (MI) และวิธี Proportioned Residual Draw Imputation (PRD) การศึกษาจากการจำลองข้อมูลด้วยโปรแกรม R กำหนดส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อน ($\pm$) เท่ากับ 5, 10 และ 15 และขนาดตัวอย่าง (n) เท่ากับ 30, 50, 100 และ 200 และร้อยละการสูญหายเท่ากับ 5, 10, 15 และ 20 เกณฑ์เปรียบเทียบประสิทธิภาพคือค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Average Mean Square Error: AMSE) ผลการวิจัยพบว่า วิธี EMMI มีประสิทธิภาพดีที่สุดสำหรับทุกระดับขนาดตัวอย่างที่ $\pm$ มีค่าเท่ากับ 5 และร้อยละการสูญหายเท่ากับ 5 วิธี MRI มีประสิทธิภาพดีกว่าวิธีอื่นที่ทุกระดับขนาดตัวอย่างเมื่อ $\pm$ มีค่าเท่ากับ 10 และร้อยละการสูญหายเท่ากับ 5 และวิธี MRI ยังคงมีประสิทธิภาพดีที่สุดเมื่อ $\pm$ มีค่าเท่ากับ 15 ในทุกระดับร้อยละการสูญหายและเกือบทุกระดับขนาดตัวอย่าง ส่วนผลการศึกษาจากข้อมูลจริงที่ $\boldsymbol{n}$ เท่ากับ 50 พบว่าวิธี MRI มีประสิทธิภาพดีที่สุดในทุกระดับร้อยละการสูญหาย

ABSTRACT

This research is to develop missing data imputation methods in dependent variable for multiple linear regression with missing at random in dependent variable, namely the Mean Regression Imputation method (MRI), the Expectation Maximization with Multiple Imputation method (EMMI) and the Nearest Average Regression Imputation method (NARI). Comparison of the efficiency of the develop methods with 6 methods, namely the Regression Imputation method (RI), the Stochastic Regression Imputation method (SRI), the K Nearest Neighbour Imputation method (KNN), the Expectation Maximization Algorithm method (EM), the Multiple Imputation method (MI) and the Proportioned Residual Draw Imputation method (PRD). The simulation study with R program where the

standard deviations of error ($\pm$) were set to be 5, 10 and 15, and sample sizes (n) were 30, 50, 100 and 200, and missing percentages were 5, 10, 15 and 20. The criteria for compare the performance is an Average Mean Square Error (AMSE). The results found that, the EMMI method has the best performance for all level of sample sizes at $\pm$ is equal to 5 and missing percentage is equal to 5. The MRI method performs better than the others at all level of sample sizes when $\pm$ is equal to 10 and missing percentage is equal to 5, and the MRI method still performs the best when $\pm$ is equal to 15 in all missing percentages and almost of all sample size levels. The result for real data at n = 50, the MRI method has the most effective in all level of missing percentages.

คำสำคัญ: การเติมข้อมูลสูญหาย การถดถอยเชิงเส้นพหุคูณ ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย
Keywords: Missing data imputation, Multiple linear regression, Average mean square error

บทนำ

การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ (Multiple linear regression analysis) เป็นการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรตาม (Dependent variable: y) กับตัวแปรอิสระ (Independent variable: x) มากกว่า 1 ตัวแปร โดยนำความสัมพันธ์ที่เป็นเชิงเส้นในค่าพารามิเตอร์มาสร้างตัวแบบทางคณิตศาสตร์ ที่เรียกว่าตัวแบบการถดถอยเชิงเส้นพหุคูณ (Multiple linear regression model) (Dielman, 2004) ปัญหาหนึ่งที่อาจเกิดขึ้นในการวิเคราะห์การถดถอยคือการสูญหายของข้อมูล (Missing data) ในตัวแปรตาม ตัวอย่างเช่น การเก็บข้อมูลจากการสำรวจ อาจเกิดปัญหาข้อมูลสูญหายเนื่องจากหน่วยตัวอย่างบางหน่วยไม่ให้คำตอบในบางประเด็นคำถาม หน่วยตัวอย่างนี้ทราบข้อมูลที่เป็นค่าของตัวแปรอิสระที่ทำการศึกษา แต่ไม่ทราบข้อมูลที่เป็นค่าของตัวแปรตาม ซึ่งคือคำตอบของประเด็นคำถามนั้น หรือตัวอย่างข้อมูลจากการทดลองที่เกิดการสูญหายเนื่องจากหน่วยทดลองเป็นสิ่งมีชีวิตแล้วเกิดการตายในระหว่างการทดลอง หน่วยตัวอย่างนี้ทราบข้อมูลที่เป็นค่าของตัวแปรอิสระ แต่ไม่ทราบข้อมูลที่เป็นค่าของตัวแปรตามเนื่องจากเกิดการตาย ในกรณีที่ไม่สามารถเก็บข้อมูลเพิ่มได้อันเนื่องจากมีข้อจำกัดในเรื่องของเวลา หรือมีข้อจำกัดจากปัจจัยอื่นๆ หากข้อมูลสมบูรณ์ที่มีอยู่มีจำนวนมากพอที่จะนำมาวิเคราะห์ ก็สามารถตัดข้อมูลที่สูญหายทิ้งไปได้ (Ignoring and discarding data) (นรุตม์, 2553) แต่บางครั้ง ข้อมูลสมบูรณ์ที่เหลืออยู่มีจำนวนไม่มาก จึงไม่เพียงพอต่อการวิเคราะห์ หากตัดข้อมูลที่ไม่สมบูรณ์ออกจะส่งผลต่อประสิทธิภาพในการวิเคราะห์ข้อมูล กรณีนี้จึงมีความจำเป็นต้องเติมค่าสูญหายให้กับตัวแปรนั้น ซึ่งมีนักวิจัยหลายคนพยายามหาวิธีจัดการกับข้อมูลสูญหายนี้

การสูญหายแบบสุ่ม (Missing at random หรือ MAR) เป็นลักษณะการสูญหายของข้อมูลที่ค่าของข้อมูลสูญหายนั้นไม่มีความสัมพันธ์กับค่าข้อมูลสูญหายอื่น แต่มีความสัมพันธ์กับข้อมูลอื่นบางข้อมูล (Little and Rubin, 2002) เมื่อเกิดข้อมูลสูญหายในค่าตัวแปรตามสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ มีนักวิจัยหลายคนได้ศึกษาวิธีการเติมข้อมูลสูญหายในลักษณะนี้

จริยา (2551) ได้เปรียบเทียบวิธีการเติมข้อมูลสูญหายในตัวแปรตาม สำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ โดยศึกษา 4 วิธีได้แก่ วิธี Loss Imputation (Loss) วิธี Mean Imputation (Mean) วิธี Regression Imputation (RI) และวิธี Multiple Imputation (MI) โดยกำหนดขนาดตัวอย่างเท่ากับ 50, 70, 100 และ 200 ส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนเท่ากับ 1, 5 และ 15 ร้อยละการสูญหายเท่ากับ 5, 10, 20 และ 30 เกณฑ์วัดประสิทธิภาพคือรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error: RMSE) ผลการวิจัยพบว่า ทุกระดับค่าส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อน กรณีขนาดตัวอย่างเท่ากับ 50 และร้อยละการสูญหายเท่ากับ 5 วิธี MI มีประสิทธิภาพดีที่สุด ส่วนร้อยละการสูญหายเท่ากับ 10, 20 และ 30 พบว่าวิธี RI มีประสิทธิภาพดีที่สุด กรณีขนาดตัวอย่างเท่ากับ 70 และ 200 ที่ทุกร้อยละการสูญหาย พบว่าวิธี RI มีประสิทธิภาพดีที่สุด กรณีขนาดตัวอย่างเท่ากับ 100 ร้อยละการสูญหายเท่ากับ 5, 10 และ 20 พบว่าวิธี RI มีประสิทธิภาพดีที่สุด ส่วนร้อยละการสูญหายเท่ากับ 30 พบว่าวิธี MI มีประสิทธิภาพดีที่สุด อุษณีย์ (2555) ได้เปรียบเทียบวิธีการเติมข้อมูลสูญหายในตัวแปรตามแบบนอนอิกนอร์เรเบิล (Nonignorable) สำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ โดยศึกษา 3 วิธีได้แก่ วิธี Expectation Maximization Algorithm

(EM) วิธี K Nearest Neighbor Imputation (KNN) และวิธี Predictive Mean Matching Imputation (PMM) โดยกำหนดขนาดตัวอย่างเท่ากับ 50, 100 และ 200 ส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนเท่ากับ 10, 30 และ 90 ร้อยละการสูญหายเท่ากับ 10, 20 และ 30 เกณฑ์วัดประสิทธิภาพคือค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองค่าเฉลี่ย (Average Mean Square Error: AMSE) ผลการวิจัยพบว่า กรณีส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนเท่ากับ 10 ที่ทุกระดับขนาดตัวอย่างและร้อยละการสูญหาย วิธี EM มีประสิทธิภาพดีที่สุด กรณีส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนเท่ากับ 30 ร้อยละการสูญหายเท่ากับ 10 ขนาดตัวอย่างเท่ากับ 50 พบว่าวิธี KNN มีประสิทธิภาพดีที่สุด ส่วนขนาดตัวอย่างเท่ากับ 100 และ 200 พบว่าวิธี EM มีประสิทธิภาพดีที่สุด แต่หากร้อยละการสูญหายเท่ากับ 30 และ 40 ที่ทุกระดับขนาดตัวอย่าง พบว่าวิธี EM มีประสิทธิภาพดีที่สุด กรณีส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนเท่ากับ 90 พบว่าวิธี KNN มีประสิทธิภาพดีที่สุดเกือบทุกสถานการณ์ เรืองลักษณ์ (2560) ได้เปรียบเทียบวิธีการเติมข้อมูลสูญหายในตัวแปรตาม 4 วิธีสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ ซึ่งศึกษาทั้งวิธีแบบเดี่ยวและแบบร่วม วิธีแบบเดี่ยวได้แก่ วิธี RI วิธี Stochastic Regression Imputation (SRI) วิธี KNN และวิธี EM ส่วนวิธีแบบร่วมได้แก่ วิธี K Nearest Regression Imputation with Equivalent Weighted (KREW) และวิธี K Nearest Stochastic Regression Imputation with Equivalent Weighted (KSEW) โดยวิธี KREW เป็นวิธีที่พัฒนามาจากการรวมกันของวิธี KNN และวิธี RI ส่วนวิธี KSEW เป็นวิธีที่พัฒนามาจากการรวมกันของวิธี KNN และวิธี SRI ทำการศึกษาที่ขนาดตัวอย่างเท่ากับ 20, 30, 50 และ 100 ส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนเท่ากับ 5, 10 และ 15 ร้อยละการสูญหายเท่ากับ 10, 20, 30 และ 40 เกณฑ์วัดประสิทธิภาพคือค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE) ผลการวิจัยพบว่า ทุกระดับของส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อน ที่ขนาดตัวอย่างเท่ากับ 20 และ 30 ทุกระดับร้อยละการสูญหาย วิธี KSEW มีประสิทธิภาพดีที่สุดเกือบทุกสถานการณ์ ส่วนที่ขนาดตัวอย่างเท่ากับ 50 และ 100 พบว่าวิธี SRI มีประสิทธิภาพดีที่สุดเกือบทุกสถานการณ์ ต่อมา Kwon and Park (2015) ได้พัฒนาวิธีการเติมข้อมูลสูญหายในตัวแปรตามสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ ชื่อว่าวิธี Proportioned Residual Draw Imputation (PRD) โดยเป็นวิธีที่คล้ายกับวิธี MI แต่ได้นำค่าสัดส่วนความคลาดเคลื่อนมาใช้สำหรับสร้างค่าแทนข้อมูลสูญหาย

จากงานวิจัยที่กล่าวมาข้างต้นซึ่งเป็นการศึกษาและพัฒนาการเติมข้อมูลสูญหายในตัวแปรตาม สำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ ซึ่งผู้วิจัยเห็นว่ายังมีแนวคิดที่จะพัฒนาวิธีการเติมข้อมูลสูญหายด้วยวิธีอื่นได้อีกที่มีลักษณะใกล้เคียงกัน ดังนั้นงานวิจัยนี้จึงนำเสนอวิธีการเติมข้อมูลสูญหายในตัวแปรตาม สำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ เมื่อตัวแปรตามมีการสูญหายแบบสุ่ม (Missing at random หรือ MAR) โดยวิธีที่นำเสนอมี 3 วิธีคือ วิธี Mean Regression Imputation (MRI) วิธี Expectation Maximization with Multiple Imputation (EMMI) และ วิธี Nearest Average Regression Imputation (NARI) และพิจารณาประสิทธิภาพด้วยค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Average Mean Square Error: AMSE) โดยนำไปเปรียบเทียบกับวิธีการเติมข้อมูลสูญหายวิธีอื่นอีก 6 วิธีดังรายละเอียดในวิธีดำเนินการวิจัย

วิธีดำเนินการวิจัย

การศึกษาจากการจำลองข้อมูล

งานวิจัยนี้ใช้โปรแกรม R ในการจำลองข้อมูล และกำหนดพารามิเตอร์และขอบเขตงานวิจัยให้เหมือนงานวิจัยอื่นที่มีลักษณะเดียวกัน เพื่อประโยชน์ในการเปรียบเทียบผลการวิจัย โดยมีวิธีการวิจัยดังนี้

1. กำหนดขนาดตัวอย่าง (*n*) ที่ศึกษา 4 ระดับคือ 30, 50, 100 และ 200 ทำการจำลองข้อมูลของตัวแปรอิสระ 3 ตัวแปรคือ x_1, x_2, x_3 โดยแต่ละตัวแปรมีจำนวน *n* ค่าและมีการแจกแจงปรกติด้วยค่าเฉลี่ยเท่ากับ 0 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 10

2. จำลองข้อมูลค่าความคลาดเคลื่อน $£_i$ จำนวน *n* ค่าให้มีการแจกแจงปรกติโดยมีค่าเฉลี่ยเท่ากับ 0 ส่วนเบี่ยงเบนมาตรฐาน ($\pm$) เท่ากับ 5, 10 และ 15

3. กำหนดค่าสัมประสิทธิ์การถดถอย $_0 = _1 = _2 = _3 = 1$ และสร้างข้อมูลของตัวแปรตาม y_i กับตัวแปรอิสระทั้ง 3 ตัวแปรภายใต้ตัวแบบการถดถอยเชิงเส้นพหุคูณ (Multiple linear regression) ด้วยรูปแบบดังนี้

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \;\; ; i = 1, 2, ..., n$$

4. สร้างการสูญหายแบบสุ่มให้กับค่าของตัวแปรตาม โดยกำหนดร้อยละการสูญหาย 4 ระดับคือ 5, 10, 15 และ 20 เมื่อเทียบกับขนาดตัวอย่างที่ศึกษา

5. ประมาณค่าข้อมูลของตัวแปรตามเพื่อแทนที่ข้อมูลที่สูญหายด้วยวิธีการเติมข้อมูลสูญหายทั้ง 9 วิธีที่ทำการศึกษา

6. นำข้อมูลในข้อ 5 ของแต่ละวิธีการเติมข้อมูลสูญหายไปสร้างตัวแบบการถดถอยเชิงเส้นพหุคูณด้วยวิธีกำลังสองน้อยที่สุดแบบสามัญ (Ordinary Least Squares method: OLS) และหาค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE) ของแต่ละวิธีจากสูตร (อุษณีย์, 2555)

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

เมื่อ y_i' คือค่าของตัวแปรตามตัวอย่างที่ i

$\hat{y}_i$ คือค่าทำนายของตัวแปรตามตัวอย่างที่ i

7. ทำซ้ำ 1,000 รอบเพื่อหาค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Average Mean Square Error: AMSE) ของแต่ละวิธีจากสูตร

$$AMSE = \frac{1}{1,000}\sum_{b=1}^{1,000} MSE_b \qquad (2)$$

8. ทำซ้ำทุกสถานการณ์ที่กำหนดในงานวิจัย และเปรียบเทียบประสิทธิภาพของวิธีการเติมข้อมูลสูญหายทั้ง 9 วิธีโดยพิจารณาจากค่า AMSE

การศึกษาจากข้อมูลจริง

สำหรับข้อมูลจริงที่นำมาใช้ในงานวิจัยคือข้อมูลคุณภาพอากาศของพื้นที่ที่มีมลพิษในประเทศอิตาลี จากฐานข้อมูลออนไลน์ UCI Machine Learning Repository โดยสุ่มตัวอย่างมา 50 ตัวอย่าง ตัวแปรตามคือปริมาณก๊าซคาร์บอนมอนอกไซด์ (CO: y) ตัวแปรอิสระ 3 ตัวแปรได้แก่ ตัวแปรอุณหภูมิ (Temperature: x_1) ตัวแปรความชื้นสัมพัทธ์ (Relative humidity: x_2) และตัวแปรความชื้นสัมบูรณ์ (Absolute humidity: x_3) โดยผู้วิจัยได้ทำการตรวจสอบความเหมาะสมของตัวแบบการถดถอยเชิงเส้นที่สร้างด้วยวิธี OLS ให้ผ่านก่อนที่จะนำไปวิเคราะห์ต่อไป ด้วยการตรวจสอบคุณสมบัติของ Residuals จากการวิเคราะห์กราฟของ Residuals ได้แก่ ตรวจสอบการแจกแจงแบบปรกติ ตรวจสอบความคงที่ของความแปรปรวน ตรวจสอบความเป็นอิสระ และตรวจสอบว่าตัวแบบการถดถอยเชิงเส้นเหมาะสมกับข้อมูลหรือไม่ และกำหนดร้อยละการสูญหายในตัวแปรตาม 4 ระดับเช่นเดียวกับการศึกษาจากการจำลองข้อมูล และเปรียบเทียบประสิทธิภาพของวิธีการเติมข้อมูลสูญหายด้วยค่า AMSE จากการทำซ้ำ 15 รอบ เนื่องจากเมื่อทดลองเพิ่มจำนวนรอบทำซ้ำมากขึ้น พบว่าให้ผลลัพธ์ไม่ต่างกัน

วิธีการเติมข้อมูลสูญหายที่นำมาเปรียบเทียบ

1. วิธี RI เป็นวิธีการเติมข้อมูลสูญหายที่ใช้ตัวแบบการถดถอยมาประมาณข้อมูลสูญหาย โดยนำเฉพาะชุดข้อมูลของตัวแปรตามและตัวแปรอิสระที่ตัวแปรตามไม่มีข้อมูลสูญหายมาสร้างตัวแบบการถดถอยด้วยวิธี OLS แล้วนำตัวแบบถดถอยที่ได้ไปทำนายข้อมูลที่สูญหายของตัวแปรตาม (Little and Rubin, 2002)

2. วิธี SRI เป็นวิธีการเติมข้อมูลสูญหายที่ใช้ตัวแบบการถดถอยจากชุดข้อมูลที่ไม่สูญหายเหมือนวิธี RI แต่จะแตกต่างจากวิธี RI คือจะเพิ่มเทอมของค่าประมาณของค่าความคลาดเคลื่อน (Residual) ที่ได้จากการสุ่มด้วยวิธีมอนติคาร์โลเข้าไปในตัวแบบการถดถอยที่ได้จากวิธี RI แล้วใช้ตัวแบบถดถอยที่มีเทอมของค่าประมาณของค่าความคลาดเคลื่อนนี้ไปทำนายข้อมูลที่สูญหายของตัวแปรตาม (Enderers, 2008)

3. วิธี KNN เป็นวิธีการเติมข้อมูลสูญหายด้วยค่าเฉลี่ยของข้อมูลตัวแปรตามเฉพาะที่ไม่สูญหายจำนวน K ค่า ที่ลักษณะของข้อมูลตัวแปรตามที่ไม่สูญหายมีความคล้ายคลึงกับข้อมูลตัวแปรตามที่สูญหายมากที่สุด โดยจะพิจารณาความคล้ายของหน่วยตัวอย่างจากระยะห่างยุคลิด (Euclidean distance) ของตัวแปรอิสระ (Jösson and Wohlin, 2006) จากสูตรตามสมการ (3) ดังนี้

$$D_{(i,j)} = \sqrt{\sum_{p=1}^{3}(x_{ip} - x_{jp})^2} \qquad (3)$$

เมื่อ $i = 1,2,\ldots,m$, $j = m+1,\ldots,n$

ซึ่งวิธี KNN มีขั้นตอนดังนี้

ขั้นตอนที่ 1 คำนวณหาค่า K โดย K มีค่าเป็นเลขคี่ที่ใกล้เคียงกับ $\sqrt{m}$ มากที่สุด เมื่อ m คือจำนวนข้อมูลที่ไม่สูญหายของตัวแปรตาม

ขั้นตอนที่ 2 หาค่าระยะห่างยุคลิด D_{ij} ค่าที่ i สำหรับข้อมูลสูญหายค่าที่ j ของตัวแปรอิสระดังสมการ (3)

ขั้นตอนที่ 3 ให้ $y_{ij} = D_{ij}$ แล้วเลือก y_{ij} ที่มีค่าน้อยที่สุดจำนวน K ค่ามาหาค่าเฉลี่ย แทนด้วย $\bar{y}_j$

ขั้นตอนที่ 4 นำค่า $\bar{y}_j$ ไปทำนายข้อมูลสูญหายค่าที่ j ของตัวแปรตาม

4. วิธี EM เป็นวิธีการเติมข้อมูลสูญหายที่ใช้กระบวนการวนซ้ำเพื่อหาค่าประมาณของพารามิเตอร์ ด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) โดยแบ่งเป็น 2 ขั้นตอนคือ ขั้นตอน E-step และ M-step โดยขั้นตอน E-step เป็นการหาค่าคาดหวังของข้อมูลสูญหายภายใต้เงื่อนไขของชุดข้อมูลที่ไม่สูญหาย เพื่อนำค่านี้ไปเติมข้อมูลสูญหายในขั้นตอน M-step ส่วนขั้นตอน M-step เป็นการประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุดจากชุดข้อมูลที่ไม่สูญหาย ซึ่งจะแทนข้อมูลสูญหายจากค่าประมาณที่ได้ในขั้นตอน E-step (Little and Rubin, 2002) โดยมีขั้นตอนดังนี้

ขั้นตอนที่ 1 จัดชุดข้อมูล $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ ออกเป็น 2 ชุดคือ ชุดที่ไม่มีข้อมูลสูญหายใน y และชุดที่มีข้อมูลสูญหายใน y ดังสมการ (4)

$$\begin{bmatrix}\mathbf{y_1}\\ \mathbf{y_2}\end{bmatrix} = \begin{bmatrix}\mathbf{X_1}\\ \mathbf{X_2}\end{bmatrix}\begin{bmatrix}0\\ 1\\ \vdots\\ k\end{bmatrix} + \begin{bmatrix}£_1\\ £_2\\ \vdots\\ £_n\end{bmatrix} \qquad (4)$$

เมื่อ $\mathbf{y_1}$ คือเวกเตอร์ขนาด $m \times 1$ ของตัวแปรตาม y ที่ไม่มีข้อมูลสูญหายส่วน $\mathbf{y_2}$ คือเวกเตอร์ขนาด $(n-m)\times 1$ ของตัวแปรตาม y ที่มีข้อมูลสูญหายสำหรับ $\mathbf{X_1}$ คือเมทริกซ์ขนาด $m \times (k+1)$ ของตัวแปรอิสระที่ตัวแปรตาม y ไม่มีข้อมูลสูญหาย และ $\mathbf{X_2}$ คือเมทริกซ์ขนาด $(n-m)\times(k+1)$ ของตัวแปรอิสระที่ตัวแปรตาม y มีข้อมูลสูญหายเมื่อ k คือจำนวนตัวแปรอิสระ และ $\boldsymbol{\beta}$ คือเวกเตอร์ของพารามิเตอร์ส่วน $\boldsymbol{\varepsilon}$ คือเวกเตอร์ของความคลาดเคลื่อน

ขั้นตอนที่ 2 ประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี OLS จากชุดข้อมูลที่ไม่สูญหายในตัวแปรตาม y จะได้ค่าสัมประสิทธิ์การถดถอยเริ่มต้น ($\hat{\boldsymbol{\beta}}^{(0)}$)

ขั้นตอนที่ 3 เข้าสู่ขั้นตอน E-step รอบที่ 1 โดยการนำ $\hat{\boldsymbol{\beta}}^{(0)}$ มาหาค่าคาดหวัง จากสูตรตามสมการ (5)

$$E(y_i|\mathbf{X},\mathbf{y_1},\hat{\boldsymbol{\beta}}^{(0)}) = \begin{cases}\mathbf{y_i} & ;i=1,2,\ldots,m\\ \hat{\boldsymbol{\beta}}_{\mathbf{0}}^{(0)} + \sum_{k=1}^{p}\hat{\boldsymbol{\beta}}_{\mathbf{k}}^{(0)}\mathbf{x_{jk}} & ;j=m+1,\ldots,n\end{cases} \qquad (5)$$

ขั้นตอนที่ 4 เข้าสู่ขั้นตอน M-step รอบที่ 1 โดยนำค่าคาดหวังที่ได้ไปประมาณค่าข้อมูลสูญหายของตัวแปรตาม y เมื่อได้ชุดข้อมูลสมบูรณ์นำข้อมูลนี้ไปคำนวณค่าสัมประสิทธิ์การถดถอย ($\hat{\boldsymbol{\beta}}^{(1)}$) ด้วยวิธี OLS

ขั้นตอนที่ 5 นำตัวแบบการถดถอยที่ได้ไปประมาณค่าข้อมูลสูญหายใหม่อีกรอบซึ่งกลับเข้าสู่ขั้นตอน E-step ตามขั้นตอนที่ 3 นั่นคือนำ ($\hat{\boldsymbol{\beta}}^{(1)}$) มาหาค่าคาดหวังใหม่ ดำเนินการขั้นตอนที่ 3 และ 4 วนซ้ำเช่นนี้ไปเรื่อยๆ จนกระทั่งได้ค่าสัมบูรณ์ของผลต่างระหว่างสัมประสิทธิ์การถดถอยในขั้นตอนที่ 3 และ 4 ทุกค่าน้อยกว่าหรือเท่ากับ 0.001 จึงจะหยุดดำเนินการ แล้วจะนำตัวแบบการถดถอยตัวแบบสุดท้ายไปประมาณค่าข้อมูลสูญหายของตัวแปรตาม y

5. วิธี MI เป็นวิธีการเติมข้อมูลสูญหายที่ใช้เทคนิคการสร้างค่าหลายค่าเพื่อแทนข้อมูลสูญหายแต่ละค่า โดยค่าที่ถูกสร้างขึ้นมามีตั้งแต่ 2 ค่าขึ้นไป (Sinharay et al., 2001) มีขั้นตอนดังนี้

ขั้นตอนที่ 1 การสร้างค่าสำหรับแทนข้อมูลสูญหายโดยการสร้างชุดข้อมูลสมบูรณ์จำนวน u ชุดจากการรวมเซตข้อมูลที่ไม่สูญหาย (y^{obs}) และเซตข้อมูลที่สูญหาย(y_j^{miss}) เมื่อ u คือจำนวนครั้งของการแทนค่ามีวิธีการดังนี้

1) นำข้อมูลตัวแปรตามและตัวแปรอิสระที่ไม่มีค่าสูญหายในตัวแปรตามมาสร้างตัวแบบการถดถอยเชิงเส้นพหุคูณด้วยวิธี OLS ได้ตัวแบบถดถอยในรูป $\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathbf{obs}}$ แล้วหาค่าความแปรปรวนของค่าความคลาดเคลื่อนจากสูตรตามสมการ (6) ดังนี้

$$\pm_1^2 = \frac{\sum_{i=1}^{m}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathbf{obs}})^2}{(m-p)} \tag{6}$$

เมื่อ p คือจำนวนสัมประสิทธิ์การถดถอย และ m คือจำนวนข้อมูลที่ไม่สูญหายในตัวแปรตาม y

2) สุ่มค่า a ซึ่งเป็นตัวแปรสุ่มที่มีการแจกแจงไคกำลังสองด้วยองศาเสรีเท่ากับ $m-p$ เพื่อมาคำนวณค่า $\pm_*^2$ จากสูตรตามสมการ (7) ดังนี้

$$\pm_*^2 = \frac{\pm_1^2 (m-p)}{a} \tag{7}$$

3) สุ่มค่า Z_j ซึ่งเป็นตัวแปรสุ่มที่มีการแจกแจงปรกติมาตรฐานมาจำนวนเท่ากับสัมประสิทธิ์การถดถอย นำ Z_j มาคำนวณค่าสัมประสิทธิ์การถดถอยค่าที่ j ใหม่จากสูตรตามสมการ (8) ดังนี้

$$\hat{}_j = \hat{}_j^{obs} + \pm_* \left(\frac{S(\hat{}_j^{obs})}{\pm_1} \right) Z_j \tag{8}$$

เมื่อ $\hat{}_j^{obs}$ และ $S(\hat{}_j^{obs})$ คือ สัมประสิทธิ์การถดถอยและค่าความคลาดเคลื่อนมาตรฐานของสัมประสิทธิ์การถดถอยตัวที่ j ตามลำดับ จากตัวแบบการถดถอยที่ปรากฏในสมการ (6)

4) สุ่มค่า Z ซึ่งเป็นตัวแปรสุ่มที่มีการแจกแจงปรกติมาตรฐานมา 1 ค่า แล้วคำนวณค่าความคลาดเคลื่อนสุ่ม $Z\pm_*$ สำหรับใช้ปรับค่าสัมประสิทธิ์การถดถอย $\hat{}_j$ ในสมการ (8) ใหม่ แล้วนำตัวแบบการถดถอยใหม่นี้ไปประมาณค่าข้อมูลสูญหายของตัวแปรตาม y ดังสมการ (9) (Rubin, 2004) จะได้ชุดข้อมูลสมบูรณ์ชุดที่ 1

$$\hat{y}_j = \hat{}_0 + \hat{}_1 x_1 + \ldots + \hat{}_k x_k + Z\pm_* \tag{9}$$

5) ทำขั้นตอนที่ 2 ถึง 4 ซ้ำจนกระทั่งได้ชุดข้อมูลสมบูรณ์ u ชุดโดยงานวิจัยนี้กำหนด u เท่ากับ 5

ขั้นตอนที่ 2 นำชุดข้อมูลสมบูรณ์แต่ละชุดมาสร้างตัวแบบการถดถอยด้วยวิธี OLS จะได้ตัวแบบการถดถอย u ตัวแบบ

ขั้นตอนที่ 3 จากตัวแบบการถดถอย u ตัวแบบ ทำการหาค่าสัมประสิทธิ์การถดถอยใหม่จากค่าเฉลี่ยของสัมประสิทธิ์การถดถอยที่ตรงกัน เช่นค่า $\hat{}_0$ ใหม่คำนวณจาก $(\hat{}_{01} + \hat{}_{02} + \ldots + \hat{}_{0u})/u$ นำค่าสัมประสิทธิ์การถดถอยใหม่ที่ได้ไปสร้างเป็นตัวแบบการถดถอยที่ใช้ประมาณค่าข้อมูลสูญหายของตัวแปรตาม y

6. วิธี PRD เป็นวิธีการเติมข้อมูลสูญหายที่พัฒนาโดย Kwon and Park (2015) ซึ่งวิธี PRD จะคล้ายกับวิธี MI คือมีการสร้างชุดข้อมูลที่สมบูรณ์ u ชุด เหมือนวิธี MI แต่วิธีนี้ได้นำเสนอค่าสัดส่วนความคลาดเคลื่อนมาใช้ในการเติมข้อมูลสูญหายของตัวแปรตาม y โดยมีขั้นตอนดังนี้

ขั้นตอนที่ 1 การสร้างค่าสำหรับแทนข้อมูลสูญหายมีวิธีการดังนี้

1) หาค่าสัมประสิทธิ์การถดถอยตัวเริ่มต้นด้วยวิธี OLS เฉพาะชุดข้อมูลที่ไม่สูญหาย ($\hat{\boldsymbol{\beta}}^{(obs)}$)

2) คำนวณหาค่าสัดส่วนความคลาดเคลื่อนจากสูตรตามสมการ (10) ดังนี้

$$\tilde{r}_i = \frac{y_i - \hat{y}_i^{obs}}{C_i - \hat{y}_i^{obs}}; i = 1, 2, \ldots, m \qquad (10)$$

โดย $y_i - \hat{y}_i^{obs} \leq C_i - \hat{y}_i^{obs}$

เมื่อ y_i คือ ข้อมูลของตัวแปรตาม y ที่ไม่สูญหายค่าที่ i ส่วน $\hat{y}_i^{obs}$ คือค่าทำนายที่ i ของตัวแปรตาม y จากตัวแบบการถดถอยในข้อ 1) ส่วน $C_i = C$ เมื่อ C คือข้อมูลของตัวแปรตาม y ที่ไม่สูญหายที่มีค่ามากที่สุด

3) สุ่มค่า $\tilde{r}_i$ จะได้ค่า $C_i - \hat{y}_i^{obs}$ และค่า $x_{i1}, x_{i2}, \ldots, x_{ik}$ ในแถวที่ตรงกับค่า $\tilde{r}_i$

4) นำค่าที่ได้จากข้อ 3) และตัวแบบการถดถอยที่ได้จากข้อ 1) ไปประมาณค่าข้อมูลสูญหายของตัวแปรตาม y จากสูตรตามสมการ (11) จะได้ชุดข้อมูลสมบูรณ์ชุดที่ 1

$$\hat{y}_j = \hat{\ }_0^{obs} + \hat{\ }_1^{obs} x_{i1} + \ldots + \hat{\ }_k^{obs} x_{ik} + \tilde{r}_i (C_i - \hat{y}_i^{obs}) \qquad (11)$$

5) ทำข้อ 2) ถึง 4) ซ้ำจนกระทั่งได้ชุดข้อมูลสมบูรณ์ u ชุด โดยงานวิจัยนี้กำหนด u เท่ากับ 5

ขั้นตอนที่ 2 ดำเนินการเหมือนวิธี MI ในขั้นตอนที่ 2 ถึง 3

วิธีการเติมข้อมูลสูญหายที่พัฒนาขึ้น

1. วิธี MRI (Mean Regression Imputation method) เป็นวิธีการเติมข้อมูลสูญหายวิธีแรกที่งานวิจัยนี้พัฒนาขึ้น โดยเป็นวิธีร่วมที่พัฒนามาจากวิธี Mean และวิธี RI มีขั้นตอนดังนี้

ขั้นตอนที่ 1 ดำเนินการเหมือนวิธี Mean คือหาค่าเฉลี่ยของตัวแปรตามเฉพาะข้อมูลที่ไม่สูญหาย แล้วนำค่าเฉลี่ยที่ได้ไปเติมข้อมูลที่สูญหายของตัวแปรตาม จะทำให้ข้อมูลที่สูญหายทุกค่าของตัวแปรตามมีค่าเดียวกันคือ $\bar{y}$

ขั้นตอนที่ 2 นำชุดข้อมูลสมบูรณ์ที่ได้ไปสร้างตัวแบบการถดถอยด้วยวิธี OLS

ขั้นตอนที่ 3 นำตัวแบบการถดถอยที่ได้ไปทำนายค่าตัวแปรตามที่สูญหายอีกครั้งซึ่งจะเห็นว่าขั้นตอนที่ 2 และ 3 ของวิธี MRI จะคล้ายกับวิธี RI แต่วิธี RI เป็นการสร้างตัวแบบการถดถอยด้วยวิธี OLS จากชุดข้อมูลเฉพาะที่ตัวแปรตามไม่สูญหายโดยไม่ได้ทำการประมาณค่าข้อมูลสูญหายเบื้องต้นขึ้นมาก่อน

2. วิธี EMMI (Expectation Maximization and Multiple Imputation method) เป็นวิธีการเติมข้อมูลสูญหายวิธีที่ 2 ที่งานวิจัยนี้พัฒนาขึ้น โดยเป็นวิธีร่วมที่พัฒนามาจากวิธี EM และวิธี MI มีขั้นตอนหลัก 2 ขั้นตอนดังนี้

ขั้นตอนที่ 1 การสร้างค่าสำหรับแทนข้อมูลสูญหาย ดำเนินการดังนี้

1) ดำเนินการตามขั้นตอนที่ 1 ถึง 5 เหมือนวิธี EM เพียงแต่ในขั้นตอนที่ 5 ของวิธี EM นั้นจะสิ้นสุดการดำเนินการ แล้วนำสัมประสิทธิ์การถดถอยสุดท้ายไปสร้างตัวแบบการถดถอยเพื่อประมาณค่าข้อมูลสูญหายของตัวแปรตาม y เลย แต่วิธี EMMI จะแตกต่างคือเมื่อเสร็จสิ้นขั้นตอนที่ 5 แล้วซึ่งจะได้ชุดของ $\hat{\ }_l$ เท่ากับจำนวนรอบคือ u ชุด จะดำเนินการต่อไปในข้อ 2)

2) ดำเนินการตามข้อ 1) ข้อ 2) และข้อ 4) ในขั้นตอนที่ 1 ของวิธี MI โดยนำค่าสัมประสิทธิ์การถดถอย $\hat{\ }_l$ จำนวน u ชุดมาสร้างตัวแบบการถดถอยตามสมการ (9) จะได้ตัวแบบการถดถอยจำนวน u ตัวแบบ นำตัวแบบการถดถอยแต่ละตัวแบบไปประมาณค่าข้อมูลสูญหายของตัวแปรตาม y นั่นคือเมื่อเสร็จสิ้นขั้นตอนนี้จะได้ชุดข้อมูลที่สมบูรณ์จำนวน u ชุด

ขั้นตอนที่ 2 ดำเนินการเหมือนวิธี MI ในขั้นตอนที่ 2 ถึง 3

3. วิธี NARI (Nearest Average Regression Imputation method) เป็นวิธีที่ใช้หลักการเติมข้อมูลสูญหายด้วยค่าเฉลี่ยของตัวแปรตามที่อยู่ใกล้เคียงร่วมกับวิธี RI โดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์มาพิจารณา มีขั้นตอนดังนี้

ขั้นตอนที่ 1 หาค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson's correlation coefficient: *r*)ระหว่างตัวแปรตามกับตัวแปรอิสระทุกตัวแปร โดยนำเฉพาะชุดข้อมูลที่ไม่สูญหายในตัวแปรตาม y มาคำนวณ แล้วเลือกคู่ลำดับ (x, y) ที่มีค่า r สูงที่สุดมาเรียงลำดับจากค่าน้อยไปหาค่ามากตามค่าของตัวแปรอิสระ x

ขั้นตอนที่ 2 จากคู่ลำดับ (x, y) ที่ถูกเลือกนำข้อมูลของตัวแปรตาม y ที่อยู่ต่ำกว่าข้อมูลสูญหาย 1 ระดับและอยู่สูงกว่าข้อมูลสูญหาย 1 ระดับมาหาค่าเฉลี่ย แล้วนำค่าเฉลี่ยที่ได้ไปประมาณข้อมูลสูญหายนั้น โดยทำการประมาณข้อมูลสูญหายของตัวแปรตาม y ทุกค่าด้วยวิธีการเดียวกันนี้ ในกรณีที่ข้อมูลสูญหายเป็นข้อมูลตัวแรกหรือข้อมูลตัวสุดท้าย จะใช้ค่าเฉลี่ยของข้อมูล 2 ค่าที่อยู่ติดกับข้อมูลสูญหายเป็นค่าประมาณข้อมูลที่สูญหายนั้น

ขั้นตอนที่ 3 นำชุดข้อมูลสมบูรณ์ที่ได้ไปทำการสร้างตัวแบบการถดถอยด้วยวิธี OLS

ขั้นตอนที่ 4 นำตัวแบบการถดถอยที่ได้ไปทำนายข้อมูลของตัวแปรตามที่สูญหาย

การพิจารณาประสิทธิภาพของวิธีการเติมข้อมูลสูญหายที่งานวิจัยนี้พัฒนาขึ้น ทำโดยเปรียบเทียบค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Average Mean Square Error: AMSE) กับวิธีการเติมข้อมูลสูญหายวิธีอื่น 6 วิธีที่ได้กล่าวไปแล้ว ได้แก่วิธี RI วิธี SRI วิธี KNN วิธี EM วิธี MI และวิธี PRD

ผลการวิจัย

ผลจากการจำลองข้อมูล

การเปรียบเทียบประสิทธิภาพของวิธีการเติมข้อมูลสูญหายทั้ง 9 วิธีด้วยการพิจารณาค่า AMSE เมื่อส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อน ($\pm$) เท่ากับ 5, 10 และ 15 ทุกระดับขนาดตัวอย่าง (n) และร้อยละการสูญหายแสดงในตารางที่ 1 ถึง 3 ตามลำดับ พบว่าวิธีการเติมข้อมูลสูญหายที่งานวิจัยนี้พัฒนาขึ้นมีประสิทธิภาพดีในหลายกรณี โดยจากตารางที่ 1 กรณี $\pm$ เท่ากับ 5 และร้อยละการสูญหายเท่ากับ 5 พบว่าวิธี EMMI ให้ค่า AMSE ต่ำสุดเท่ากับวิธี RI และวิธี EM ในทุกระดับค่า n ส่วนที่ร้อยละการสูญหายระดับอื่น วิธี EMMI ก็ให้ค่า AMSE ต่ำใกล้เคียงกับวิธี RI และวิธี EM ซึ่งเป็น 2 วิธีที่ให้ค่า AMSE ต่ำสุดเท่ากัน จากตารางที่ 2 กรณี $\pm$ เท่ากับ 10 และร้อยละการสูญหายเท่ากับ 5 พบว่าวิธี MRI ให้ค่า AMSE ต่ำสุดในทุกระดับค่า n และวิธี EMMI ก็ให้ค่า AMSE ต่ำใกล้เคียงกับวิธี MRI จากตารางที่ 3 กรณี $\pm$ เท่ากับ 15 พบว่าวิธี MRI ให้ค่า AMSE ต่ำสุดเกือบทุกระดับค่า n และร้อยละการสูญหาย นอกจากนั้น เมื่อพิจารณาจากทุกตาราง พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ทุกวิธีจะให้ค่า AMSE ลดลง เนื่องจากการเพิ่มขึ้นของขนาดตัวอย่างจะช่วยให้ค่าความคลาดเคลื่อนจากการทำนายลดลง ส่วนเมื่อค่า $\pm$ เพิ่มขึ้น ร้อยละการสูญหายเพิ่มขึ้น พบว่าจะมีผลทำให้ค่า AMSE ของทุกวิธีเพิ่มขึ้น

ผลจากข้อมูลจริง

ผลการเปรียบเทียบประสิทธิภาพของวิธีการเติมข้อมูลสูญหายจากข้อมูลคุณภาพอากาศของพื้นที่ที่มีมลพิษในประเทศอิตาลีด้วยค่า AMSE แสดงดังในตารางที่ 4 โดยศึกษาเฉพาะกรณีที่ n เท่ากับ 50 ซึ่งพบว่าให้ผลเหมือนกรณีศึกษาจากการจำลองข้อมูลที่ $\pm$ เท่ากับ 15

ตารางที่ 1 ค่า AMSE ของวิธีการเติมข้อมูลสูญหาย 9 วิธี เมื่อส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อน (±) เท่ากับ 5 ในแต่ละระดับขนาดตัวอย่าง (*n*) และแต่ละระดับร้อยละการสูญหาย (%)

n	%	วิธีการเติมข้อมูลสูญหาย								
		RI	SRI	KNN	EM	MI	PRD	MRI	EMMI	NARI
30	5	***3.731***	4.023	4.788	***3.731***	3.737	4.058	4.015	***3.731***	4.034
	10	***3.888***	4.342	5.707	***3.888***	3.897	4.590	4.504	3.948	4.618
	15	***4.208***	5.257	8.145	***4.208***	4.244	6.306	6.189	4.353	6.558
	20	***4.415***	5.867	9.677	***4.415***	4.475	7.750	7.482	4.682	7.942
50	5	***2.178***	2.365	2.656	***2.178***	***2.178***	2.234	2.230	***2.178***	2.286
	10	***2.291***	2.726	3.383	***2.291***	2.299	2.582	2.546	2.308	2.642
	15	***2.448***	3.187	4.804	***2.448***	2.463	3.480	3.452	2.513	3.579
	20	***2.570***	3.774	6.144	***2.570***	2.603	4.549	4.503	2.704	4.664
100	5	***1.048***	1.156	1.249	***1.048***	***1.048***	1.071	1.069	***1.048***	1.083
	10	***1.112***	1.417	1.687	***1.112***	1.115	1.249	1.242	1.115	1.289
	15	***1.177***	1.818	2.314	***1.177***	***1.177***	1.639	1.617	1.208	1.686
	20	***1.274***	2.321	3.210	***1.274***	1.297	2.401	2.384	1.370	2.441
200	5	***0.527***	0.601	0.624	***0.527***	***0.527***	***0.527***	***0.527***	***0.527***	0.544
	10	***0.554***	0.865	0.887	***0.554***	0.556	0.636	0.633	0.556	0.650
	15	***0.583***	1.243	1.293	***0.583***	0.588	0.893	0.883	0.602	0.883
	20	***0.622***	1.653	1.895	***0.622***	0.630	1.447	1.428	0.671	1.389

หมายเหตุ: ตัวหนาและเอียงหมายถึงค่า AMSE ต่ำที่สุด

ตารางที่ 2 ค่า AMSE ของวิธีการเติมข้อมูลสูญหาย 9 วิธี เมื่อส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อน (±) เท่ากับ 10 ในแต่ละระดับขนาดตัวอย่าง (*n*) และแต่ละระดับร้อยละการสูญหาย (%)

n	%	วิธีการเติมข้อมูลสูญหาย								
		RI	SRI	KNN	EM	MI	PRD	MRI	EMMI	NARI
30	5	14.852	16.094	15.408	14.852	14.832	14.810	***14.767***	14.921	15.183
	10	15.571	17.770	16.713	15.571	***15.562***	15.721	15.615	15.640	16.269
	15	***16.987***	21.876	19.375	***16.987***	***16.987***	17.782	17.651	17.524	18.918
	20	***17.740***	22.999	21.229	***17.740***	17.851	19.423	19.104	18.744	20.757
50	5	***8.662***	9.410	9.153	***8.662***	8.665	8.718	***8.662***	8.665	8.901
	10	9.096	10.394	9.900	9.096	***9.080***	9.231	9.193	9.147	9.604
	15	***9.868***	12.673	11.603	***9.868***	9.889	10.321	10.278	10.029	11.223
	20	***10.537***	16.144	13.259	***10.537***	10.649	11.712	11.538	11.076	12.742
100	5	4.207	4.614	4.335	4.207	4.207	***4.202***	***4.202***	4.207	4.261
	10	***4.465***	5.717	4.953	***4.465***	***4.465***	4.557	4.556	4.506	4.717
	15	***4.747***	7.344	5.699	***4.747***	4.797	5.066	5.067	4.875	5.394
	20	***5.073***	9.835	6.775	***5.073***	5.133	6.010	5.974	5.394	6.485
200	5	***2.069***	2.419	2.170	***2.069***	***2.069***	***2.069***	***2.069***	***2.069***	2.104
	10	***2.179***	3.354	2.475	***2.179***	2.187	2.227	2.224	2.207	2.325
	15	***2.319***	4.953	2.961	***2.319***	2.340	2.560	2.554	2.355	2.749
	20	***2.476***	6.659	3.615	***2.476***	2.516	3.151	3.148	2.668	3.433

หมายเหตุ: ตัวหนาและเอียงหมายถึงค่า AMSE ต่ำที่สุด

ตารางที่ 3 ค่า AMSE ของวิธีการเติมข้อมูลสูญหาย 9 วิธี เมื่อส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อน (±) เท่ากับ 15 ในแต่ละระดับขนาดตัวอย่าง (*n*) และแต่ละระดับร้อยละการสูญหาย (%)

n	%	วิธีการเติมข้อมูลสูญหาย								
		RI	SRI	KNN	EM	MI	PRD	MRI	EMMI	NARI
30	5	33.349	35.930	***32.915***	33.349	33.348	32.982	***32.915***	33.518	33.962
	10	34.662	39.232	34.374	34.662	34.898	34.315	***34.186***	35.113	35.758
	15	38.463	46.937	38.468	38.463	38.714	38.015	***37.784***	40.252	41.231
	20	40.302	53.382	40.274	40.302	41.085	39.676	***39.403***	42.135	44.141
50	5	19.559	20.674	19.532	19.559	19.571	***19.362***	***19.362***	19.556	19.845
	10	20.540	24.447	20.522	20.540	20.579	20.264	***20.218***	20.638	21.113
	15	22.363	30.478	22.650	22.363	22.592	21.950	***21.856***	23.094	23.566
	20	23.466	35.132	24.344	23.466	23.594	23.407	***23.201***	24.534	25.571
100	5	9.699	10.333	9.636	9.699	9.708	***9.625***	***9.625***	9.699	9.752
	10	10.423	13.387	10.470	10.423	10.407	10.263	***10.261***	10.533	10.703
	15	11.122	17.009	11.340	11.122	11.231	10.911	***10.888***	11.443	11.733
	20	11.952	21.819	12.545	11.952	12.115	11.971	***11.875***	12.560	13.298
200	5	4.758	5.521	4.758	4.758	4.758	***4.733***	***4.733***	4.758	4.794
	10	5.040	7.480	5.157	5.040	5.056	***5.007***	***5.007***	5.078	5.216
	15	***5.282***	11.136	5.671	***5.282***	5.325	5.378	5.358	5.456	5.806
	20	***5.648***	15.762	6.428	***5.648***	5.745	6.024	6.010	6.167	6.741

หมายเหตุ: ตัวหนาและเอียงหมายถึงค่า AMSE ต่ำที่สุด

ตารางที่ 4 ค่า AMSE ของวิธีการเติมข้อมูลสูญหาย 9 วิธี สำหรับข้อมูลคุณภาพอากาศของพื้นที่ที่มีมลพิษในประเทศอิตาลี แต่ละระดับร้อยละการสูญหาย (%) เมื่อขนาดตัวอย่าง *n* เท่ากับ 50

%	วิธีการเติมข้อมูลสูญหาย								
	RI	SRI	KNN	EM	MI	PRD	MRI	EMMI	NARI
5	0.621	0.004	***0.003***	0.621	0.621	***0.003***	***0.003***	0.621	0.621
10	0.623	0.029	***0.011***	0.623	0.623	***0.011***	***0.011***	0.623	0.623
15	0.640	0.093	0.037	0.640	0.655	0.035	***0.033***	0.641	0.632
20	0.643	0.104	0.051	0.643	0.645	0.052	***0.050***	0.644	0.637

หมายเหตุ: ตัวหนาและเอียงหมายถึงค่า AMSE ต่ำที่สุด

สรุปผลการวิจัยและวิจารณ์ผล

งานวิจัยนี้ได้พัฒนาวิธีการเติมข้อมูลสูญหาย 3 วิธี คือวิธี MRI วิธี EMMI และวิธี NARI เมื่อตัวแปรตามมีการสูญหายแบบสุ่ม สำหรับการถดถอยเชิงเส้นพหุคูณ โดยเปรียบเทียบประสิทธิภาพวิธีการเติมข้อมูลสูญหายที่พัฒนาขึ้นกับอีก 6 วิธีด้วยค่า AMSE กรณีศึกษาจากการจำลองข้อมูล ผลการวิจัยสรุปดังในตารางที่ 5 ซึ่งพบว่าวิธีการเติมข้อมูลสูญหายที่งานวิจัยนี้พัฒนาขึ้นมีประสิทธิภาพดีในหลายกรณีดังนี้ กรณี ± เท่ากับ 5 พบว่าวิธี EMMI มีประสิทธิภาพดีที่สุดที่ทุกระดับ *n* เมื่อร้อยละการสูญหายเท่ากับ 5 โดยมีประสิทธิภาพดีเทียบเท่ากับวิธีอื่นบางวิธีรวมถึงวิธี MRI ด้วยที่มีประสิทธิภาพดีที่ *n* เท่ากับ 200 กรณีที่± เท่ากับ 10 พบว่าวิธี MRI มีประสิทธิภาพดีที่สุดที่ทุกระดับ *n* เมื่อร้อยละการสูญหายเท่ากับ 5 ในทำนองเดียวกันวิธี MRI ก็มีประสิทธิภาพดีเทียบเท่ากับวิธีอื่นบางวิธีรวมถึงวิธี EMMI ด้วยที่มีประสิทธิภาพดีที่ *n* เท่ากับ 200 กรณี± เท่ากับ 15 พบว่าวิธี MRI มีประสิทธิภาพดีที่สุดที่ *n* เท่ากับ 30, 50 และ 100 ในทุกระดับร้อยละการสูญหายส่วน *n* เท่ากับ 200 วิธี MRI มีประสิทธิภาพดีที่สุดเมื่อระดับร้อยละการสูญหายเท่ากับ 5 และ 10 จะเห็นว่า แม้ว่าวิธีการเติมข้อมูลสูญหายที่งานวิจัยนี้พัฒนาขึ้นไม่ได้มีประสิทธิภาพดีที่สุดทุกกรณี แต่ก็มีประสิทธิภาพดีกว่าหลายวิธี

ที่นำมาเปรียบเทียบ เช่น กรณี ± เท่ากับ 5 ก็พบว่าวิธี EMMI มีประสิทธิภาพดีรองลงมาเป็นอันดับ 2 หรือ 3 ในทุกระดับ *n* และร้อยละการสูญหาย ส่วนที่ ± เท่ากับ 10 ก็พบว่าวิธี EMMI มีประสิทธิภาพดีรองลงมาเป็นอันดับ 2 หรือ 3 เกือบทุกระดับ *n* และเกือบทุกระดับร้อยละการสูญหาย

กรณีศึกษาจากข้อมูลจริง พบว่าวิธีการเติมข้อมูลสูญหายที่งานวิจัยนี้พัฒนาขึ้นคือวิธี MRI มีประสิทธิภาพดีที่สุดในทุกระดับร้อยละการสูญหาย ซึ่งให้ผลสอดคล้องกับผลการศึกษาจากการจำลองข้อมูลที่ ± เท่ากับ 15 และ *n* เท่ากับ 50 ทุกระดับร้อยละการสูญหาย

จากผลการศึกษาในงานวิจัยนี้ พบว่ามีความสอดคล้องกับงานวิจัยของจริยา (2551) ที่ว่าวิธี RI เป็นวิธีการเติมข้อมูลสูญหายที่มีประสิทธิภาพเมื่อ ± เท่ากับ 5 และ *n* เท่ากับ 50, 100, 200 และร้อยละการสูญหายเท่ากับ 10 และ 20 ซึ่งจริยาไม่ได้ทำการศึกษาเปรียบเทียบกับวิธี EM ส่วนเมื่อเปรียบเทียบกับงานวิจัยของอุษณีย์ (2555) พบว่าผลการวิจัยมีความสอดคล้องกันคือ วิธี EM เป็นวิธีการเติมข้อมูลสูญหายที่มีประสิทธิภาพเมื่อ ± เท่ากับ 10 และ *n* เท่ากับ 50, 100, 200 และร้อยละการสูญหายเท่ากับ 10 และ 20 ซึ่งอุษณีย์ไม่ได้ทำการศึกษาเปรียบเทียบกับวิธี RI

ตารางที่ 5 วิธีการเติมข้อมูลสูญหายที่มีประสิทธิภาพในแต่ละระดับค่าส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อน(±)ขนาดตัวอย่าง (*n*) และร้อยละการสูญหาย (%)

n	%	± = 5	± = 10	± = 15
30	5	RI , EM , ***EMMI***	***MRI***	KNN , ***MRI***
	10	RI , EM	MI	***MRI***
	15	RI , EM	RI , EM , MI	***MRI***
	20	RI , EM	RI , EM	***MRI***
50	5	RI , EM , MI , ***EMMI***	RI , EM , ***MRI***	PRD , ***MRI***
	10	RI , EM	MI	***MRI***
	15	RI , EM	RI , EM	***MRI***
	20	RI , EM	RI , EM	***MRI***
100	5	RI , EM , MI , ***EMMI***	PRD , ***MRI***	PRD , ***MRI***
	10	RI , EM	RI , EM , MI	***MRI***
	15	RI , EM , MI	RI , EM	***MRI***
	20	RI , EM	RI , EM	***MRI***
200	5	RI , EM , MI , PRD , ***MRI*** , ***EMMI***	RI , EM , MI , PRD ***MRI*** , ***EMMI***	PRD , ***MRI***
	10	RI , EM	RI , EM , MI	***MRI***
	15	RI , EM , MI	RI , EM	***MRI***
	20	RI , EM	RI , EM	***MRI***

หมายเหตุ: ตัวหนาและเอียงหมายถึงวิธีการเติมข้อมูลสูญหายที่งานวิจัยนี้พัฒนาขึ้น

เอกสารอ้างอิง

จริยา แสงสุวรรณ. (2551). การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยพหุคูณ. วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต ,มหาวิทยาลัยเกษตรศาสตร์.กรุงเทพฯ: 98 หน้า.

นรุตม์ บุตรพลอย. (2553). การประยุกต์ Soft Computing และ k-Nearest Neighbor เพื่อใช้ประมาณค่าสูญหายของข้อมูล. National Conference on Information Technology. 28: 25-29.

เรืองลักษณ์ หลำใจซื่อ. (2560). การเปรียบเทียบวิธีการประมาณค่าสูญหายสาหรับการวิเคราะห์การถดถอยเมื่อตัวแปรตามมีการสูญหายแบบสุ่ม. วารสารวิทยาศาสตร์และเทคโนโลยี 25(5): 766-777.

อุษณีย์ วงศ์อามาตย์. (2555). การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิลในการวิเคราะห์การถดถอยเชิงเส้นพหุ. วิทยานิพนธ์ปริญญาสถิติศาสตรมหาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย. กรุงเทพฯ: 104 หน้า.

Barnett, V. (2002). Sample Servey Principle & Methods. New York: Arnold. pp. 82-89.

Dielman, T.E. (2004). Applied Regression Analysis.Boston: Brooks/Cole. pp. 171-220.

Enderers, C.K. (2008). Applied Missing Data Analysis. New York: The Guilford Press. pp. 46-48.

Jösson, P. and Wohlin,C. (2006). Benchmarking K Nearest Neighbor Imputation with Homogeneous Likert Data**.** Empirical Software Engineering 11(3): 463-489.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. New York: John Wiley & Sons. pp. 24-27.

Kwon, T.Y. and Park, Y. (2015). A new multiple imputation method for bounded missing values. Statistics and Probability Letters 107: 204-209.

Rubin, D.B. (2004). Multiple Imputation forNonresponse in Surveys.New York: John Wiley & Sons. pp. 15-22.

Sinharay, S., Stern Hal, S. and Russell, D. (2001). The Use Multiple Imputation for the Analysis of Missing Data.Psychological Methods 6: 317-329.

Vito, S.D. (2016). Air Quality Data Set. from: https://archive.ics.uci.edu/ml/datasets/Air+Quality. Retrieved 12 September 2018.

❑❑❑❑❑